



ENCYCLOPEDIA
OF
FINANCIAL MODELS



I

FRANK J. FABOZZI, EDITOR

ENCYCLOPEDIA
OF
FINANCIAL MODELS

Volume I

ENCYCLOPEDIA
OF
FINANCIAL MODELS

Volume I

FRANK J. FABOZZI, EDITOR



WILEY

John Wiley & Sons, Inc.

Cover image (top): © Jamie Farrant/iStockphoto.
Cover image (bottom) (gold background): © kyoshino/iStockphoto.
Cover image (bottom) (numbers): © Dimitris Stephanides/iStockphoto.
Cover design: Michael J. Freeland.

Copyright © 2013 by Frank J. Fabozzi. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit www.wiley.com.

ISBN: 978-1-118-00673-3 (3 v. set : cloth)

ISBN: 978-1-118-01032-7 (v. 1: cloth)

ISBN: 978-1-118-53976-7 (ebk.)

ISBN: 978-1-118-53985-9 (ebk.)

ISBN: 978-1-118-53986-6 (ebk.)

ISBN: 978-1-118-18236-5 (ebk.)

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

About the Editor

Frank J. Fabozzi is Professor of Finance at EDHEC Business School and a member of the EDHEC Risk Institute. Prior to joining EDHEC in August 2011, he held various professorial positions in finance at Yale University's School of Management from 1994 to 2011 and from 1986 to 1992 was a visiting professor of finance and accounting at MIT's Sloan School of Management. From 2008 to 2011, he was an affiliated professor in the Institute of Statistics, Econometrics, and Mathematical Finance at the University of Karlsruhe in Germany. Prior to 1986 he held professorial positions at Lafayette College, Fordham University, Queens College (CUNY), and Hofstra University. From 2003 to 2011, he served on Princeton University's Advisory Council for the Department of Operations Research and Financial Engineering and since then has been a visiting fellow in that department.

Professor Fabozzi is the editor of the *Journal of Portfolio Management*, as well as on the editorial board of the *Journal of Fixed Income*, *Journal of Asset Management*, *Quantitative Finance*, *Review of Futures Markets*, *Journal of Mathematical Finance*, *Journal of Structured Finance*, *Annals of Financial Economics*, and *Theoretical Economic Letters*.

He has authored and edited a number of books in asset management and quantitative finance. His coauthored books in quantitative finance include *A Probability Metrics Approach to Financial Risk Measures* (2011), *Financial Modeling with Lévy Processes and Volatility Clustering* (2011), *Quantitative Equity Investing: Techniques and Strategies* (2010), *Probability and Statistics for Finance* (2010), *Simulation and Optimization Modeling in Finance* (2010), *Bayesian Methods in Finance* (2008), *Advanced Stochastic Models, Risk Assessment, and Portfolio Optimization: The Ideal Risk* (2008), *Financial Econometrics: From Basics to Advanced Modeling Techniques* (2007), *Robust Portfolio Optimization and Management* (2007), and *Mathematics of Financial Modeling and Investment Management* (2004). His books in applied mathematics include *The Methods of Distances in the Theory of Probability and Statistics* (2013) and *Robust and Non-Robust Models in Statistics* (2009). He coauthored three monographs for the Research Foundation of the CFA Institute: *The Impact of the Financial Crisis on the Asset Management Industry* (2010), *Challenges in Quantitative Equity Management* (2008), and *Trends in Quantitative Finance* (2006).

Professor Fabozzi's research papers have appeared in numerous journals, including *Journal of Finance*, *Journal of Finance and Quantitative Analysis*, *Econometric Theory*, *Operations Research*, *Journal of Banking and Finance*, *Journal of Economic Dynamics and Control*, *Studies in Nonlinear Dynamics and Econometrics*, *European Journal of Operational Research*, *Annals of Operations Research*, *Quantitative Finance*, *European Financial Management*, and *The Econometric Journal*. His 2010 article published in *European Financial Management* with Professors Robert Shiller, and Radu Tunaru, "Property Derivatives for Managing European Real-Estate Risk," received the Best Paper Award and his paper with the same coauthors entitled "A Pricing Framework for Real Estate Derivatives" was awarded

Best Research Paper at the 10th Research Conference Campus for Finance held annually at WHU Otto Beisheim School of Management, Vallendar, Germany. An article coauthored with Dr. Sergio Focardi, "An Autoregressive Conditional Duration Model of Credit Risk Contagion," published in 2005 in *Journal of Risk Finance* was the winner of the 2006 Outstanding Paper by Emerald Literati Network.

He has received several awards and honors for his body of work. In 1994 he was awarded an Honorary Doctorate of Humane Letters from Nova Southeastern University. In 2002 he was inducted into the Fixed Income Analysts Society's Hall of Fame, established by the society "to recognize the lifetime achievements of outstanding practitioners in the advancement of the analysis of fixed-income securities and portfolios." In 2007 he was the recipient of the C. Stewart Sheppard Award given by the CFA Institute "in recognition of outstanding contribution to continuing education in the CFA profession." He was the cover story in the July 1999 issue of *Bloomberg Magazine* entitled "The Boswell of Bonds."

Professor Fabozzi was the co-founder of Information Management Network (now a subsidiary of Euromoney), a conference company specializing in financial topics. He is a trustee for the BlackRock family of closed-end funds where he is the chair of the performance committee and a member of the audit committee. He was a director of Guardian Mutual Funds and Guardian Annuity Funds.

He earned both an M.A. and B.A. in economics and statistics in June 1970 from the City College of New York and elected to Phi Beta Kappa in 1969. He earned a Ph.D. in Economics in September 1972 from the City University of New York. Professor Fabozzi holds two professional designations: Chartered Financial Analyst (1977) and Certified Public Accountant (1982).

Contents

Contributors	xi		
Preface	xvii		
Guide to the <i>Encyclopedia of Financial Models</i>	xxxiii		
Index	569		
Volume I			
Asset Allocation	1		
Mean-Variance Model for Portfolio Selection	3		
Principles of Optimization for Portfolio Selection	21		
Asset Allocation and Portfolio Construction Techniques in Designing the Performance-Seeking Portfolio	35		
Asset Pricing Models	47		
General Principles of Asset Pricing	49		
Capital Asset Pricing Models	65		
Modeling Asset Price Dynamics	79		
Arbitrage Pricing: Finite-State Models	99		
Arbitrage Pricing: Continuous-State, Continuous-Time Models	121		
Bayesian Analysis and Financial Modeling Applications	137		
Basic Principles of Bayesian Analysis	139		
Introduction to Bayesian Inference	151		
Bayesian Linear Regression Model	163		
Bayesian Estimation of ARCH-Type Volatility Models	175		
Bayesian Techniques and the Black-Litterman Model	189		
Bond Valuation		207	
Basics of Bond Valuation		209	
Relative Value Analysis of Fixed-Income Products		225	
Yield Curves and Valuation Lattices		235	
Using the Lattice Model to Value Bonds with Embedded Options, Floaters, Options, and Caps/Floors		243	
Understanding the Building Blocks for OAS Models		257	
Quantitative Models to Value Convertible Bonds		271	
Quantitative Approaches to Inflation-Indexed Bonds		277	
Credit Risk Modeling		297	
An Introduction to Credit Risk Models		299	
Default Correlation in Intensity Models for Credit Risk Modeling		313	
Structural Models in Credit Risk Modeling		341	
Modeling Portfolio Credit Risk		361	
Simulating the Credit Loss Distribution		377	
Managing Credit Spread Risk Using Duration Times Spread (DTS)		391	
Credit Spread Decomposition		401	
Credit Derivatives and Hedging Credit Risk		407	
Derivatives Valuation		421	
No-Arbitrage Price Relations for Forwards, Futures, and Swaps		423	
No-Arbitrage Price Relations for Options		437	
Introduction to Contingent Claims Analysis		457	
Black-Scholes Option Pricing Model		465	

Pricing of Futures/Forwards and Options	477	Classification and Regression Trees and Their Use in Financial Modeling	375
Pricing Options on Interest Rate Instruments	489	Applying Cointegration to Problems in Finance	383
Basics of Currency Option Pricing Models	507	Nonlinearity and Nonlinear Econometric Models in Finance	401
Credit Default Swap Valuation	525	Robust Estimates of Betas and Correlations	437
Valuation of Fixed Income Total Return Swaps	541	Working with High-Frequency Data	449
Pricing of Variance, Volatility, Covariance, and Correlation Swaps	545	Financial Modeling Principles	465
Modeling, Pricing, and Risk Management of Assets and Derivatives in Energy and Shipping	555	Milestones in Financial Modeling	467
		From Art to Financial Modeling	479
		Basic Data Description for Financial Modeling and Analysis	485
		Time Series Concepts, Representations, and Models	501
		Extracting Risk-Neutral Density Information from Options Market Prices	521
		Financial Statement Analysis	529
		Financial Statements	531
		Financial Ratio Analysis	545
		Cash-Flow Analysis	565
		Finite Mathematics for Financial Modeling	579
		Important Functions and Their Features	581
		Time Value of Money	595
		Fundamentals of Matrix Algebra	621
		Difference Equations	629
		Differential Equations	643
		Partial Differential Equations in Finance	659
		Model Risk and Selection	689
		Model Risk	691
		Model Selection and Its Pitfalls	699
		Managing the Model Risk with the Methods of the Probabilistic Decision Theory	719
		Fat-Tailed Models for Risk Estimation	731
		Volume III	
		Mortgage-Backed Securities Analysis and Valuation	1
		Valuing Mortgage-Backed and Asset-Backed Securities	3
		The Active-Passive Decomposition Model for MBS	17
		Analysis of Nonagency Mortgage-Backed Securities	29
Volume II			
Equity Models and Valuation	1		
Dividend Discount Models	3		
Discounted Cash Flow Methods for Equity Valuation	15		
Relative Valuation Methods for Equity Analysis	33		
Equity Analysis in a Complex Market	47		
Equity Portfolio Selection Models in Practice	61		
Basics of Quantitative Equity Investing	89		
Quantitative Equity Portfolio Management	107		
Forecasting Stock Returns	121		
Factor Models for Portfolio Construction	135		
Factor Models	137		
Principal Components Analysis and Factor Analysis	153		
Multifactor Equity Risk Models and Their Applications	171		
Factor-Based Equity Portfolio Construction and Analysis	195		
Cross-Sectional Factor-Based Models and Trading Strategies	213		
The Fundamentals of Fundamental Factor Models	243		
Multifactor Equity Risk Models and Their Applications	255		
Multifactor Fixed Income Risk Models and Their Applications	267		
Financial Econometrics	293		
Scope and Methods of Financial Econometrics	295		
Regression Analysis: Theory and Estimation	305		
Categorical and Dummy Variables in Regression Models	333		
Quantile Regression	353		
ARCH/GARCH Models in Applied Financial Econometrics	359		

Measurements of Prepayments for Residential Mortgage-Backed Securities	47	Back-Testing Market Risk Models	361
Prepayments and Factors Influencing the Return of Principal for Residential Mortgage-Backed Securities	65	Estimating Liquidity Risks	371
Operational Risk	79	Estimate of Downside Risk with Fat-Tailed and Skewed Models	381
Operational Risk	81	Moving Average Models for Volatility and Correlation, and Covariance Matrices	395
Operational Risk Models	91	Software for Financial Modeling	415
Modeling Operational Loss Distributions	103	Introduction to Financial Model Building with MATLAB	417
Optimization Tools	121	Introduction to Visual Basic for Applications	449
Introduction to Stochastic Programming and Its Applications to Finance	123	Stochastic Processes and Tools	469
Robust Portfolio Optimization	137	Stochastic Integrals	471
Probability Theory	149	Stochastic Differential Equations	485
Concepts of Probability Theory	151	Stochastic Processes in Continuous Time	495
Discrete Probability Distributions	165	Conditional Expectation and Change of Measure	507
Continuous Probability Distributions	195	Change of Time Methods	519
Continuous Probability Distributions with Appealing Statistical Properties	207	Term Structure Modeling	531
Continuous Probability Distributions Dealing with Extreme Events	227	The Concept and Measures of Interest Rate Volatility	533
Stable and Tempered Stable Distributions	241	Short-Rate Term Structure Models	543
Fat Tails, Scaling, and Stable Laws	259	Static Term Structure Modeling in Discrete and Continuous Time	559
Copulas	283	The Dynamic Term Structure Model	575
Applications of Order Statistics to Risk Management Problems	289	Essential Classes of Interest Rate Models and Their Use	593
Risk Measures	297	A Review of No Arbitrage Interest Rate Models	603
Measuring Interest Rate Risk: Effective Duration and Convexity	299	Trading Cost Models	621
Yield Curve Risk Measures	307	Modeling Market Impact Costs	623
Value-at-Risk	319	Volatility	635
Average Value-at-Risk	331	Monte Carlo Simulation in Finance	637
Risk Measures and Portfolio Selection	349	Stochastic Volatility	653

Contributors

Yves Achdou, PhD

Professor, Lab. Jacques-Louis Lions, University Paris-Diderot, Paris, France

Irene Aldridge

Managing Partner, Able Alpha Trading

Carol Alexander, PhD

Professor of Finance, University of Sussex

Andrew Alford, PhD

Managing Director, Quantitative Investment Strategies, Goldman Sachs Asset Management

Noël Amenc, PhD

Professor of Finance, EDHEC Business School, Director, EDHEC-Risk Institute

Bala Arshanapalli, PhD

Professor of Finance, Indiana University Northwest

David Audley, PhD

Senior Lecturer, The Johns Hopkins University

Jennifer Bender, PhD

Vice President, MSCI

William S. Berliner

Executive Vice President, Manhattan Advisory Services Inc.

Anand K. Bhattacharya, PhD

Professor of Finance Practice, Department of Finance, W. P. Carey School of Business, Arizona State University

Michele Leonardo Bianchi, PhD

Research Analyst, Specialized Intermediaries Supervision Department, Bank of Italy

Olivier Bokanowski

Associate Professor, Lab. Jacques-Louis Lions, University Paris-Diderot, Paris, France

Gerald W. Buetow Jr., PhD, CFA

President and Founder, BFRC Services, LLC

Paul Bukowski, CFA, FCAS

Executive President, Head of Equities, Hartford Investment Management

Joseph A. Cerniglia

Visiting Researcher, Courant Institute of Mathematical Sciences, New York University

Ren-Raw Chen

Professor of Finance, Graduate School of Business, Fordham University

Anna Chernobai, PhD

Assistant Professor of Finance, M. J. Whitman School of Management, Syracuse University

Richard Chin

Investment Manager, New York Life Investments

António Baldaque da Silva
Managing Director, Barclays

Siddhartha G. Dastidar, PhD, CFA
Vice President, Barclays

Arik Ben Dor, PhD
Managing Director, Barclays

Michael Dorigan, PhD
Senior Quantitative Analyst, PNC Capital
Advisors

Kevin Dowd, PhD
Partner, Cobden Partners, London

Pamela P. Drake, PhD, CFA
J. Gray Ferguson Professor of Finance, College
of Business, James Madison University

Lev Dynkin, PhD
Managing Director, Barclays

Brian Eales
Academic Leader (Retired), London Metropolitan
University

Abel Elizalde, PhD
Credit Derivatives Strategy, J.P. Morgan

Robert F. Engle, PhD
Michael Armellino Professorship in the Man-
agement of Financial Services and Director of
the Volatility Institute, Leonard N. Stern School
of Business, New York University

Frank J. Fabozzi, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

Peter Fitton
Manager, Scientific Development, CreditXpert
Inc.

Sergio M. Focardi, PhD
Partner, The Intertek Group

Radu Găbudean, PhD
Vice President, Barclays

Vacslav S. Glukhov, PhD
Head of Quantitative Strategies and Data Ana-
lytics, Liquidnet Europe Ltd, London, United
Kingdom

Felix Goltz, PhD
Head of Applied Research, EDHEC-Risk
Institute

Chris Gowlland, CFA
Senior Quantitative Analyst, Delaware Invest-
ments

Biliana S. Güner
Assistant Professor of Statistics and Economet-
rics, Özyeğin University, Turkey

Francis Gupta, PhD
Director, Index Research & Design, Dow Jones
Indexes

Markus Höchstötter, PhD
Assistant Professor, University of Karlsruhe

John S. J. Hsu, PhD
Professor of Statistics and Applied Probability,
University of California, Santa Barbara

Jay Hyman, PhD
Managing Director, Barclays, Tel Aviv

Bruce I. Jacobs, PhD
Principal, Jacobs Levy Equity Management

Robert R. Johnson, PhD, CFA
Independent Financial Consultant,
Charlottesville, VA

Frank J. Jones, PhD
Professor, Accounting and Finance Depart-
ment, San Jose State University and Chairman,
Investment Committee, Private Ocean Wealth
Management

Robert Jones, CFA
Chairman, Arwen Advisors, and Chairman and
CIO, Systems Two Advisors

Andrew Kalotay, PhD

President, Andrew Kalotay Associates

Young Shin Kim, PhD

Research Assistant Professor, School of Economics and Business Engineering, University of Karlsruhe and KIT

Petter N. Kolm, PhD

Director of the Mathematics in Finance Masters Program and Clinical Associate Professor, Courant Institute of Mathematical Sciences, New York University

Glen A. Larsen Jr., PhD CFA

Professor of Finance, Indiana University Kelley School of Business—Indianapolis

Anthony Lazanas

Managing Director, Barclays

Arturo Leccadito, PhD

Business Administration Department, Università della Calabria

Tony Lelièvre, PhD

Professor, CERMICS, Ecole des Ponts Paristech, Marne-la-Vallée, France

Alexander Levin, PhD

Director, Financial Engineering, Andrew Davidson & Co., Inc.

Kenneth N. Levy, CFA

Principal, Jacobs Levy Equity Management

Terence Lim, PhD, CFA

CEO, Arwen Advisors

Peter C. L. Lin

PhD Candidate, The Johns Hopkins University

Steven V. Mann, PhD

Professor of Finance, Moore School of Business, University of South Carolina

Harry M. Markowitz, PhD

Consultant and Nobel Prize Winner, Economics, 1990

Lionel Martellini, PhD

Professor of Finance, EDHEC Business School, Scientific Director, EDHEC-Risk Institute

James F. McNatt, CFA

Executive Vice President, ValueWealth Services

Christian Menn, Dr Rer Pol

Managing Partner, RIVACON

Ivan Mitov

Head of Quantitative Research, FinAnalytica

Edwin H. Neave

Professor Emeritus, School of Business, Queen's University, Kingston, Ontario

William Nelson, PhD

Professor of Finance, Indiana University Northwest

Frank Nielsen

Managing Director of Quantitative Research, Fidelity Investments - Global Asset Allocation

Philip O. Obazee

Senior Vice President and Head of Derivatives, Delaware Investments

Dominic O'Kane, PhD

Affiliated Professor of Finance, EDHEC Business School, Nice, France

Dessislava A. Pachamanova

Associate Professor of Operations Research, Babson College

Bruce D. Phelps

Managing Director, Barclays

Thomas K. Philips, PhD

Regional Head of Investment Risk and Performance, BNP Paribas Investment Partners

David Philpotts

QEP Global Equities, Schroder Investment Management, Sydney, Australia

Wesley Phoa

Senior Vice President, Capital International Research, Inc.

Svetlozar T. Rachev, PhD Dr Sci

Frey Family Foundation Chair Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

Boryana Racheva-Yotova, PhD

President, FinAnalytica

Shrikant Ramanmurthy

Consultant, New York, NY

Srichander Ramaswamy, PhD

Senior Economist, Bank for International Settlements, Basel, Switzerland

Patrice Retkowsky

Senior Research Engineer, EDHEC-Risk Institute

Paul Sclavounos

Department of Mechanical Engineering, Massachusetts Institute of Technology

Shani Shamah

Consultant, RBC Capital Markets

Koray D. Simsek, PhD

Associate Professor, Sabanci School of Management, Sabanci University

James Sochacki

Professor of Applied Mathematics, James Madison University

Arne D. Staal

Director, Barclays

Maxwell J. Stevenson, PhD

Discipline of Finance, Business School, University of Sydney, Australia

Filippo Stefanini

Head of Hedge Funds and Manager Selection, Eurizon Capital SGR

Stoyan V. Stoyanov, PhD

Professor of Finance at EDHEC Business School and Head of Research for EDHEC Risk Institute-Asia

Anatoliy Swishchuk, PhD

Professor of Mathematics and Statistics, University of Calgary

Ruey S. Tsay, PhD

H.G.B. Alexander Professor of Econometrics and Statistics, University of Chicago Booth School of Business

Radu S. Tunaru

Professor of Quantitative Finance, Business School, University of Kent

Cenk Ural, PhD

Vice President, Barclays

Donald R. Van Deventer, PhD

Chairman and Chief Executive Officer, Kamakura Corporation

Raman Vardharaj

Vice President, Oppenheimer Funds

Robert E. Whaley, PhD

Valere Blair Potter Professor of Management and Co-Director of the Financial Markets Research Center, Owen Graduate School of Management, Vanderbilt University

Mark B. Wickard

Senior Vice President/Corporate Cash
Investment Advisor, Morgan Stanley Smith
Bamey

James X. Xiong, PhD, CFA

Senior Research Consultant, Ibbotson
Associates, A Morningstar Company

Guofu Zhou

Frederick Bierman and James E. Spears Profes-
sor of Finance, Olin Business School, Washing-
ton University in St. Louis

Min Zhu

Business School, Queensland University of
Technology, Australia

Preface

It is often said that investment management is an art, not a science. However, since the early 1990s the market has witnessed a progressive shift toward a more industrial view of the investment management process. There are several reasons for this change. First, with globalization the universe of investable assets has grown many times over. Asset managers might have to choose from among several thousand possible investments from around the globe. Second, institutional investors, often together with their consultants, have encouraged asset management firms to adopt an increasingly structured process with documented steps and measurable results. Pressure from regulators and the media is another factor. Finally, the sheer size of the markets makes it imperative to adopt safe and repeatable methodologies.

In its modern sense, financial modeling is the design (or engineering) of financial instruments and portfolios of financial instruments that result in predetermined cash flows contingent upon different events. Broadly speaking, financial models are employed to manage investment portfolios and risk. The objective is the transfer of risk from one entity to another via appropriate financial arrangements. Though the aggregate risk is a quantity that cannot be altered, risk can be transferred if there is a willing counterparty.

Financial modeling came to the forefront of finance in the 1980s, with the broad diffusion

of derivative instruments. However, the concept and practice of financial modeling are quite old. The notion of the diversification of risk (central to modern risk management) and the quantification of insurance risk (a requisite for pricing insurance policies) were already understood, at least in practical terms, in the 14th century. The rich epistolary of Francesco Datini, a 14th-century merchant, banker, and insurer from Prato (Tuscany, Italy), contains detailed instructions to his agents on how to diversify risk and insure cargo.

What is specific to modern financial modeling is the quantitative management of risk. Both the pricing of contracts and the optimization of investments require some basic capabilities of statistical modeling of financial contingencies. It is the size, diversity, and efficiency of modern competitive markets that makes the use of financial modeling imperative.

This three-volume encyclopedia offers not only coverage of the fundamentals and advances in financial modeling but provides the mathematical and statistical techniques needed to develop and test financial models, as well as the practical issues associated with implementation. The encyclopedia offers the following unique features:

- The entries for the encyclopedia were written by experts from around the world. This diverse collection of expertise has created the most definitive coverage of established and

cutting-edge financial models, applications, and tools in this ever-evolving field.

- The series emphasizes both technical and managerial issues. This approach provides researchers, educators, students, and practitioners with a balanced understanding of the topics and the necessary background to deal with issues related to financial modeling.
- Each entry follows a format that includes the author, entry abstract, introduction, body, listing of key points, notes, and references. This enables readers to pick and choose among various sections of an entry, and creates consistency throughout the entire encyclopedia.
- The numerous illustrations and tables throughout the work highlight complex topics and assist further understanding.
- Each volume includes a complete table of contents and index for easy access to various parts of the encyclopedia.

TOPIC CATEGORIES

As is the practice in the creation of an encyclopedia, the topic categories are presented alphabetically. The topic categories and a brief description of each topic follow.

VOLUME I

Asset Allocation

A major activity in the investment management process is establishing policy guidelines to satisfy the investment objectives. Setting policy begins with the asset allocation decision. That is, a decision must be made as to how the funds to be invested should be distributed among the major asset classes (e.g., equities, fixed income, and alternative asset classes). The term “asset allocation” includes (1) policy asset allocation, (2) dynamic asset allocation, and (3) tactical asset allocation. Policy asset allocation decisions can loosely be characterized as long-term asset allocation decisions, in which the investor seeks to assess an appropriate long-term “normal” asset mix that represents an ideal blend of controlled risk and enhanced return. In dynamic asset allocation the asset mix (i.e., the

allocation among the asset classes) is mechanically shifted in response to changing market conditions. Once the policy asset allocation has been established, the investor can turn his or her attention to the possibility of active departures from the normal asset mix established by policy. If a decision to deviate from this mix is based upon rigorous objective measures of value, it is often called tactical asset allocation. The fundamental model used in establishing the policy asset allocation is the mean-variance portfolio model formulated by Harry Markowitz in 1952, popularly referred to as the theory of portfolio selection and modern portfolio theory.

Asset Pricing Models

Asset pricing models seek to formalize the relationship that should exist between asset returns and risk if investors behave in a hypothesized manner. At its most basic level, asset pricing is mainly about transforming asset payoffs into prices. The two most well-known asset pricing models are the arbitrage pricing theory and the capital asset pricing model. The fundamental theorem of asset pricing asserts the equivalence of three key issues in finance: (1) absence of arbitrage; (2) existence of a positive linear pricing rule; and (3) existence of an investor who prefers more to less and who has maximized his or her utility. There are two types of arbitrage opportunities. The first is paying nothing today and obtaining something in the future, and the second is obtaining something today and with no future obligations. Although the principle of absence of arbitrage is fundamental for understanding asset valuation in a competitive market, there are well-known limits to arbitrage resulting from restrictions imposed on rational traders, and, as a result, pricing inefficiencies may exist for a period of time.

Bayesian Analysis and Financial Modeling Applications

Financial models describe in mathematical terms the relationships between financial random variables through time and/or across assets. The fundamental assumption is that the

model relationship is valid independent of the time period or the asset class under consideration. Financial data contain both meaningful information and random noise. An adequate financial model not only extracts optimally the relevant information from the historical data but also performs well when tested with new data. The uncertainty brought about by the presence of data noise makes imperative the use of statistical analysis as part of the process of financial model building, model evaluation, and model testing. Statistical analysis is employed from the vantage point of either of the two main statistical philosophical traditions—frequentist and Bayesian. An important difference between the two lies with the interpretation of the concept of probability. As the name suggests, advocates of the frequentist approach interpret the probability of an event as the limit of its long-run relative frequency (i.e., the frequency with which it occurs as the amount of data increases without bound). Since the time financial models became a mainstream tool to aid in understanding financial markets and formulating investment strategies, the framework applied in finance has been the frequentist approach. However, strict adherence to this interpretation is not always possible in practice. When studying rare events, for instance, large samples of data may not be available, and in such cases proponents of frequentist statistics resort to theoretical results. The Bayesian view of the world is based on the subjectivist interpretation of probability: Probability is subjective, a degree of belief that is updated as information or data are acquired. Only in the last two decades has Bayesian statistics started to gain greater acceptance in financial modeling, despite its introduction about 250 years ago. It has been the advancements of computing power and the development of new computational methods that have fostered the growing use of Bayesian statistics in financial modeling.

Bond Valuation

The value of any financial asset is the present value of its expected future cash flows. To value

a bond (also referred to as a fixed-income security), one must be able to estimate the bond's remaining cash flows and identify the appropriate discount rate(s) at which to discount the cash flows. The traditional approach to bond valuation is to discount every cash flow with the same discount rate. Simply put, the relevant term structure of interest rate used in valuation is assumed to be flat. This approach, however, permits opportunities for arbitrage. Alternatively, the arbitrage-free valuation approach starts with the premise that a bond should be viewed as a portfolio or package of zero-coupon bonds. Moreover, each of the bond's cash flows is valued using a unique discount rate that depends on the term structure of interest rates and when in time the cash flow is. The relevant set of discount rates (that is, spot rates) is derived from an appropriate term structure of interest rates and when used to value risky bonds augmented with a suitable risk spread or premium. Rather than modeling to calculate the fair value of its price, the market price can be taken as given so as to compute a yield measure or a spread measure. Popular yield measures are the yield to maturity, yield to call, yield to put, and cash flow yield. Nominal spread, static (or zero-volatility) spread, and option-adjusted spread are popular relative value measures quoted in the bond market. Complications in bond valuation arise when a bond has one or more embedded options such as call, put, or conversion features. For bonds with embedded options, the financial modeling draws from options theory, more specifically, the use of the lattice model to value a bond with embedded options.

Credit Risk Modeling

Credit risk is a broad term used to refer to three types of risk: default risk, credit spread risk, and downgrade risk. Default risk is the risk that the counterparty to a transaction will fail to satisfy the terms of the obligation with respect to the timely payment of interest and repayment of the amount borrowed. The counterparty could be the issuer of a debt obligation or an entity on

the other side of a private transaction such as a derivative trade or a collateralized loan agreement (i.e., a repurchase agreement or a securities lending agreement). The default risk of a counterparty is often initially gauged by the credit rating assigned by one of the three rating companies—Standard & Poor’s, Moody’s Investors Service, and Fitch Ratings. Although default risk is the one that most market participants think of when reference is made to credit risk, even in the absence of default, investors are concerned about the decline in the market value of their portfolio bond holdings due to a change in credit spread or the price performance of their holdings relative to a bond index. This risk is due to an adverse change in credit spreads, referred to as credit spread risk, or when it is attributed solely to the downgrade of the credit rating of an entity, it is called downgrade risk. Financial modeling of credit risk is used (1) to measure, monitor, and control a portfolio’s credit risk, and (2) to price credit risky debt instruments. There are two general categories of credit risk models: structural models and reduced-form models. There is considerable debate as to which type of model is the best to employ.

Derivatives Valuation

A derivative instrument is a contract whose value depends on some underlying asset. The term “derivative” is used to describe this product because its value is derived from the value of the underlying asset. The underlying asset, simply referred to as the “underlying,” can be either a commodity, a financial instrument, or some reference entity such as an interest rate or stock index, leading to the classification of commodity derivatives and financial derivatives. Although there are close conceptual relations between derivative instruments and cash market instruments such as debt and equity, the two classes of instruments are used differently: Debt and equity are used primarily for raising funds from investors, while derivatives are primarily

used for dividing up and trading risks. Moreover, debt and equity are direct claims against a firm’s assets, while derivative instruments are usually claims on a third party. A derivative’s value depends on the value of the underlying, but the derivative instrument itself represents a claim on the “counterparty” to the trade. Derivatives instruments are classified in terms of their payoff characteristics: linear and nonlinear payoffs. The former, also referred to as symmetric payoff derivatives, includes forward, futures, and swap contracts while the latter include options. Basically, a linear payoff derivative is a risk-sharing arrangement between the counterparties since both are sharing the risk regarding the price of the underlying. In contrast, nonlinear payoff derivative instruments (also referred to as asymmetric payoff derivatives) are insurance arrangements because one party to the trade is willing to insure the counterparty of a minimum or maximum (depending on the contract) price. The amount received by the insuring party is referred to as the contract price or premium. Derivative instruments are used for controlling risk exposure with respect to the underlying. Hedging is a special case of risk control where a party seeks to eliminate the risk exposure. Derivative valuation or pricing is developed based on no-arbitrage price relations, relying on the assumption that two perfect substitutes must have the same price.

VOLUME II

Difference Equations and Differential Equations

The tools of linear difference equations and differential equations have found many applications in finance. A difference equation is an equation that involves differences between successive values of a function of a discrete variable. A function of such a variable is one that provides a rule for assigning values in sequences to it. The theory of linear difference equations covers three areas: solving difference equations, describing the behavior

of difference equations, and identifying the equilibrium (or critical value) and stability of difference equations. Linear difference equations are important in the context of dynamic econometric models. Stochastic models in finance are expressed as linear difference equations with random disturbances added. Understanding the behavior of solutions of linear difference equations helps develop intuition for the behavior of these models. In nontechnical terms, differential equations are equations that express a relationship between a function and one or more derivatives (or differentials) of that function. The relationship between difference equations and differential equations is that the latter are invaluable for modeling situations in finance where there is a continually changing value. The problem is that not all changes in value occur continuously. If the change in value occurs incrementally rather than continuously, then differential equations have their limitations. Instead, a financial modeler can use difference equations, which are recursively defined sequences. It would be difficult to overemphasize the importance of differential equations in financial modeling where they are used to express laws that govern the evolution of price probability distributions, the solution of economic variational problems (such as intertemporal optimization), and conditions for continuous hedging (such as in the Black-Scholes option pricing model). The two broad types of differential equations are ordinary differential equations and partial differential equations. The former are equations or systems of equations involving only one independent variable. Another way of saying this is that ordinary differential equations involve only total derivatives. Partial differential equations are differential equations or systems of equations involving partial derivatives. When one or more of the variables is a stochastic process, we have the case of stochastic differential equations and the solution is also a stochastic process. An assumption must be made about what is driving noise in a stochastic differential

equation. In most applications, it is assumed that the noise term follows a Gaussian random variable, although other types of random variables can be assumed.

Equity Models and Valuation

Traditional fundamental equity analysis involves the analysis of a company's operations for the purpose of assessing its economic prospects. The analysis begins with the financial statements of the company in order to investigate the earnings, cash flow, profitability, and debt burden. The fundamental analyst will look at the major product lines, the economic outlook for the products (including existing and potential competitors), and the industries in which the company operates. The result of this analysis will be the growth prospects of earnings. Based on the growth prospects of earnings, a fundamental analyst attempts to determine the fair value of the stock using one or more equity valuation models. The two most commonly used approaches for valuing a firm's equity are based on discounted cash flow and relative valuation models. The principal idea underlying discounted cash flow models is that what an investor pays for a share of stock should reflect what is expected to be received from it—return on the investor's investment. What an investor receives are cash dividends in the future. Therefore, the value of a share of stock should be equal to the present value of all the future cash flows an investor expects to receive from that share. To value stock, therefore, an investor must project future cash flows, which, in turn, means projecting future dividends. Popular discounted cash flow models include the basic dividend discount model, which assumes a constant dividend growth, and the multiple-phase models, which include the two-stage dividend growth model and the stochastic dividend discount models. Relative valuation methods use multiples or ratios—such as price/earnings, price/book, or price/free cash flow—to determine whether a stock is trading at higher or lower multiples than its peers.

There are two critical assumptions in using relative valuation: (1) the universe of firms selected to be included in the peer group are in fact comparable, and (2) the average multiple across the universe of firms can be treated as a reasonable approximation of “fair value” for those firms. This second assumption may be problematic during periods of market panic or euphoria. Managers of quantitative equity firms employ techniques that allow them to identify attractive stock candidates, focusing not on a single stock as is done with traditional fundamental analysis but rather on stock characteristics in order to explain why one stock outperforms another stock. They do so by statistically identifying a group of characteristics to create a quantitative selection model. In contrast to the traditional fundamental stock selection, quantitative equity managers create a repeatable process that utilizes the stock selection model to identify attractive stocks. Equity portfolio managers have used various statistical models for forecasting returns and risk. These models, referred to as predictive return models, make conditional forecasts of expected returns using the current information set. Predictive return models include regressive models, linear autoregressive models, dynamic factor models, and hidden-variable models.

Factor Models and Portfolio Construction

Quantitative asset managers typically employ multifactor risk models for the purpose of constructing and rebalancing portfolios and analyzing portfolio performance. A multifactor risk model, or simply factor model, attempts to estimate and characterize the risk of a portfolio, either relative to a benchmark such as a market index or in absolute value. The model allows the decomposition of risk factors into a systematic and an idiosyncratic component. The portfolio’s risk exposure to broad risk factors is captured by the systematic risk. For equity portfolios these are typically fundamental factors (e.g., market capitalization and value

vs. growth), technical (e.g., momentum), and industry/sector/country. For fixed-income portfolios, systematic risk captures a portfolio’s exposure to broad risk factors such as the term structure of interest rates, credit spreads, optionality (call and prepayment), credit, and sectors. The portfolio’s systematic risk depends not only on its exposure to these risk factors but also the volatility of the risk factors and how they correlate with each other. In contrast to systematic risk, idiosyncratic risk captures the uncertainty associated with news affecting the holdings of individual issuers in the portfolio. In equity portfolios, idiosyncratic risk can be easily diversified by reducing the importance of individual issuers in the portfolio. Because of the larger number of issuers in bond indexes, however, this is a difficult task. There are different types of factor models depending on the factors. Factors can be exogenous variables or abstract variables formed by portfolios. Exogenous factors (or known factors) can be identified from traditional fundamental analysis or from economic theory that suggests macroeconomic factors. Abstract factors, also called unidentified or latent factors, can be determined with the statistical tool of factor analysis or principal component analysis. The simplest type of factor models is where the factors are assumed to be known or observable, so that time-series data are those factors that can be used to estimate the model. The four most commonly used approaches for the evaluation of return premiums and risk characteristics to factors are portfolio sorts, factor models, factor portfolios, and information coefficients. Despite its use by quantitative asset managers, the basic building blocks of factor models used by model builders and by traditional fundamental analysts are the same: They both seek to identify the drivers of returns for the asset class being analyzed.

Financial Econometrics

Econometrics is the branch of economics that draws heavily on statistics for testing and

analyzing economic relationships. The economic equivalent of the laws of physics, econometrics represents the quantitative, mathematical laws of economics. Financial econometrics is the econometrics of financial markets. It is a quest for models that describe financial time series such as prices, returns, interest rates, financial ratios, defaults, and so on. Although there are similarities between financial econometric models and models of the physical sciences, there are two important differences. First, the physical sciences aim at finding immutable laws of nature; econometric models model the economy or financial markets—artifacts subject to change. Because the economy and financial markets are artifacts subject to change, econometric models are not unique representations valid throughout time; they must adapt to the changing environment. Second, while basic physical laws are expressed as differential equations, financial econometrics uses both continuous-time and discrete-time models.

Financial Modeling Principles

The origins of financial modeling can be traced back to the development of mathematical equilibrium at the end of the nineteenth century, followed in the beginning of the twentieth century with the introduction of sophisticated mathematical tools for dealing with the uncertainty of prices and returns. In the 1950s and 1960s, financial modelers had tools for dealing with probabilistic models for describing markets, the principles of contingent claims analysis, an optimization framework for portfolio selection based on mean and variance of asset returns, and an equilibrium model for pricing capital assets. The 1970s ushered in models for pricing contingent claims and a new model for pricing capital assets based on arbitrage pricing. Consequently, by the end of the 1970s, the frameworks for financial modeling were well known. It was the advancement of computing power and refinements of the theories to take into account real-world market imperfections and

conventions starting in the 1980s that facilitated implementation and broader acceptance of mathematical modeling of financial decisions. The diffusion of low-cost high-performance computers has allowed the broad use of numerical methods, the landscape of financial modeling. The importance of finding closed-form solutions and the consequent search for simple models has been dramatically reduced. Computationally intensive methods such as Monte Carlo simulations and the numerical solution of differential equations are now widely used. As a consequence, it has become feasible to represent prices and returns with relatively complex models. Nonnormal probability distributions have become commonplace in many sectors of financial modeling. It is fair to say that the key limitation of financial modeling is now the size of available data samples or training sets, not the computations; it is the data that limit the complexity of estimates. Mathematical modeling has also undergone major changes. Techniques such as equivalent martingale methods are being used in derivative pricing, and cointegration, the theory of fat-tailed processes, and state-space modeling (including ARCH/GARCH and stochastic volatility models) are being used in financial modeling.

Financial Statement Analysis

Much of the financial data that are used in constructing financial models for forecasting and valuation purposes draw from the financial statements that companies are required to provide to investors. The four basic financial statements are the balance sheet, the income statement, the statement of cash flows, and the statement of shareholders' equity. It is important to understand these data so that the information conveyed by them is interpreted properly in financial modeling. The financial statements are created using several assumptions that affect how to use and interpret the financial data. The analysis of financial statements involves the selection, evaluation, and

interpretation of financial data and other pertinent information to assist in evaluating the operating performance and financial condition of a company. The operating performance of a company is a measure of how well a company has used its resources—its assets, both tangible and intangible—to produce a return on its investment. The financial condition of a company is a measure of its ability to satisfy its obligations, such as the payment of interest on its debt in a timely manner. There are many tools available in the analysis of financial information. These tools include financial ratio analysis and cash flow analysis. Cash flows are essential ingredients in valuation. Therefore, understanding past and current cash flows may help in forecasting future cash flows and, hence, determine the value of the company. Moreover, understanding cash flow allows the assessment of the ability of a firm to maintain current dividends and its current capital expenditure policy without relying on external financing. Financial modelers must understand how to use these financial ratios and cash flow information in the most effective manner in building models.

Finite Mathematics and Basic Functions for Financial Modeling

The collection of mathematical tools that does not include calculus is often referred to as “finite mathematics.” This includes matrix algebra, probability theory, and statistical analysis. Ordinary algebra deals with operations such as addition and multiplication performed on individual numbers. In financial modeling, it is useful to consider operations performed on ordered arrays of numbers. Ordered arrays of numbers are called vectors and matrices while individual numbers are called scalars. Probability theory is the mathematical approach to formalize the uncertainty of events. Even though a decision maker may not know which one of the set of possible events may finally occur, with probability theory a decision maker has the means of providing each event with

a certain probability. Furthermore, it provides the decision maker with the axioms to compute the probability of a composed event in a unique way. The rather formal environment of probability theory translates in a reasonable manner to the problems related to risk and uncertainty in finance such as, for example, the future price of a financial asset. Today, investors may be aware of the price of a certain asset, but they cannot say for sure what value it might have tomorrow. To make a prudent decision, investors need to assess the possible scenarios for tomorrow’s price and assign to each scenario a probability of occurrence. Only then can investors reasonably determine whether the financial asset satisfies an investment objective included within a portfolio. Probability models are theoretical models of the occurrence of uncertain events. In contrast, statistics is about empirical data and can be broadly defined as a set of methods used to make inferences from a known sample to a larger population that is in general unknown. In finance, a particular important example is making inferences from the past (the known sample) to the future (the unknown population). There are important mathematical functions with which the financial modeler should be acquainted. These include the continuous function, the indicator function, the derivative of a function, the monotonic function, and the integral, as well as special functions such as the characteristic function of random variables and the factorial, the gamma, beta, and Bessel functions.

Liquidity and Trading Costs

In broad terms, liquidity refers to the ability to execute a trade or liquidate a position with little or no cost or inconvenience. Liquidity depends on the market where a financial instrument is traded, the type of position traded, and sometimes the size and trading strategy of an individual trade. Liquidity risks are those associated with the prospect of imperfect market liquidity and can relate to risk of loss or

risk to cash flows. There are two main aspects to liquidity risk measurement: the measurement of liquidity-adjusted measures of market risk and the measurement of liquidity risks per se. Market practitioners often assume that markets are liquid—that is, that they can liquidate or unwind positions at going market prices—usually taken to be the mean of bid and ask prices—without too much difficulty or cost. This assumption is very convenient and provides a justification for the practice of marking positions to market prices. However, it is often empirically questionable, and the failure to allow for liquidity can undermine the measurement of market risk. Because liquidity risk is a major risk factor in its own right, portfolio managers and traders will need to measure this risk in order to formulate effective portfolio and trading strategies. A considerable amount of work has been done in the equity market in estimating liquidity risk. Because transaction costs are incurred when buying or selling stocks, poorly executed trades can adversely impact portfolio returns and therefore relative performance. Transaction costs are classified as explicit costs such as brokerage and taxes, and implicit costs, which include market impact cost, price movement risk, and opportunity cost. Broadly speaking, market impact cost is the price that a trader has to pay for obtaining liquidity in the market and is a key component of trading costs that must be modeled so that effective trading programs for executing trades can be developed. Typical forecasting models for market impact costs are based on a statistical factor approach where the independent variables are trade-based factors or asset-based factors.

VOLUME III

Model Risk and Selection

Model risk is the risk of error in pricing or risk-forecasting models. In practice, model risk arises because (1) any model involves simpli-

fication and calibration, and both of these require subjective judgments that are prone to error, and/or (2) a model is used inappropriately. Although model risk cannot be avoided, there are many ways in which financial modelers can manage this risk. These include (1) recognizing model risk, (2) identifying, evaluating, and checking the model's key assumption, (3) selecting the simplest reasonable model, (4) resisting the temptation to ignore small discrepancies in results, (5) testing the model against known problems, (6) plotting results and employing nonparametric statistics, (7) back-testing and stress-testing the model, (8) estimating model risk quantitatively, and (9) reevaluating models periodically. In financial modeling, model selection requires a blend of theory, creativity, and machine learning. The machine-learning approach starts with a set of empirical data that the financial modeler wants to explain. Data are explained by a family of models that include an unbounded number of parameters and are able to fit data with arbitrary precision. There is a trade-off between model complexity and the size of the data sample. To implement this trade-off, ensuring that models have forecasting power, the fitting of sample data is constrained to avoid fitting noise. Constraints are embodied in criteria such as the Akaike information criterion or the Bayesian information criterion. Economic and financial data are generally scarce given the complexity of their patterns. This scarcity introduces uncertainty as regards statistical estimates obtained by the financial modeler. It means that the data might be compatible with many different models with the same level of statistical confidence. Methods of probabilistic decision theory can be used to deal with model risk due to uncertainty regarding the model's parameters. Probabilistic decision making starts from the Bayesian inference process and involves computer simulations in all realistic situations. Since a risk model is typically a combination of a probability distribution model and a risk measure, a critical assumption is the probability distribution assumed for

the random variable of interest. Too often, the Gaussian distribution is the model of choice. Empirical evidence supports the use of probability distributions that exhibit fat tails such as the Student's t distribution and its asymmetric version and the Pareto stable class of distributions and their tempered extensions. Extreme value theory offers another approach for risk modeling.

Mortgage-Backed Securities Analysis and Valuation

Mortgage-backed securities are fixed-income securities backed by a pool of mortgage loans. Residential mortgage-backed securities (RMBS) are backed by a pool of residential mortgage loans (one-to-four family dwellings). The RMBS market includes agency RMBS and nonagency RMBS. The former are securities issued by the Government National Mortgage Association (Ginnie Mae), Fannie Mae, and Freddie Mac. Agency RMBS include passthrough securities, collateralized mortgage obligations, and stripped mortgage-backed securities (interest-only and principal-only securities). The valuation of RMBS is complicated due to prepayment risk, a form of call risk. In contrast, nonagency RMBS are issued by private entities, have no implicit or explicit government guarantee, and therefore require one or more forms of credit enhancement in order to be assigned a credit rating. The analysis of nonagency RMBS must take into account both prepayment risk and credit risk. The most commonly used method for valuing RMBS is the Monte Carlo method, although other methods have garnered favor, in particular the decomposition method. The analysis of RMBS requires an understanding of the factors that impact prepayments.

Operational Risk

Operational risk has been regarded as a mere part of a financial institution's "other" risks. However, failures of major financial entities

have made regulators and investors aware of the importance of this risk. In general terms, operational risk is the risk of loss resulting from inadequate or failed internal processes, people, or systems or from external events. This risk encompasses legal risks, which includes, but is not limited to, exposure to fines, penalties, or punitive damages resulting from supervisory actions, as well as private settlements. Operational risk can be classified according to several principles: nature of the loss (internally inflicted or externally inflicted), direct losses or indirect losses, degree of expectancy (expected or unexpected), risk type, event type or loss type, and by the magnitude (or severity) of loss and the frequency of loss. Operational risk can be the cause of reputational risk, a risk that can occur when the market reaction to an operational loss event results in reduction in the market value of a financial institution that is greater than the amount of the initial loss. The two principal approaches in modeling operational loss distributions are the nonparametric approach and the parametric approach. It is important to employ a model that captures tail events, and for this reason in operational risk modeling, distributions that are characterized as light-tailed distributions should be used with caution. The models that have been proposed for assessing operational risk can be broadly classified into top-down models and bottom-up models. Top-down models quantify operational risk without attempting to identify the events or causes of losses. Bottom-up models quantify operational risk on a micro level, being based on identified internal events. The obstacle hindering the implementation of these models is the scarcity of available historical operational loss data.

Optimization Tools

Optimization is an area in applied mathematics that, most generally, deals with efficient algorithms for finding an optimal solution among a set of solutions that satisfy given constraints. Mathematical programming, a management

science tool that uses mathematical optimization models to assist in decision making, includes linear programming, integer programming, mixed-integer programming, nonlinear programming, stochastic programming, and goal programming. Unlike other mathematical tools that are available to decision makers such as statistical models (which tell the decision maker what occurred in the past), forecasting models (which tell the decision maker what might happen in the future), and simulation models (which tell the decision maker what will happen under different conditions), mathematical programming models allow the decision maker to identify the “best” solution. Markowitz’s mean-variance model for portfolio selection is an example of an application of one type of mathematical programming (quadratic programming). Traditional optimization modeling assumes that the inputs to the algorithms are certain, but there are also branches of optimization such as robust optimization that study the optimal decision under uncertainty about the parameters of the problem. Stochastic programming deals with both the uncertainty about the parameters and a multiperiod decision-making framework.

Probability Distributions

In financial models where the outcome of interest is a random variable, an assumption must be made about the random variable’s probability distribution. There are two types of probability distributions: discrete and continuous. Discrete probability distributions are needed whenever the random variable is to describe a quantity that can assume values from a countable set, either finite or infinite. A discrete probability distribution (or law) is quite intuitive in that it assigns certain values, positive probabilities, adding up to one, while any other value automatically has zero probability. Continuous probability distributions are needed when the random variable of interest can assume any value inside of one or more

intervals of real numbers such as, for example, any number greater than zero. Asset returns, for example, whether measured monthly, weekly, daily, or at an even higher frequency are commonly modeled as continuous random variables. In contrast to discrete probability distributions that assign positive probability to certain discrete values, continuous probability distributions assign zero probability to any single real number. Instead, only entire intervals of real numbers can have positive probability such as, for example, the event that some asset return is not negative. For each continuous probability distribution, this necessitates the so-called probability density, a function that determines how the entire probability mass of one is distributed. The density often serves as the proxy for the respective probability distribution. To model the behavior of certain financial assets in a stochastic environment, a financial modeler can usually resort to a variety of theoretical distributions. Most commonly, probability distributions are selected that are analytically well known. For example, the normal distribution (a continuous distribution)—also called the Gaussian distribution—is often the distribution of choice when asset returns are modeled. Or the exponential distribution is applied to characterize the randomness of the time between two successive defaults of firms in a bond portfolio. Many other distributions are related to them or built on them in a well-known manner. These distributions often display pleasant features such as stability under summation—meaning that the return of a portfolio of assets whose returns follow a certain distribution again follows the same distribution. However, one has to be careful using these distributions since their advantage of mathematical tractability is often outweighed by the fact that the stochastic behavior of the true asset returns is not well captured by these distributions. For example, although the normal distribution generally renders modeling easy because all moments of the distribution exist, it fails to reflect stylized facts commonly encountered in

asset returns—namely, the possibility of very extreme movements and skewness. To remedy this shortcoming, probability distributions accounting for such extreme price changes have become increasingly popular. Some of these distributions concentrate exclusively on the extreme values while others permit any real number, but in a way capable of reflecting market behavior. Consequently, a financial modeler has available a great selection of probability distributions to realistically reproduce asset price changes. Their common shortcoming is generally that they are mathematically difficult to handle.

Risk Measures

The standard assumption in financial models is that the distribution for the return on financial assets follows a normal (or Gaussian) distribution and therefore the standard deviation (or variance) is an appropriate measure of risk in the portfolio selection process. This is the risk measure that is used in the well-known Markowitz portfolio selection model (that is, mean-variance model), which is the foundation for modern portfolio theory. Mounting evidence since the early 1960s strongly suggests that return distributions do not follow a normal distribution, but instead exhibit heavy tails and, possibly, skewness. The “tails” of the distribution are where the extreme values occur, and these extreme values are more likely than would be predicted by the normal distribution. This means that between periods where the market exhibits relatively modest changes in prices and returns, there will be periods where there are changes that are much higher (that is, crashes and booms) than predicted by the normal distribution. This is of major concern to financial modelers in seeking to generate probability estimates for financial risk assessment. To more effectively implement portfolio selection, researchers have proposed alternative risk measures. These risk measures fall into

two disjointed categories: dispersion measures and safety-first measures. Dispersion measures include mean standard deviation, mean absolute deviation, mean absolute moment, index of dissimilarity, mean entropy, and mean colog. Safety-first risk measures include classical safety first, value-at-risk, average value-at-risk, expected tail loss, MiniMax, lower partial moment, downside risk, probability-weighted function of deviations below a specified target return, and power conditional value-at-risk. Despite these alternative risk measures, the most popular risk measure used in financial modeling is volatility as measured by the standard deviation. There are different types of volatility: historical, implied volatility, level-dependent volatility, local volatility, and stochastic volatility (e.g., jump-diffusion volatility). There are risk measures commonly used for bond portfolio management. These measures include duration, convexity, key rate duration, and spread duration.

Software for Financial Modeling

The development of financial models requires the modeler to be familiar with spreadsheets such as Microsoft Excel and/or a platform to implement concepts and algorithms such as the Palisade Decision Tools Suite and other Excel-based software (mostly @RISK1, Solver2, VBA3), and MATLAB. Financial modelers can choose one or the other, depending on their level of familiarity and comfort with spreadsheet programs and their add-ins versus programming environments such as MATLAB. Some tasks and implementations are easier in one environment than in the other. MATLAB is a modeling environment that allows for input and output processing, statistical analysis, simulation, and other types of model building for the purpose of analysis of a situation. MATLAB uses a number-array-oriented programming language, that is, a programming language in which vectors and matrices

are the basic data structures. Reliable built-in functions, a wide range of specialized toolboxes, easy interface with widespread software like Microsoft Excel, and beautiful graphing capabilities for data visualization make implementation with MATLAB efficient and useful for the financial modeler. Visual Basic for Applications (VBA) is a programming language environment that allows Microsoft Excel users to automate tasks, create their own functions, perform complex calculations, and interact with spreadsheets. VBA shares many of the same concepts as object-oriented programming languages. Despite some important limitations, VBA does add useful capabilities to spreadsheet modeling, and it is a good tool to know because Excel is the platform of choice for many finance professionals.

Stochastic Processes and Tools

Stochastic integration provides a coherent way to represent that instantaneous uncertainty (or volatility) cumulates over time. It is thus fundamental to the representation of financial processes such as interest rates, security prices, or cash flows. Stochastic integration operates on stochastic processes and produces random variables or other stochastic processes. Stochastic integration is a process defined on each path as the limit of a sum. However, these sums are different from the sums of the Riemann-Lebesgue integrals because the paths of stochastic processes are generally not of bounded variation. Stochastic integrals in the sense of Itô are defined through a process of approximation by (1) defining Brownian motion, which is the continuous limit of a random walk, (2) defining stochastic integrals for elementary functions as the sums of the products of the elementary functions multiplied by the increments of the Brownian motion, and (3) extending this definition to any function through approximating sequences. The major application of integration to financial modeling involves stochastic

integrals. An understanding of stochastic integrals is needed to understand an important tool in contingent claims valuation: stochastic differential equations. The dynamic of financial asset returns and prices can be expressed using a deterministic process if there is no uncertainty about its future behavior, or, with a stochastic process, in the more likely case when the value is uncertain. Stochastic processes in continuous time are the most used tool to explain the dynamic of financial assets returns and prices. They are the building blocks to construct financial models for portfolio optimization, derivatives pricing, and risk management. Continuous-time processes allow for more elegant theoretical modeling compared to discrete time models, and many results proven in probability theory can be applied to obtain a simple evaluation method.

Statistics

Probability models are theoretical models of the occurrence of uncertain events. In contrast, statistics is about empirical data and can be broadly defined as a set of methods used to make inferences from a known sample to a larger population that is in general unknown. In finance, a particular important example is making inferences from the past (the known sample) to the future (the unknown population). In statistics, probabilistic models are applied using data so as to estimate the parameters of these models. It is not assumed that all parameter values in the model are known. Instead, the data for the variables in the model to estimate the value of the parameters are used and then applied to test hypotheses or make inferences about their estimated values. In financial modeling, the statistical technique of regression models is the workhorse. However, because regression models are part of the field of financial econometrics, this topic is covered in that topic category. Understanding dependences or functional links between variables is a key theme in

financial modeling. In general terms, functional dependencies are represented by dynamic models. Many important models are linear models whose coefficients are correlation coefficients. In many instances in financial modeling, it is important to arrive at a quantitative measure of the strength of dependencies. The correlation coefficient provides such a measure. In many instances, however, the correlation coefficient might be misleading. In particular, there are cases of nonlinear dependencies that result in a zero correlation coefficient. From the point of view of financial modeling, this situation is particularly dangerous as it leads to substantially underestimated risk. Different measures of dependence have been proposed, in particular copula functions. The copula overcomes the drawbacks of the correlation as a measure of dependency by allowing for a more general measure than linear dependence, allowing for the modeling of dependence for extreme events, and being indifferent to continuously increasing transformations. Another essential tool in financial modeling, because it allows the incorporation of uncertainty in financial models and consideration of additional layers of complexity that are difficult to incorporate in analytical models, is Monte Carlo simulation. The main idea of Monte Carlo simulation is to represent the uncertainty in market variables through scenarios, and to evaluate parameters of interest that depend on these market variables in complex ways. The advantage of such an approach is that it can easily capture the dynamics of underlying processes and the otherwise complex effects of interactions among market variables. A substantial amount of research in recent years has been dedicated to making scenario generation more accurate and efficient, and a number of sophisticated computational techniques are now available to the financial modeler.

Term Structure Modeling

The arbitrage-free valuation approach to the valuation of option-free bonds, bonds with em-

bedded options, and option-type derivative instruments requires that a financial instrument be viewed as a package of zero-coupon bonds. Consequently, in financial modeling, it is essential to be able to discount each expected cash flow by the appropriate interest rate. That rate is referred to as the spot rate. The term structure of interest rates provides the relationship between spot rates and maturity. Because of its role in valuation of cash bonds and option-type derivatives, the estimation of the term structure of interest rates is of critical importance as an input into a financial model. In addition to its role in valuation modeling, term structure models are fundamental to expressing value, risk, and establishing relative value across the spectrum of instruments found in the various interest-rate or bond markets. The term structure is most often specified for a specific market such as the U.S. Treasury market, the bond market for double-A rated financial institutions, the interest rate market for LIBOR, and swaps. Static models of the term structure are characterizations that are devoted to relationships based on a given market and do not serve future scenarios where there is uncertainty. Standard static models include those known as the spot yield curve, discount function, par yield curve, and the implied forward curve. Instantiations of these models may be found in both a discrete- and continuous-time framework. An important consideration is establishing how these term structure models are constructed and how to transform one model into another. In modeling the behavior of interest rates, stochastic differential equations (SDEs) are commonly used. The SDEs used to model interest rates must capture the market properties of interest rates such as mean reversion and/or a volatility that depends on the level of interest rates. For a one-factor model, the SDE is used to model the behavior of the short-term rate, referred to as simply the "short rate." The addition of another factor (i.e., a two-factor model) involves extending the SDE to represent the behavior of the short rate and a long-term rate (i.e., long rate).

The entries can serve as material for a wide spectrum of courses, such as the following:

- Financial engineering
- Financial mathematics
- Financial econometrics
- Statistics with applications in finance
- Quantitative asset management
- Asset and derivative pricing
- Risk management

Frank J. Fabozzi
Editor, *Encyclopedia of Financial Models*

Guide to the *Encyclopedia of Financial Models*

The *Encyclopedia of Financial Models* provides comprehensive coverage of the field of financial modeling. This reference work consists of three separate volumes and 127 entries. Each entry provides coverage of the selected topic intended to inform a broad spectrum of readers ranging from finance professionals to academicians to students to fiduciaries. To derive the greatest possible benefit from the *Encyclopedia of Financial Models*, we have provided this guide. It explains how the information within the encyclopedia can be located.

ORGANIZATION

The *Encyclopedia of Financial Models* is organized to provide maximum ease of use for its readers.

Table of Contents

A complete table of contents for the entire encyclopedia appears in the front of each volume. This list of titles represents topics that have been carefully selected by the editor, Frank J. Fabozzi. The Preface includes a more detailed description of the volumes and the topic categories that the entries are grouped under.

Index

A Subject Index for the entire encyclopedia is located at the end of each volume. The sub-

jects in the index are listed alphabetically and indicate the volume and page number where information on this topic can be found.

Entries

Each entry in the *Encyclopedia of Financial Models* begins on a new page, so that the reader may quickly locate it. The author's name and affiliation are displayed at the beginning of the entry. All entries in the encyclopedia are organized according to a standard format, as follows:

- Title and author
- Abstract
- Introduction
- Body
- Key points
- Notes
- References

Abstract

The abstract for each entry gives an overview of the topic, but not necessarily the content of the entry. This is designed to put the topic in the context of the entire *Encyclopedia*, rather than give an overview of the specific entry content.

Introduction

The text of each entry begins with an introductory section that defines the topic under

discussion and summarizes the content. By reading this section, the reader gets a general idea about the content of a specific entry.

Body

The body of each entry explains the purpose, theory, and math behind each model.

Key Points

The key points section provides in bullet point format a review of the materials discussed in

each entry. It imparts to the reader the most important issues and concepts discussed.

Notes

The notes provide more detailed information and citations of further readings.

References

The references section lists the publications cited in the entry.

ENCYCLOPEDIA
OF
FINANCIAL MODELS

Volume I

Asset Allocation

Mean-Variance Model for Portfolio Selection

FRANK J. FABOZZI, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

HARRY M. MARKOWITZ, PhD
Consultant

PETTER N. KOLM, PhD
Director of the Mathematics in Finance M.S. Program and Clinical Associate Professor,
Courant Institute of Mathematical Sciences, New York University

FRANCIS GUPTA, PhD
Director, Index Research & Design, Dow Jones Indexes

Abstract: The theory of portfolio selection together with capital asset pricing theory provides the foundation and the building blocks for the management of portfolios. The goal of portfolio selection is the construction of portfolios that maximize expected returns consistent with individually acceptable levels of risk. Using both historical data and investor expectations of future returns, portfolio selection uses modeling techniques to quantify expected portfolio returns and acceptable levels of portfolio risk and provides methods to select an optimal portfolio.

The theory of portfolio selection presented in this entry, often referred to as *mean-variance portfolio analysis* or simply mean-variance analysis, is a normative theory. A normative theory is one that describes a standard or norm of behavior that investors should pursue in constructing a portfolio rather than a prediction concerning actual behavior.

Asset pricing theory goes on to formalize the relationship that should exist between asset returns and risk if investors behave in a hypothesized manner. In contrast to a normative

theory, asset pricing theory is a positive theory—a theory that hypothesizes how investors behave rather than how investors should behave. Based on that hypothesized behavior of investors, a model that provides the expected return (a key input for constructing portfolios based on mean-variance analysis) is derived and is called an asset pricing model.

Together, portfolio selection theory and asset pricing theory provide a framework to specify and measure investment risk and to develop relationships between expected asset return and

risk (and hence between risk and required return on an investment). However, it is critically important to understand that portfolio selection is a theory that is independent of any theories about asset pricing. The validity of portfolio selection theory does not rest on the validity of asset pricing theory.

It would not be an overstatement to say that modern portfolio theory has revolutionized the world of investment management. Allowing managers to quantify the investment risk and expected return of a portfolio has provided the scientific and objective complement to the subjective art of investment management. More importantly, whereas at one time the focus of portfolio management used to be the risk of individual assets, the theory of portfolio selection has shifted the focus to the risk of the entire portfolio. This theory shows that it is possible to combine risky assets and produce a portfolio whose expected return reflects its components, but with considerably lower risk. In other words, it is possible to construct a portfolio whose risk is smaller than the sum of all its individual parts!

Though practitioners realized that the risks of individual assets were related, before modern portfolio theory, they were unable to formalize how combining these assets into a portfolio impacted the risk at the entire portfolio level, or how the addition of a new asset would change the return–risk characteristics of the portfolio. This is because practitioners were unable to quantify the returns and risks of their investments. Furthermore, in the context of the entire portfolio, they were also unable to formalize the interaction of the returns and risks across asset classes and individual assets. The failure to quantify these important measures and formalize these important relationships made the goal of constructing an optimal portfolio highly subjective and provided no insight into the return investors could expect and the risk they were undertaking. The other drawback before the advent of the theory of portfolio selection and asset pricing theory was that there was no mea-

surement tool available to investors for judging the performance of their investment managers.

SOME BASIC CONCEPTS

Portfolio theory draws on concepts from two fields: financial economic theory and probability and statistical theory. This section presents the concepts from financial economic theory used in portfolio theory. While many of the concepts presented here have a more technical or rigorous definition, the purpose is to keep the explanations simple and intuitive so that the importance and contribution of these concepts to the development of modern portfolio theory can be appreciated.

Utility Function and Indifference Curves

There are many situations where entities (i.e., individuals and firms) face two or more choices. The economic “theory of choice” uses the concept of a utility function to describe the way entities make decisions when faced with a set of choices. A utility function assigns a (numeric) value to all possible choices faced by the entity. The higher the value of a particular choice, the greater the utility derived from that choice. The choice that is selected is the one that results in the maximum utility given a set of constraints faced by the entity.

In portfolio theory too, entities are faced with a set of choices. Different portfolios have different levels of expected return and risk. Typically, the higher the level of expected return, the larger the risk. Entities are faced with the decision of choosing a portfolio from the set of all possible risk–return combinations, where when they like return, they dislike risk. Therefore, entities obtain different levels of utility from different risk–return combinations. The utility obtained from any possible risk–return combination is expressed by the utility function. Put simply, the utility function expresses the

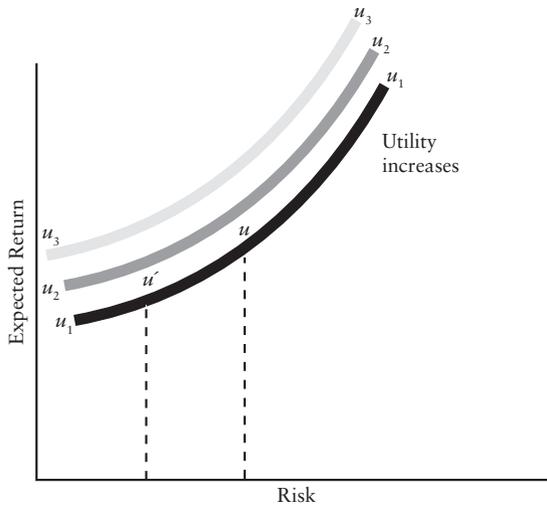


Figure 1 Indifference Curves

preferences of entities over perceived risk and expected return combinations.

A utility function can be expressed in graphical form by a set of indifference curves. Figure 1 shows indifference curves labeled u_1 , u_2 , and u_3 . By convention, the horizontal axis measures risk and the vertical axis measures expected return. Each curve represents a set of portfolios with different combinations of risk and return. All the points on a given indifference curve indicate combinations of risk and expected return that will give the same level of utility to a given investor. For example, on utility curve u_1 , there are two points u and u' , with u having a higher expected return than u' , but also having a higher risk. Because the two points lie on the same indifference curve, the investor has an equal preference for (or is indifferent to) the two points, or, for that matter, any point on the curve. The (positive) slope of an indifference curve reflects the fact that, to obtain the same level of utility, the investor requires a higher expected return in order to accept higher risk.

For the three indifference curves shown in Figure 1, the utility the investor receives is greater the further the indifference curve is from the horizontal axis because that curve represents a higher level of return at every level

of risk. Thus, for the three indifference curves shown in the figure, u_3 has the highest utility and u_1 the lowest.

The Set of Efficient Portfolios and the Optimal Portfolio

Portfolios that provide the largest possible expected return for given levels of risk are called *efficient portfolios*. To construct an efficient portfolio, it is necessary to make some assumption about how investors behave when making investment decisions. One reasonable assumption is that investors are risk averse. A risk-averse investor is an investor who, when faced with choosing between two investments with the same expected return but two different risks, prefers the one with the lower risk.

In selecting portfolios, an investor seeks to maximize the expected portfolio return given his tolerance for risk. (Alternatively stated, an investor seeks to minimize the risk that he is exposed to given some target expected return.) Given a choice from the set of efficient portfolios, an *optimal portfolio* is the one that is most preferred by the investor.

Risky Assets vs. Risk-Free Assets

A risky asset is one for which the return that will be realized in the future is uncertain. For example, an investor who purchases the stock of Pfizer Corporation today with the intention of holding it for some finite time does not know what return will be realized at the end of the holding period. The return will depend on the price of Pfizer's stock at the time of sale and on the dividends that the company pays during the holding period. Thus, Pfizer stock, and indeed the stock of all companies, is a risky asset.

Securities issued by the U.S. government are also risky. For example, an investor who purchases a U.S. government bond that matures in 30 years does not know the return that will be realized if this bond is held for only one year. This is because changes in interest rates in that year will affect the price of the bond one year

from now and that will impact the return on the bond over that year.

There are assets, however, for which the return that will be realized in the future is known with certainty today. Such assets are referred to as risk-free or riskless assets. The risk-free asset is commonly defined as a short-term obligation of the U.S. government. For example, if an investor buys a U.S. government security that matures in one year and plans to hold that security for one year, then there is no uncertainty about the return that will be realized. The investor knows that in one year, the maturity date of the security, the government will pay a specific amount to retire the debt. Notice how this situation differs for the U.S. government security that matures in 30 years. While the 1-year and the 30-year securities are obligations of the U.S. government, the former matures in one year so that there is no uncertainty about the return that will be realized. In contrast, while the investor knows what the government will pay at the end of 30 years for the 30-year bond, he does not know what the price of the bond will be one year from now.

MEASURING A PORTFOLIO'S EXPECTED RETURN

We are now ready to define the actual and expected return of a risky asset and a portfolio of risky assets.

Measuring Single-Period Portfolio Return

The actual return on a portfolio of assets over some specific time period is straightforward to calculate using the formula:

$$R_p = w_1 R_1 + w_2 R_2 + \dots + w_G R_G \quad (1)$$

where

R_p = rate of return on the portfolio over the period

R_g = rate of return on asset g over the period

w_g = weight of asset g in the portfolio (i.e., market value of asset g as a proportion of the market value of the total portfolio) at the beginning of the period

G = number of assets in the portfolio

In shorthand notation, equation (1) can be expressed as follows:

$$R_p = \sum_{g=1}^G w_g R_g \quad (2)$$

Equation (2) states that the return on a portfolio (R_p) of G assets is equal to the sum over all individual assets' weights in the portfolio times their respective return. The portfolio return R_p is sometimes called the holding period return or the ex post return.

For example, consider the following portfolio consisting of three assets:

Asset	Market Value at the Beginning of Holding Period	Rate of Return over Holding Period
1	\$6 million	12%
2	\$8 million	10%
3	\$11 million	5%

The portfolio's total market value at the beginning of the holding period is \$25 million. Therefore,

$$w_1 = \$6 \text{ million} / \$25 \text{ million} = 0.24, \text{ or } 24\% \text{ and } R_1 = 12\%$$

$$w_2 = \$8 \text{ million} / \$25 \text{ million} = 0.32, \text{ or } 32\% \text{ and } R_2 = 10\%$$

$$w_3 = \$11 \text{ million} / \$25 \text{ million} = 0.44, \text{ or } 44\% \text{ and } R_3 = 5\%$$

Notice that the sum of the weights is equal to 1. Substituting into equation (1), we get the holding period portfolio return,

$$R_p = 0.24(12\%) + 0.32(10\%) + 0.44(5\%) = 8.28\%$$

The Expected Return of a Portfolio of Risky Assets

Equation (1) shows how to calculate the actual return of a portfolio over some specific time period. In portfolio management, the investor also

wants to know the expected (or anticipated) return from a portfolio of risky assets. The expected portfolio return is the weighted average of the expected return of each asset in the portfolio. The weight assigned to the expected return of each asset is the percentage of the market value of the asset to the total market value of the portfolio. That is,

$$E(R_p) = w_1 E(R_1) + w_2 E(R_2) + \dots + w_G E(R_G) \quad (3)$$

The $E()$ signifies expectations, and $E(R_p)$ is sometimes called the *ex ante* return, or the expected portfolio return over some specific time period.

The expected return, $E(R_i)$, on a risky asset i is calculated as follows. First, a probability distribution for the possible rates of return that can be realized must be specified. A probability distribution is a function that assigns a probability of occurrence to all possible outcomes for a random variable. Given the probability distribution, the expected value of a random variable is simply the weighted average of the possible outcomes, where the weight is the probability associated with the possible outcome.

In our case, the random variable is the uncertain return of asset i . Having specified a probability distribution for the possible rates of return, the expected value of the rate of return for asset i is the weighted average of the possible outcomes. Finally, rather than use the term “expected value of the return of an asset,” we simply use the term “expected return.” Mathematically, the expected return of asset i is expressed as

$$E(R_i) = p_1 R_1 + p_2 R_2 + \dots + p_N R_N \quad (4)$$

where

R_n = the n th possible rate of return for asset i

p_n = the probability of attaining the rate of return R_n for asset i

N = the number of possible outcomes for the rate of return

How do we specify the probability distribution of returns for an asset? We shall see later

Table 1 Probability Distribution for the Rate of Return for Stock XYZ

n	Rate of Return	Probability of Occurrence
1	12%	0.18
2	10%	0.24
3	8%	0.29
4	4%	0.16
5	-4%	0.13
Total		1.00

on in this entry that in most cases the probability distribution of returns is based on long-run historical returns. If there is no reason to believe that future long-run returns should differ significantly from historical long-run returns, then probabilities assigned to different return outcomes based on the historical long-run performance of an uncertain investment could be a reasonable estimate for the probability distribution. However, for the purpose of illustration, assume that an investor is considering an investment, stock XYZ, which has a probability distribution for the rate of return for some time period as given in Table 1. The stock has five possible rates of return and the probability distribution specifies the likelihood of occurrence (in a probabilistic sense) of each of the possible outcomes.

Substituting into equation (4) we get

$$\begin{aligned} E(R_{XYZ}) &= 0.18(12\%) + 0.24(10\%) + 0.29(8\%) \\ &\quad + 0.16(4\%) + 0.13(-4\%) \\ &= 7\% \end{aligned}$$

Thus, 7% is the expected return or mean of the probability distribution for the rate of return on stock XYZ.

MEASURING PORTFOLIO RISK

Investors have used a variety of definitions to describe risk. Markowitz (1952, 1959) quantified the concept of risk using the well-known statistical measure: the standard deviation and the variance. The former is the intuitive concept. For most probability density functions, about

95% of the outcomes fall in the range defined by two standard deviations above and below the mean. Variance is defined as the square of the standard deviation. Computations are simplest in terms of variance. Therefore, it is convenient to compute the variance of a portfolio and then take its square root to obtain standard deviation.

Variance and Standard Deviation as a Measure of Risk

The variance of a random variable is a measure of the dispersion or variability of the possible outcomes around the expected value (mean). In the case of an asset's return, the variance is a measure of the dispersion of the possible rate of return outcomes around the expected return.

The equation for the variance of the expected return for asset i , denoted $\text{var}(R_i)$, is

$$\text{var}(R_i) = p_1[r_1 - E(R_i)]^2 + p_2[r_2 - E(R_i)]^2 + \dots + p_N[r_N - E(R_i)]^2$$

or

$$\text{var}(R_i) = \sum_{n=1}^N p_n[r_n - E(R_i)]^2 \quad (5)$$

Using the probability distribution of the return for stock XYZ, we can illustrate the calculation of the variance:

$$\begin{aligned} \text{var}(R_{XYZ}) &= 0.18(12\% - 7\%)^2 + 0.24(10\% - 7\%)^2 \\ &\quad + 0.29(8\% - 7\%)^2 + 0.16(4\% - 7\%)^2 \\ &\quad + 0.13(-4\% - 7\%)^2 = 24.1\% \end{aligned}$$

The variance associated with a distribution of returns measures the tightness with which the distribution is clustered around the mean or expected return. Markowitz argued that this tightness or variance is equivalent to the uncertainty or riskiness of the investment. If an asset is riskless, it has an expected return dispersion of zero. In other words, the return (which is also the expected return in this case) is certain, or guaranteed.

Since the variance is squared units, as we know from earlier in this section, it is common to see the variance converted to the standard deviation by taking the positive square root:

$$SD(R_i) = \sqrt{\text{Var}(R_i)}$$

For stock XYZ, then, the standard deviation is

$$SD(R_{XYZ}) = \sqrt{24.1\%} = 4.9\%$$

The variance and standard deviation are conceptually equivalent; that is, the larger the variance or standard deviation, the greater the investment risk. (A criticism of the variance or standard deviation as a measure is discussed later in this entry.)

Measuring the Portfolio Risk of a Two-Asset Portfolio

Equation (5) gives the variance for an individual asset's return. The variance of a portfolio consisting of two assets is a little more difficult to calculate. It depends not only on the variance of the two assets, but also upon how closely the returns of one asset track those of the other asset. The formula is

$$\begin{aligned} \text{var}(R_p) &= w_i^2 \text{var}(R_i) + w_j^2 \text{var}(R_j) \\ &\quad + 2w_i w_j \text{cov}(R_i, R_j) \end{aligned} \quad (6)$$

where

$\text{cov}(R_i, R_j)$ = covariance between the return for assets i and j

In words, equation (6) states that the variance of the portfolio return is the sum of the squared weighted variances of the two assets plus two times the weighted covariance between the two assets. We will see that this equation can be generalized to the case where there are more than two assets in the portfolio.

Covariance

The covariance has a precise mathematical translation. Its practical meaning is the degree to which the returns of two assets covary or change together. The covariance is not expressed in a particular unit, such as dollars or percent. A positive covariance means the returns on two assets tend to move or change in

Table 2 Probability Distribution for the Rate of Return for Asset XYZ and Asset ABC

<i>n</i>	Rate of Return for Asset XYZ	Rate of Return for Asset ABC	Probability of Occurrence
1	12%	21%	0.18
2	10%	14%	0.24
3	8%	9%	0.29
4	4%	4%	0.16
5	-4%	-3%	0.13
Total			1.00
Expected return	7.0%	10.0%	
Variance	24.1%	53.6%	
Standard deviation	4.9%	7.3%	

the same direction, while a negative covariance means the returns tend to move in opposite directions. The covariance between any two assets *i* and *j* is computed using the following formula:

$$\begin{aligned} \text{cov}(R_i, R_j) = & p_1[r_{i1} - E(R_i)][r_{j1} - E(R_j)] \\ & + p_2[r_{i2} - E(R_i)][r_{j2} - E(R_j)] + \dots \\ & + p_N[r_{iN} - E(R_i)][r_{jN} - E(R_j)] \end{aligned} \quad (7)$$

where

- r_{in} = the *n*th possible rate of return for asset *i*
- r_{jn} = the *n*th possible rate of return for asset *j*
- p_n = the probability of attaining the rate of return r_{in} and r_{jn} for assets *i* and *j*
- N = the number of possible outcomes for the rate of return

To illustrate the calculation of the covariance between two assets, we use the two stocks in Table 2. The first is stock XYZ from Table 1 that we used earlier to illustrate the calculation of the expected return and the standard deviation. The other hypothetical stock is stock ABC, whose data are shown in Table 2. Substituting the data for the two stocks from Table 2 in equation (7), the covariance between stocks XYZ and ABC is calculated as follows:

$$\begin{aligned} \text{cov}(R_{XYZ}, R_{ABC}) &= 0.18(12\% - 7\%)(21\% - 10\%) \\ &+ 0.24(10\% - 7\%)(14\% - 10\%) + 0.29(8\% \\ &- 7\%)(9\% - 10\%) + 0.16(4\% - 7\%)(4\% - 10\%) \\ &+ 0.13(-4\% - 7\%)(-3\% - 10\%) = 0.3396\% \end{aligned}$$

Relationship between Covariance and Correlation

The correlation is related to the covariance between the expected returns for two assets. Specifically, the correlation between the returns for assets *i* and *j* is defined as the covariance of the two assets divided by the product of their standard deviations:

$$\text{cor}(R_i, R_j) = \text{cov}(R_i, R_j) / [SD(R_i)SD(R_j)] \quad (8)$$

Dividing the covariance between the returns of two assets by the product of their standard deviations results in the correlation between the returns of the two assets. Because the correlation is a standardized number (i.e., it has been corrected for differences in the standard deviation of the returns), the correlation is comparable across different assets. The correlation between the returns for stock XYZ and stock ABC is

$$\text{cor}(R_{XYZ}, R_{ABC}) = 0.3396\% / (4.9\% \times 7.3\%) \approx 0.95$$

The correlation coefficient can have values ranging from +1.0, denoting perfect comovement in the same direction, to -1.0, denoting perfect comovement in the opposite direction. Also note that because the standard deviations are always positive, the correlation can only be negative if the covariance is a negative number. A correlation of zero implies that the returns are uncorrelated.

Measuring the Risk of a Portfolio Consisting of More than Two Assets

So far we have defined the risk of a portfolio consisting of two assets. The extension to three assets—*i*, *j*, and *k*—is as follows:

$$\begin{aligned} \text{var}(R_p) = & w_i^2 \text{var}(R_i) + w_j^2 \text{var}(R_j) + w_k^2 \text{var}(R_k) \\ & + 2w_i w_j \text{cov}(R_i, R_j) + 2w_i w_k \text{cov}(R_i, R_k) \\ & + 2w_j w_k \text{cov}(R_j, R_k) \end{aligned} \quad (9)$$

In words, equation (9) states that the variance of the portfolio return is the sum of the squared weighted variances of the individual assets plus two times the sum of the weighted pairwise

covariances of the assets. In general, for a portfolio with G assets, the portfolio variance is given by

$$\begin{aligned} \text{var}(R_p) = & \sum_{g=1}^G w_g^2 \text{var}(R_g) \\ & + \sum_{\substack{g=1 \\ \text{and} \\ h \neq g}}^G \sum_{h=1}^G w_g w_h \text{cov}(R_g, R_h) \end{aligned} \quad (10)$$

PORTFOLIO DIVERSIFICATION

Often, one hears investors talking about diversifying their portfolio. By this an investor means constructing a portfolio in such a way as to reduce portfolio risk without sacrificing return. This is certainly a goal that investors should seek. However, the question is how to do this in practice.

Some investors would say that including assets across all asset classes could diversify a portfolio. For example, a investor might argue that a portfolio should be diversified by investing in stocks, bonds, and real estate. While that might be reasonable, two questions must be addressed in order to construct a diversified portfolio. First, how much should be invested in each asset class? Should 40% of the portfolio be in stocks, 50% in bonds, and 10% in real estate, or is some other allocation more appropriate? Second, given the allocation, which specific stocks, bonds, and real estate should the investor select?

Some investors who focus only on one asset class such as common stock argue that such portfolios should also be diversified. By this they mean that an investor should not place all funds in the stock of one corporation, but rather should include stocks of many corporations. Here, too, several questions must be answered in order to construct a diversified portfolio. First, which corporations should be represented in the portfolio? Second, how much

of the portfolio should be allocated to the stocks of each corporation?

Prior to the development of portfolio theory, while investors often talked about diversification in these general terms, they did not possess the analytical tools by which to answer the questions posed above. For example, in 1945, Leavens (1945, p. 473) wrote:

An examination of some fifty books and articles on investment that have appeared during the last quarter of a century shows that most of them refer to the desirability of diversification. The majority, however, discuss it in general terms and do not clearly indicate why it is desirable.

Leavens illustrated the benefits of diversification on the assumption that risks are independent. However, in the last paragraph of his article, he cautioned:

The assumption, mentioned earlier, that each security is acted upon by independent causes, is important, although it cannot always be fully met in practice. Diversification among companies in one industry cannot protect against unfavorable factors that may affect the whole industry; additional diversification among industries is needed for that purpose. Nor can diversification among industries protect against cyclical factors that may depress all industries at the same time.

A major contribution of the theory of portfolio selection is that using the concepts discussed above, a quantitative measure of the diversification of a portfolio is possible, and it is this measure that can be used to achieve the maximum diversification benefits.

The Markowitz diversification strategy is primarily concerned with the degree of covariance between asset returns in a portfolio. Indeed a key contribution of Markowitz diversification is the formulation of an asset's risk in terms of a portfolio of assets, rather than in isolation. Markowitz diversification seeks to combine assets in a portfolio with returns that are less than perfectly positively correlated, in an effort to lower portfolio risk (variance) without sacrificing return. It is the concern for maintaining return while lowering risk through an analysis of the covariance between asset returns that separates Markowitz diversification from a naive

approach to diversification and makes it more effective.

Markowitz diversification and the importance of asset correlations can be illustrated with a simple two-asset portfolio example. To do this, we first show the general relationship between the risk of a two-asset portfolio and the correlation of returns of the component assets. Then we look at the effects on portfolio risk of combining assets with different correlations.

Portfolio Risk and Correlation

In our two-asset portfolio, assume that asset C and asset D are available with expected returns and standard deviations as shown:

Asset	$E(R)$	$SD(R)$
Asset C	12%	30%
Asset D	18%	40%

If an equal 50% weighting is assigned to both stocks C and D, the expected portfolio return can be calculated as shown:

$$E(R_p) = 0.50(12\%) + 0.50(18\%) = 15\%$$

The variance of the return on the two-stock portfolio from equation (6), using decimal form rather than percentage form for the standard deviation inputs, is

$$\begin{aligned} \text{var}(R_p) &= w_C^2 \text{var}(R_C) + w_D^2 \text{var}(R_D) \\ &\quad + 2w_C w_D \text{cov}(R_C, R_D) \\ &= (0.5)^2(0.30)^2 + (0.5)^2(0.40)^2 \\ &\quad + 2(0.5)(0.5) \text{cov}(R_C, R_D) \end{aligned}$$

From equation (8),

$$\text{cor}(R_C, R_D) = \text{cov}(R_C, R_D) / [SD(R_C)SD(R_D)]$$

so

$$\text{cov}(R_C, R_D) = SD(R_C)SD(R_D)\text{cor}(R_C, R_D)$$

Since $SD(R_C) = 0.30$ and $SD(R_D) = 0.40$, then

$$\text{cov}(R_C, R_D) = (0.30)(0.40) \text{cor}(R_C, R_D)$$

Substituting into the expression for $\text{var}(R_p)$, we get

$$\begin{aligned} \text{var}(R_p) &= (0.5)^2(0.30)^2 + (0.5)^2(0.40)^2 \\ &\quad + 2(0.5)(0.5)(0.30)(0.40)\text{cor}(R_C, R_D) \end{aligned}$$

Taking the square root of the variance gives

$$\begin{aligned} SD(R_p) &= \sqrt{(0.5)^2(0.30)^2 + (0.5)^2(0.40)^2 \\ &\quad + 2(0.5)(0.5)(0.30)(0.40)\text{cor}(R_C, R_D)} \\ &= \sqrt{0.0625 + (0.06)\text{cor}(R_C + R_D)} \end{aligned} \tag{11}$$

The Effect of the Correlation of Asset Returns on Portfolio Risk

How would the risk change for our two-asset portfolio with different correlations between the returns of the component stocks? Let's consider the following three cases for $\text{cor}(R_C, R_D)$: +1.0, 0, and -1.0. Substituting into equation (11) for these three cases of $\text{cor}(R_C, R_D)$, we get:

$\text{cor}(R_C, R_D)$	$E(R_p)$	$SD(R_p)$
+1.0	15%	35%
0.0	15%	25%
-1.0	15%	5%

As the correlation between the expected returns on stocks C and D decreases from +1.0 to 0.0 to -1.0, the standard deviation of the expected portfolio return also decreases from 35% to 5%. However, the expected portfolio return remains 15% for each case.

This example clearly illustrates the effect of Markowitz diversification. The principle of Markowitz diversification states that as the correlation (covariance) between the returns for assets that are combined in a portfolio decreases, so does the variance (hence the standard deviation) of the return for the portfolio.

The good news is that investors can maintain expected portfolio return and lower portfolio risk by combining assets with lower (and preferably negative) correlations. However, the bad news is that very few assets have small

Table 3 Portfolio Expected Returns and Standard Deviations for Five Mixes of Assets C and D
 Asset C: $E(R_C) = 12\%$, $SD(R_C) = 30\%$
 Asset D: $E(R_D) = 18\%$, and $SD(R_D) = 40\%$
 Correlation between Assets C and D = $\text{cor}(R_C, R_D) = -0.5$

Portfolio	Proportion of Asset C	Proportion of Asset D	$E(R_p)$	$SD(R_p)$
1	100%	0%	12.0%	30.0%
2	75%	25%	13.5%	19.5%
3	50%	50%	15.0%	18.0%
4	25%	75%	16.5%	27.0%
5	0%	100%	18.0%	40.0%

to negative correlations with other assets! The problem, then, becomes one of searching among large numbers of assets in an effort to discover the portfolio with the minimum risk at a given level of expected return or, equivalently, the highest expected return at a given level of risk.

The stage is now set for a discussion of efficient portfolios and their construction.

CHOOSING A PORTFOLIO OF RISKY ASSETS

Diversification in the manner suggested by Markowitz leads to the construction of portfolios that have the highest expected return for a given level of risk. Such portfolios are called *efficient portfolios*.

Constructing Efficient Portfolios

The technique of constructing efficient portfolios from large groups of stocks requires a massive number of calculations. In a portfolio of G securities, there are $(G^2 - G)/2$ unique covariances to estimate. Hence, for a portfolio of just 50 securities, there are 1,225 covariances that must be calculated. For 100 securities, there are 4,950. Furthermore, in order to solve for the portfolio that minimizes risk for each level of return, a mathematical technique called quadratic programming must be used. A discussion of this technique is beyond the scope of this entry. However, it is possible to illustrate the general idea of the construction of efficient portfolios by

referring again to the simple two-asset portfolio consisting of assets C and D.

Recall that for two assets, C and D , $E(R_C) = 12\%$, $SD(R_C) = 30\%$, $E(R_D) = 18\%$, and $SD(R_D) = 40\%$. We now further assume that $\text{cor}(R_C, R_D) = -0.5$. Table 3 presents the expected portfolio return and standard deviation for five different portfolios made up of varying proportions of C and D .

Feasible and Efficient Portfolios

A feasible portfolio is any portfolio that an investor can construct given the assets available. The five portfolios presented in Table 3 are all feasible portfolios. The collection of all feasible portfolios is called the *feasible set of portfolios*. With only two assets, the feasible set of portfolios is graphed as a curve, which represents those combinations of risk and expected return that are attainable by constructing portfolios from all possible combinations of the two assets.

Figure 2 presents the feasible set of portfolios for all combinations of assets C and D. As mentioned earlier, the portfolio mixes listed in Table 3 belong to this set and are shown by the points 1 through 5, respectively. Starting from 1 and proceeding to 5, asset C goes from 100% to 0%, while asset D goes from 0% to 100%—therefore, all possible combinations of C and D lie between portfolios 1 and 5, or on the curve labeled 1–5. In the case of two assets, any risk–return combination not lying on this curve is not attainable since there is no mix of assets C and D that will result in that risk–return

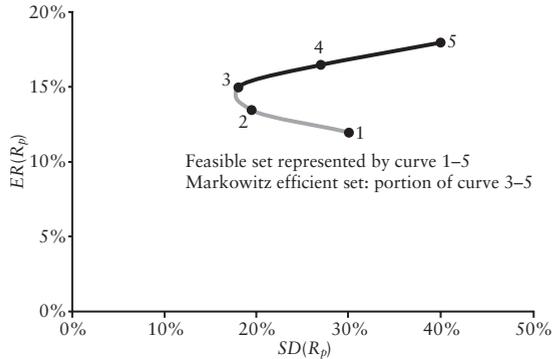


Figure 2 Feasible and Efficient Portfolios for Assets C and D

combination. Consequently, the curve 1–5 can also be thought of as the feasible set.

In contrast to a feasible portfolio, an *efficient portfolio* is one that gives the highest expected return of all feasible portfolios with the same risk. An efficient portfolio is also said to be a *mean-variance efficient portfolio*. Thus, for each level of risk there is an efficient portfolio. The collection of all efficient portfolios is called the *efficient set*.

The efficient set for the feasible set presented in Figure 2 is differentiated by the bold curve section 3–5. Efficient portfolios are the combinations of assets C and D that result in the risk–return combinations on the bold section of the curve. These portfolios offer the highest expected return at a given level of risk. Notice that two of our five portfolio mixes—portfolio 1 with $E(R_p) = 12\%$ and $SD(R_p) = 20\%$ and portfolio 2 with $E(R_p) = 13.5\%$ and $SD(R_p) = 19.5\%$ —are not included in the efficient set. This is because there is at least one portfolio in the efficient set (for example, portfolio 3) that has a higher expected return and lower risk than both of them. We can also see that portfolio 4 has a higher expected return and lower risk than portfolio 1. In fact, the whole curve section 1–3 is not efficient. For any given risk–return combination on this curve section, there is a combination (on the curve section 3–5) that has the same risk and a higher return, or the same return and a lower risk, or both. In other words, for any

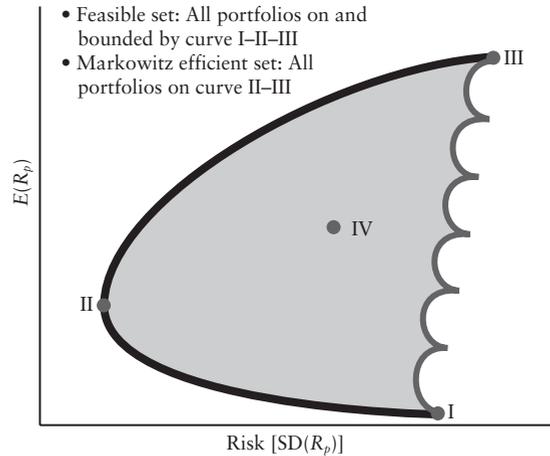


Figure 3 Feasible and Efficient Portfolios with More Than Two Assets^a

^aThe picture is for illustrative purposes only. The actual shape of the feasible region depends on the returns and risks of the assets chosen and the correlation among them.

portfolio that results in the return/risk combination on the curve section 1–3 (excluding portfolio 3), there exists a portfolio that dominates it by having the same return and lower risk, or the same risk and a higher return, or a lower risk and a higher return. For example, portfolio 4 dominates portfolio 1, and portfolio 3 dominates both portfolios 1 and 2.

Figure 3 shows the feasible and efficient sets when there are more than two assets. In this case, the feasible set is not a line, but an area. This is because, unlike the two-asset case, it is possible to create asset portfolios that result in risk–return combinations that not only result in combinations that lie on the curve I–II–III, but all combinations that lie in the shaded area. However, the efficient set is given by the curve II–III. It is easily seen that all the portfolios on the efficient set dominate the portfolios in the shaded area.

The efficient set of portfolios is sometimes called the *efficient frontier* because graphically all the efficient portfolios lie on the boundary of the set of feasible portfolios that have the maximum return for a given level of risk. Any risk–return combination above the efficient frontier cannot

be achieved, while risk–return combinations of the portfolios that make up the efficient frontier dominate those that lie below the efficient frontier.

Choosing the Optimal Portfolio in the Efficient Set

Now that we have constructed the efficient set of portfolios, the next step is to determine the optimal portfolio.

Since all portfolios on the efficient frontier provide the greatest possible return at their level of risk, an investor or entity will want to hold one of the portfolios on the efficient frontier. Notice that the portfolios on the efficient frontier represent trade-offs in terms of risk and return. Moving from left to right on the efficient frontier, the risk increases, but so does the expected return. The question is which one of those portfolios should an investor hold? The best portfolio to hold of all those on the efficient frontier is the *optimal portfolio*.

Intuitively, the optimal portfolio should depend on the investor’s preference over different risk–return trade-offs. As explained earlier, this preference can be expressed in terms of a utility function.

In Figure 4, three indifference curves representing a utility function and the efficient frontier are drawn on the same diagram. An indifference curve indicates the combinations of risk and expected return that give the same level of utility. Moreover, the farther the indifference curve from the horizontal axis, the higher the utility.

From Figure 4, it is possible to determine the optimal portfolio for the investor with the indifference curves shown. Remember that the investor wants to get to the highest indifference curve achievable given the efficient frontier. Given that requirement, the optimal portfolio is represented by the point where an indifference curve is tangent to the efficient frontier. In Figure 4, that is the portfolio P_{MEF}^* . For example, suppose that P_{MEF}^* corresponds to portfolio 4

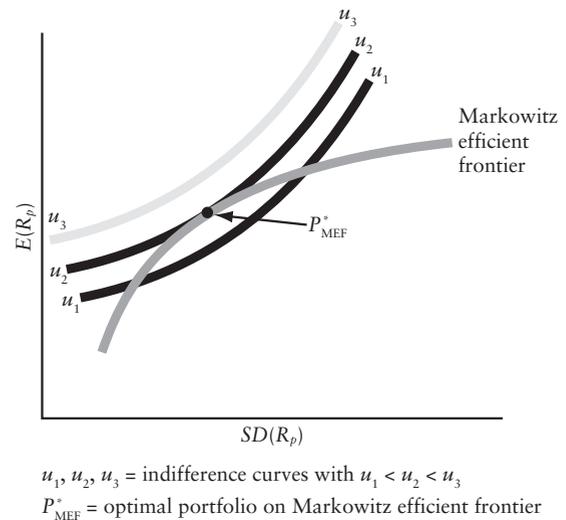


Figure 4 Selection of the Optimal Portfolio

in Figure 2. We know from Table 3 that this portfolio is made up of 25% of asset C and 75% of asset D, with an $E(R_p) = 16.5\%$ and $SD(R_p) = 27.0\%$.

Consequently, for the investor’s preferences over risk and return as determined by the shape of the indifference curves represented in Figure 4, and expectations for asset C and D inputs (returns and variance-covariance) represented in Table 3, portfolio 4 is the optimal portfolio because it maximizes the investor’s utility. If this investor had a different preference for expected risk and return, there would have been a different optimal portfolio.

At this point in our discussion, a natural question is how to estimate an investor’s utility function so that the indifference curves can be determined. Economists in the field of behavioral and experimental economics have conducted a vast amount of research in the area of utility functions. Though the assumption sounds reasonable that individuals should possess a function that maps the different preference choices they face, the research shows that it is not so straightforward to assign an individual with a specific utility function. This is because preferences may be dependent on circumstances, and those may change with time.

Table 4 Annualized Expected Returns, Standard Deviations, and Correlations between the Four Country Equity Indexes: Australia, Austria, Belgium, and Canada

Expected Returns	Standard Deviation	Correlations	1	2	3	4
7.9%	19.5%	Australia	1	1		
7.9%	18.2%	Austria	2	0.24	1	
9.0%	18.3%	Belgium	3	0.25	0.47	1
7.1%	16.5%	Canada	4	0.22	0.14	0.25
						1

The inability to assign an investor with a specific utility function does not imply that the theory is irrelevant. Once the efficient frontier is constructed, it is possible for the investor to subjectively evaluate the trade-offs for the different return–risk outcomes and choose the efficient portfolio that is appropriate given his or her tolerance to risk.

Example Using the MSCI World Country Indexes

Now that we know how to calculate the optimal portfolios and the efficient frontier, let us take a look at a practical example. We start the example using only four assets and later show these results change as more assets are included. The four assets are the four country equity indexes in the MSCI World Index for Australia, Austria, Belgium, and Canada.

Let us assume that we are given the annualized expected returns, standard deviations, and correlations between these countries as presented in Table 4. The expected returns vary from 7.1% to 9%, whereas the standard deviations range from 16.5% to 19.5%. Furthermore, we observe that the four country indexes are not highly correlated with each other—the highest correlation, 0.47, is between Austria and Belgium. Therefore, we expect to see some benefits of portfolio diversification.

Figure 5 shows the efficient frontier for the four assets. We observe that the four assets, represented by the diamond-shaped marks, are all below the efficient frontier. This means that for a targeted expected portfolio return, the mean-variance portfolio has a lower standard deviation. A utility maximizing investor, measuring

utility as the trade-off between expected return and standard deviation, will prefer a portfolio on the efficient frontier over any of the individual assets.

The portfolio at the leftmost end of the efficient frontier (marked with a solid circle in Figure 5) is the portfolio with the smallest obtainable standard deviation. It is called the *global minimum variance* (GMV) portfolio.

Increasing the Asset Universe

We know that by introducing more (low correlating) assets, for a targeted expected portfolio return, we should be able to decrease the standard deviation of the portfolio. In Table 5, the assumed annualized expected returns, stan-

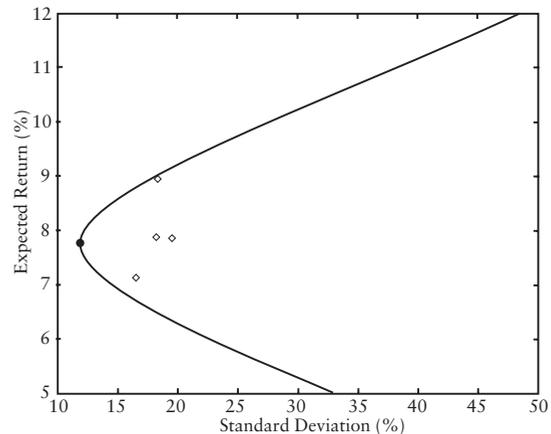


Figure 5 The Mean-Variance Efficient Frontier of Country Equity Indexes of Australia, Austria, Belgium, and Canada

Note: Constructed using the data in Table 4. The expected return and standard deviation combination of each country index is represented by a diamond-shaped mark. The global minimum variance portfolio (GMV) is represented by a solid circle. The portfolios on the curves above the GMV portfolio constitute the efficient frontier.

Table 5 Annualized Expected Returns, Standard Deviations, and Correlations between 18 Countries in the MSCI World Index

Expected Returns	Standard Deviation	Correlations	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
7.9%	19.5%	Australia	1																	
7.9%	18.2%	Austria	2	0.24	1															
9.0%	18.3%	Belgium	3	0.25	0.47	1														
7.1%	16.5%	Canada	4	0.22	0.14	0.25	1													
12.0%	18.4%	Denmark	5	0.24	0.44	0.48	0.21	1												
10.3%	20.4%	France	6	0.22	0.41	0.56	0.35	0.45	1											
9.5%	21.8%	Germany	7	0.26	0.48	0.57	0.35	0.48	0.65	1										
12.0%	28.9%	Hong Kong	8	0.31	0.17	0.17	0.19	0.18	0.22	0.24	1									
11.6%	23.3%	Italy	9	0.20	0.36	0.42	0.22	0.38	0.47	0.47	0.16	1								
9.5%	22.1%	Japan	10	0.32	0.28	0.28	0.18	0.28	0.27	0.29	0.24	0.21	1							
10.9%	19.7%	Netherlands	11	0.26	0.38	0.57	0.39	0.45	0.67	0.67	0.24	0.44	0.28	1						
7.9%	22.7%	Norway	12	0.33	0.37	0.41	0.27	0.41	0.45	0.47	0.21	0.32	0.28	0.50	1					
7.6%	21.5%	Singapore	13	0.34	0.22	0.23	0.20	0.22	0.22	0.26	0.44	0.19	0.34	0.24	0.28	1				
9.9%	20.8%	Spain	14	0.26	0.42	0.50	0.27	0.43	0.57	0.54	0.20	0.48	0.25	0.51	0.39	0.25	1			
16.2%	23.5%	Sweden	15	0.27	0.34	0.42	0.31	0.42	0.53	0.53	0.23	0.41	0.27	0.51	0.43	0.27	0.49	1		
10.7%	17.9%	Switzerland	16	0.26	0.47	0.59	0.32	0.49	0.64	0.69	0.23	0.45	0.32	0.67	0.48	0.25	0.53	0.51	1	
9.8%	18.5%	United Kingdom	17	0.25	0.34	0.47	0.38	0.40	0.58	0.53	0.22	0.40	0.28	0.68	0.43	0.24	0.46	0.45	0.57	1
10.5%	16.5%	United States	18	0.05	0.05	0.21	0.62	0.11	0.29	0.29	0.13	0.17	0.08	0.32	0.15	0.12	0.22	0.26	0.31	1

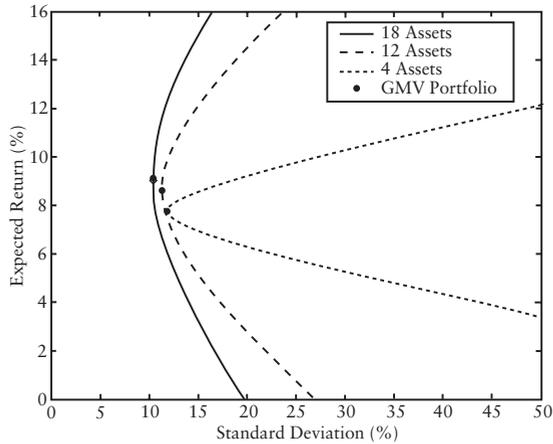


Figure 6 The Efficient Frontier Widens as the Number of Low Correlated Assets Increase
Note: The efficient frontiers have been constructed with 4, 12, and 18 countries (from the innermost to the outermost frontier) from the MSCI World Index. The portfolios on the curves above the GMV portfolio constitute the efficient frontiers for the three cases.

standard deviations, and correlations of 18 countries in the MSCI World Index are presented.

Figure 6 illustrates how the efficient frontier moves outwards and upwards as we go from 4 to 12 assets and then to 18 assets. By increasing the number of investment opportunities, we increase the level of possible diversification thereby making it possible to generate a higher level of return at each level of risk.

Adding Short Selling Constraints

So far in this section, our theoretical derivations imposed no restrictions on the portfolio weights other than having them add up to one. In particular, we allowed the portfolio weights to take on both positive and negative values; that is, we did not restrict short selling. In practice, many portfolio managers cannot sell assets short. This could be for investment policy or legal reasons, or sometimes just because particular asset classes are difficult to sell short such real estate. In Figure 7, we see the effect of not allowing for short selling. Since we are restricting the opportunity set by constraining all the weights to be

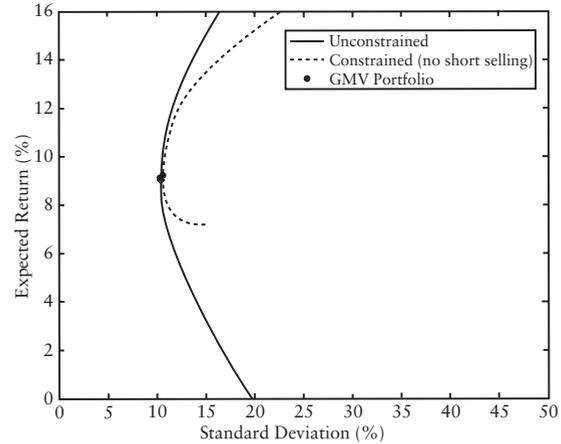


Figure 7 The Effect of Restricting Short Selling: Constrained versus Unconstrained Efficient Frontiers Constructed from 18 Countries from the MSCI World Index
Note: The portfolios on the curves above the GMV portfolio constitute the efficient frontiers.

positive, the resulting efficient frontier is inside the unconstrained efficient frontier.

ROBUST PORTFOLIO OPTIMIZATION

Despite the great influence and theoretical impact of modern portfolio theory, today full risk–return optimization at the asset level is primarily done only at the more quantitatively oriented asset management firms. The availability of quantitative tools is not the issue—today’s optimization technology is mature and much more user-friendly than it was at the time Markowitz first proposed the theory of portfolio selection—yet many asset managers avoid using the quantitative portfolio allocation framework altogether.

A major reason for the reluctance of portfolio managers to apply quantitative risk–return optimization is that they have observed that it may be unreliable in practice. Specifically, mean–variance optimization (or any measure of risk for that matter) is very sensitive to changes

in the inputs (in the case of mean-variance optimization, such inputs include the expected return and variance of each asset and the asset covariance between each pair of assets). While it can be difficult to make accurate estimates of these inputs, estimation errors in the forecasts significantly impact the resulting portfolio weights. As a result, the optimal portfolios generated by the mean-variance analysis generally have extreme or counterintuitive weights for some assets.¹ Such examples, however, are not necessarily a sign that the theory of portfolio selection is flawed; rather, that when used in practice, the mean-variance analysis as presented by Markowitz has to be modified in order to achieve reliability, stability, and robustness with respect to model and estimation errors.

It goes without saying that advances in the mathematical and physical sciences have had a major impact upon finance. In particular, mathematical areas such as probability theory, statistics, econometrics, operations research, and mathematical analysis have provided the necessary tools and discipline for the development of modern financial economics. Substantial advances in the areas of robust estimation and robust optimization were made during the 1990s, and have proven to be of great importance for the practical applicability and reliability of portfolio management and optimization.

Any statistical estimate is subject to error—that is, estimation error. A robust estimator is a statistical estimation technique that is less sensitive to outliers in the data and is not driven by one particular set of observations of the data. For example, in practice, it is undesirable that one or a few extreme returns have a large impact on the estimation of the average return of a stock. Nowadays, statistical techniques such as Bayesian analysis and robust statistics are more commonplace in asset management. Taking it one step further, practitioners are starting to incorporate the uncertainty introduced by estimation errors directly into the optimization process. This is very different from traditional mean-variance analysis, where one solves the

portfolio optimization problem as a problem with deterministic inputs (i.e., inputs that are assumed to be known with certainty), without taking the estimation errors into account. In particular, the statistical precision of individual estimates is explicitly incorporated into the portfolio allocation process. Providing this benefit is the underlying goal of robust portfolio optimization.²

Modern robust optimization techniques allow a portfolio manager to solve the robust version of the portfolio optimization problem in about the same time as needed for the traditional mean-variance portfolio optimization problem. The robust approach explicitly uses the distribution from the estimation process to find a robust portfolio in a single optimization, thereby directly incorporating uncertainty about inputs in the optimization process. As a result, robust portfolios are less sensitive to estimation errors than other portfolios, and often perform better than optimal portfolios determined by traditional mean-variance portfolios. Moreover, the robust optimization framework offers greater flexibility and many new interesting applications. For instance, robust portfolio optimization can exploit the notion of statistically equivalent portfolios. This concept is important in large-scale portfolio management involving many complex constraints such as transaction costs, turnover, or market impact. Specifically, with robust optimization, a portfolio manager can find the best portfolio that (1) minimizes trading costs with respect to the current holdings and (2) has an expected portfolio return and variance that are statistically equivalent to those of the classical mean-variance portfolio.³

KEY POINTS

- Markowitz quantified the concept of diversification through the statistical notion of the covariances between individual securities that make up a portfolio and the overall standard deviation of the portfolio.

- A basic assumption behind modern portfolio theory is that an investor's preferences over portfolios with different expected returns and variances can be represented by a function (utility function).
- The basic principle underlying modern portfolio theory is that for a given level of expected return an investor would choose the portfolio with the minimum variance from among the set of all possible portfolios.
- Minimum variance portfolios are called mean-variance efficient portfolios. The set of all mean-variance efficient portfolios is called the efficient frontier. The portfolio on the efficient frontier with the smallest variance is called the global minimum variance portfolio (GMVP).
- The efficient frontier moves outwards and upwards as the number of (not perfectly correlated) securities increases. The efficient frontier shrinks as constraints are imposed upon the portfolio.
- An advancement in the theory of portfolio selection is the development of estimation techniques that generate more robust mean-variance estimates along with optimization techniques that result in optimized portfolios being more robust to the mean-variance estimates used.

NOTES

1. See Best and Grauer (1991) and Chopra and Ziemba (1993).
2. There are two approaches that have been suggested for dealing with this problem. One is the application of estimation by using a statistical technique known as Bayes analysis. (See Rachev, Hsu, Bagasheva, and Fabozzi, 2008.) The Black-Litterman model uses this approach. (See Black and Litterman, 1990.) The other approach is using a resampling methodology as suggested by Michaud (2001). A study by Markowitz and Usmen (2003) found that the resampled approach is superior to that of a Bayesian approach.
3. For a discussion of these models, see Fabozzi, Kolm, Pachamanova, and Focardi (2007).

REFERENCES

- Best, M. J., and Grauer, R. R. (1991). On the sensitivity of mean-variance efficient portfolios to changes in asset means: Some analytical and computational results. *Review of Financial Studies* 4: 315–342.
- Black, R., and Litterman, R. (1990). Asset allocation: Combining investor views with market equilibrium. Goldman, Sachs & Co., *Fixed Income Research*.
- Chopra, V. K., and Ziemba, W. T. (1993). The effect of errors in means, variances, and covariances on optimal portfolio choice. *Journal of Portfolio Management* 19: 6–11.
- Fabozzi, F. J., Kolm, P. N., Pachamanova, D., and Focardi, S. M. (2007). *Robust Portfolio Optimization and Management*. Hoboken, NJ: John Wiley & Sons.
- Leavens, D. H. (1945). Diversification of investments. *Trusts and Estates* 80: 469–473.
- Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance* 7: 77–91.
- Markowitz, H. M. (1959). *Portfolio Selection: Efficient Diversification of Investments*, Cowles Foundation Monograph 16. New York: John Wiley & Sons, 1959.
- Markowitz, H. M., and Usmen, N. (2003). Diffuse priors vs. resampled frontiers: An experiment. *Journal of Investment Management* 1: 9–25.
- Michaud, R. (2001). *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation*. New York: Oxford University Press.
- Rachev, S. T., Hsu, J., Bagasheva, B., and Fabozzi, F. J. (2008). *Bayesian Methods in Finance*. Hoboken, NJ: John Wiley & Sons.

Principles of Optimization for Portfolio Selection

STOYAN V. STOYANOV, PhD

Professor of Finance at EDHEC Business School and Head of Research for EDHEC Risk Institute-Asia

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair-Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: The mathematical theory of optimization has a natural application in the field of finance. From a general perspective, the behavior of economic agents in the face of uncertainty involves balancing expected risks and expected rewards. For example, the portfolio choice problem concerns the optimal trade-off between risk and reward. A portfolio is said to be optimal in the sense that it is the best portfolio among many alternative ones. The criterion that measures the quality of a portfolio relative to the others is known as the objective function in optimization theory. The set of portfolios among which we are choosing is called the “set of feasible solutions” or the “set of feasible points.”

In *optimization* theory there is a distinction between two types of optimization problems depending on whether the set of feasible solutions is constrained or unconstrained. If the optimization problem is a constrained one, then the *set of feasible solutions* is defined by means of certain linear and/or nonlinear equalities and inequalities. These functions are often said to be forming the *constraint set*. Furthermore, a distinction is made between the types of optimization problems depending on the

assumed properties of the *objective function* and the functions in the constraint set, such as *linear problems*, *quadratic problems*, and *convex problems*. The solution methods vary with respect to the particular optimization problem type as there are efficient algorithms prepared for particular problem types.

In this chapter, we describe the basic types of optimization problems and remark on the methods for their solution. Boyd and Vandenberghe (2004) and Ruszczyński (2006)

provide more detailed information on the topic.

UNCONSTRAINED OPTIMIZATION

When there are no constraints imposed on the set of feasible solutions, we have an *unconstrained optimization* problem. Thus, the goal is to maximize or to minimize the objective function with respect to the function arguments without any limits on their values. We consider directly the n -dimensional case; that is, the domain of the objective function f is the n -dimensional space and the function values are real numbers, $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Maximization is denoted by

$$\max f(x_1, \dots, x_n)$$

and minimization by

$$\min f(x_1, \dots, x_n)$$

A more compact form is commonly used; for example

$$\min_{x \in \mathbb{R}^n} f(x) \quad (1)$$

denotes that we are searching for the minimal value of the function $f(x)$ by varying x in the entire n -dimensional space \mathbb{R}^n . A solution to problem (1) is a value of $x = x^0$ for which the minimum of f is attained:

$$f_0 = f(x^0) = \min_{x \in \mathbb{R}^n} f(x)$$

Thus, the vector x_0 is such that the function takes a larger value than f_0 for any other vector x ,

$$f(x^0) \leq f(x), x \in \mathbb{R}^n \quad (2)$$

Note that there may be more than one vector x^0 satisfying the inequality in (2) and, therefore, the argument for which f_0 is achieved may not be unique. If (2) holds, then the function is said to attain its global minimum at x^0 . If the inequality in (2) holds for x belonging only to a small neighborhood of x^0 and not to the entire space \mathbb{R}^n , then the objective function is said to have a

local minimum at x^0 . This is usually denoted by

$$f(x^0) \leq f(x)$$

for all x such that $\|x - x^0\|_2 < \epsilon$ where $\|x - x^0\|_2$ stands for the Euclidean distance between the vectors x and x^0 ,

$$\|x - x^0\|_2 = \sqrt{\sum_{i=1}^n (x_i - x_i^0)^2}$$

and ϵ is some positive number. A local minimum may not be global as there may be vectors outside the small neighborhood of x_0 for which the objective function attains a smaller value than $f(x_0)$. Figure 3 shows the graph of a function with two local maxima, one of which is the global maximum.

There is a connection between minimization and maximization. Maximizing the objective function is the same as minimizing the negative of the objective function and then changing the sign of the minimal value:

$$\max_{x \in \mathbb{R}^n} f(x) = -\min_{x \in \mathbb{R}^n} [-f(x)]$$

This relationship is illustrated in Figure 1. As a consequence, problems for maximization can be stated in terms of function minimization and vice versa.

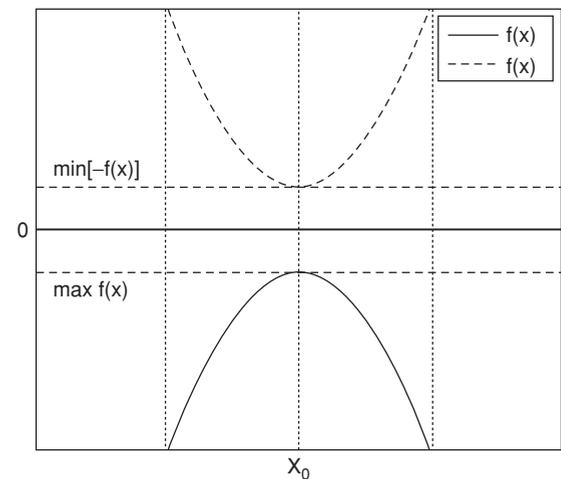


Figure 1 The Relationship between Minimization and Maximization for a One-Dimensional Function

Minima and Maxima of a Differentiable Function

If the second derivatives of the objective function exist, then its local maxima and minima, often called generically local extrema, can be characterized. Denote by $\nabla f(x)$ the vector of the first partial derivatives of the objective function evaluated at x ,

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \dots, \frac{\partial f(x)}{\partial x_n} \right)$$

This vector is called the function gradient. At each point x of the domain of the function, it shows the direction of greatest rate of increase of the function in a small neighborhood of x . If for a given x the gradient equals a vector of zeros,

$$\nabla f(x) = (0, \dots, 0)$$

then the function does not change in a small neighborhood of $x \in \mathbb{R}^n$. It turns out that all points of local extrema of the objective function are characterized by a zero gradient. As a result, the points yielding the local extrema of the objective function are among the solutions of the system of equations,

$$\begin{cases} \frac{\partial f(x)}{\partial x_1} = 0 \\ \dots \\ \frac{\partial f(x)}{\partial x_n} = 0 \end{cases} \quad (3)$$

The system of equations (3) is often referred to as representing the first-order condition for the objective function extrema. However, it is only a necessary condition; that is, if the gradient is zero at a given point in the n -dimensional space, then this point may or may not be a point of a local extremum for the function. An illustration is given in Figures 2 and 3. Figure 2 shows the graph of a two-dimensional function and Figure 3 contains the contour lines of the function with the gradient calculated at a grid of points. There are three points marked with a black dot which have a zero gradient. The middle point is not a point of a local maximum even though it has a zero gradient. This point is called a *sad-*

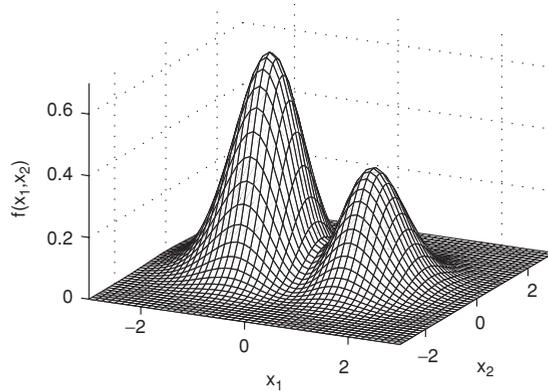


Figure 2 A Function $f(x_1, x_2)$ with Two Local Maxima

dle point, since the graph resembles the shape of a saddle in a neighborhood of it. The left and the right points are where the function has two local maxima corresponding to the two peaks visible on the top plot. The right peak is a local maximum which is not the global one and the left peak represents the global maximum.

This example demonstrates that the first-order conditions are generally insufficient to characterize the points of local extrema. The additional condition which identifies which of the

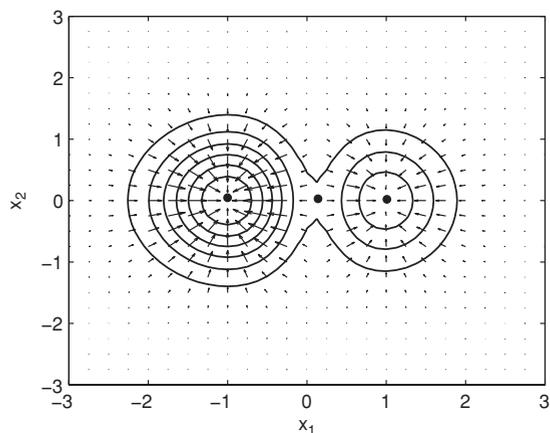


Figure 3 The Contour Lines of $f(x_1, x_2)$ Together with the Gradient Evaluated at a Grid of Points

Note: The middle black point shows the position of the saddle point between the two local maxima.

zero-gradient points are points of local minimum or maximum is given through the matrix of second derivatives,

$$H = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{pmatrix} \quad (4)$$

which is called the Hessian matrix or just the Hessian. The Hessian is a symmetric matrix because the order of differentiation is insignificant:

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}$$

The additional condition is known as the second-order condition. We will not provide the second-order condition for functions of n -dimensional arguments because it is rather technical and goes beyond the scope of the entry. We only state it for two-dimensional functions.

In the case $n = 2$, the following conditions hold:

- If $\nabla f(x_1, x_2) = (0, 0)$ at a given point (x_1, x_2) and the determinant of the Hessian matrix evaluated at (x_1, x_2) is positive, then the function has:

A local maximum in (x_1, x_2) if

$$\frac{\partial^2 f(x_1, x_2)}{\partial x_1^2} < 0 \quad \text{or} \quad \frac{\partial^2 f(x_1, x_2)}{\partial x_2^2} < 0$$

A local minimum in (x_1, x_2) if

$$\frac{\partial^2 f(x_1, x_2)}{\partial x_1^2} > 0 \quad \text{or} \quad \frac{\partial^2 f(x_1, x_2)}{\partial x_2^2} > 0$$

- If $\nabla f(x_1, x_2) = (0, 0)$ at a given point (x_1, x_2) and the determinant of the Hessian matrix evaluated at (x_1, x_2) is negative, then the function f has a saddle point in (x_1, x_2) .
- If $\nabla f(x_1, x_2) = (0, 0)$ at a given point (x_1, x_2) and the determinant of the Hessian matrix evaluated at (x_1, x_2) is zero, then no conclusion can be drawn.

Convex Functions

We just demonstrated that the first-order conditions are insufficient in the general case to describe the local extrema. However, when certain assumptions are made for the objective function, the first-order conditions can become sufficient. Furthermore, for certain classes of functions, the local extrema are necessarily global. Therefore, solving the first-order conditions we obtain the global extremum.

A general class of functions with nice *optimal* properties is the class of *convex functions*. Not only are the convex functions easy to optimize but also they have important application in risk management. (See Chapter 6 in Rachev, Stoyanov, and Fabozzi [2008] for a discussion of general measures of risk.) It turns out that the property which guarantees that diversification is possible appears to be exactly the convexity property. As a consequence, a measure of risk is necessarily a convex functional.

A *function* in mathematics can be viewed as a rule assigning to each element of a set D a single element of a set C . The set D is called the domain of f and the set C is called the codomain of f . A *functional* is a special kind of a function which takes other functions as its argument and returns numbers as output; that is, its domain is a set of functions. For example, the definite integral can be viewed as a functional because it assigns a real number to a function—the corresponding area below the function graph. A risk measure can also be viewed as a functional because it assigns a number to a random variable. Any random variable is mathematically described as a certain function the domain of which is a set of outcomes Ω .

Precisely, a function $f(x)$ is called a convex function if it satisfies the property: For a given $\alpha \in [0, 1]$ and all $x^1 \in \mathbb{R}^n$ and $x^2 \in \mathbb{R}^n$ in the function domain,

$$f(\alpha x^1 + (1 - \alpha)x^2) \leq \alpha f(x^1) + (1 - \alpha)f(x^2) \quad (5)$$

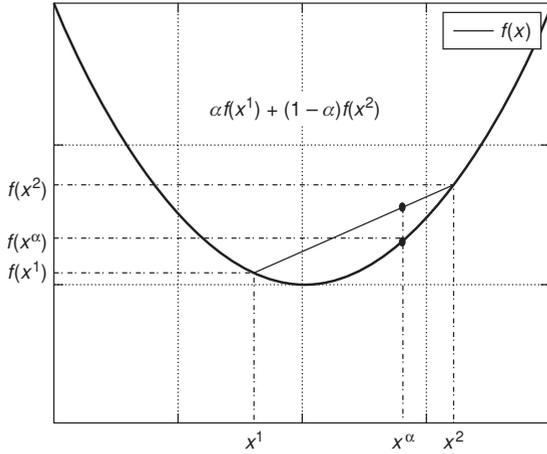


Figure 4 Illustration of the Definition of a Convex Function in the One-Dimensional Case
 Note: On the plot, $x_\alpha = \alpha x^1 + (1 - \alpha)x^2$.

The definition is illustrated in Figure 4. Basically, if a function is convex, then a straight line connecting any two points on the graph lies “above” the graph of the function.

There is a related term to convex functions. A function f is called *concave* if the negative of f is convex. In effect, a function is concave if it satisfies the property: For a given $\alpha \in [0, 1]$ and all $x^1 \in \mathbb{R}^n$ and $x^2 \in \mathbb{R}^n$ in the function domain,

$$f(\alpha x^1 + (1 - \alpha)x^2) \geq \alpha f(x^1) + (1 - \alpha)f(x^2)$$

If the domain D of a convex function is not the entire space \mathbb{R}^n , then the set D satisfies the property

$$\alpha x^1 + (1 - \alpha)x^2 \in D \quad (6)$$

where $x^1 \in D$, $x^2 \in D$, and $0 \leq \alpha \leq 1$. The sets which satisfy (6) are called convex sets. Thus, the domains of convex (and concave) functions should be convex sets. Geometrically, a set is convex if it contains the straight line connecting any two points belonging to the set. Rockafellar (1997) provides detailed information on the implications of convexity in optimization theory.

We summarize several important properties of convex functions:

- Not all convex functions are differentiable. If a convex function is two times continuously differentiable, then the corresponding Hessian defined in (4) is a positive semidefinite matrix. (A matrix H is a positive semidefinite matrix if $x'Hx \geq 0$ for all $x \in \mathbb{R}^n$ and $x \neq (0, \dots, 0)$.)
- All convex functions are continuous if considered in an open set.
- The sublevel sets

$$L_c = \{x : f(x) \leq c\} \quad (7)$$

where c is a constant, are convex sets if f is a convex function. The converse is not true in general. Later, we provide more information about nonconvex functions with convex sublevel sets.

- The local minima of a convex function are global. If a convex function f is twice continuously differentiable, then the global minimum is obtained in the points solving the first-order condition

$$\nabla f(x) = 0$$

- A sum of convex functions is a convex function:

$$f(x) = f_1(x) + f_2(x) + \dots + f_k(x)$$

if $f_i, i = 1, \dots, k$ are convex functions.

A simple example of a convex function is the linear function

$$f(x) = a'x, \quad x \in \mathbb{R}^n$$

where $a \in \mathbb{R}^n$ is a vector of constants. In fact, the linear function is the only function which is both convex and concave. In finance, if we consider a portfolio of assets, then the expected portfolio return is a linear function of portfolio weights, in which the coefficients equal the expected asset returns.

As a more involved example, consider the following function:

$$f(x) = \frac{1}{2}x'Cx, \quad x \in \mathbb{R}^n \quad (8)$$

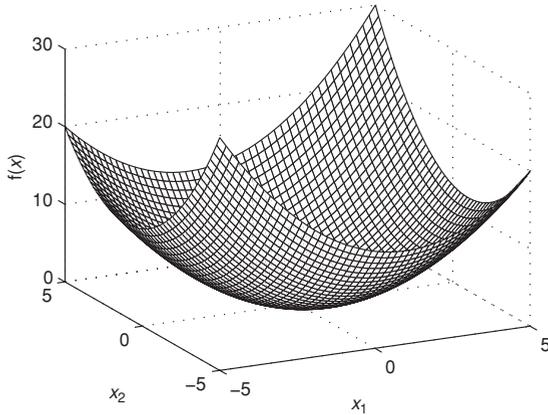


Figure 5 The Surface of a Two-Dimensional Convex Quadratic Function

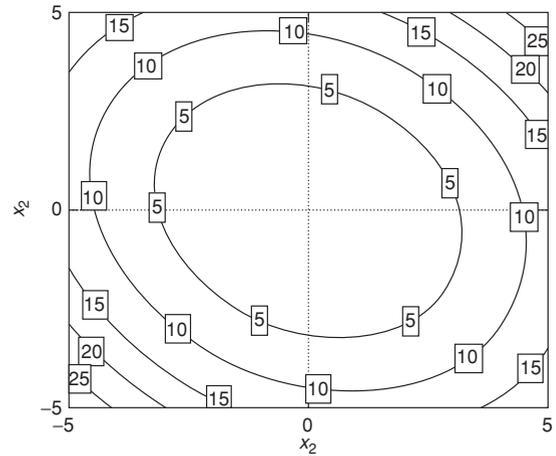


Figure 6 The Contour Lines of a Two-Dimensional Convex Quadratic Function

where $C = \{c_{ij}\}_{i,j=1}^n$ is an $n \times n$ symmetric matrix. In portfolio theory, the variance of portfolio return is a similar function of portfolio weights. In this case, C is the covariance matrix. The function defined in (8) is called a quadratic function because writing the definition in terms of the components of the argument X , we obtain

$$f(x) = \frac{1}{2} \left[\sum_{i=1}^n c_{ii} x_i^2 + \sum_{i \neq j} c_{ij} x_i x_j \right]$$

which is a quadratic function of the components $x_i, i = 1, \dots, n$. The function in (8) is convex if and only if the matrix C is positive semidefinite. In fact, in this case the matrix C equals the Hessian matrix, $C = H$. Since the matrix C contains all parameters, we say that the quadratic function is defined by the matrix C .

Figures 5–8 illustrate the surface and contour lines of a convex and nonconvex two-dimensional quadratic function. The contour lines of the convex function are concentric ellipses and a sublevel set L_c is represented by the points inside some ellipse. The point $(0, 0)$ in Figure 8 is a saddle point. The convex quadratic function is defined by the matrix

$$C = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$$

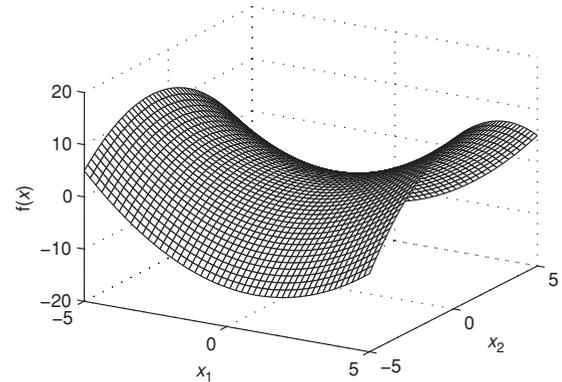


Figure 7 The Surface of a Nonconvex Two-Dimensional Quadratic Function

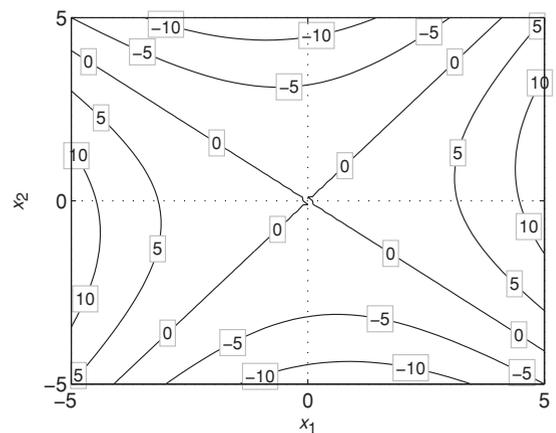


Figure 8 The Contour Lines of a Nonconvex Two-Dimensional Quadratic Function

and the nonconvex quadratic function is defined by the matrix

$$C = \begin{pmatrix} -1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$$

A property of convex functions is that the sum of convex functions is a convex function. As a result of the preceding analysis, the function

$$f(x) = \lambda x' C x - a' x \tag{9}$$

where $\lambda > 0$ and C is a positive semidefinite matrix, is a convex function as a sum of two convex functions. In the mean-variance efficient frontier, as formulated by Markowitz (1952), we find functions similar to (9). Let us use the properties of convex functions in order to solve the unconstrained problem of minimizing the function in (9):

$$\min_{x \in \mathbb{R}^n} \lambda x' C x - a' x$$

This function is differentiable and we can search for the global minimum by solving the first-order conditions:

$$\nabla f(x) = 2\lambda C x - a = 0$$

Therefore, the value of x minimizing the objective function equals

$$x^0 = \frac{1}{2\lambda} C^{-1} a$$

where C^{-1} denotes the inverse of the matrix C .

Quasi-Convex Functions

Besides convex functions, there are other classes of functions with convenient optimal properties. An example of such a class is the class of *quasi-convex functions*. Formally, a function is called quasi-convex if all sublevel sets defined in (7) are convex sets. Alternatively, a function $f(x)$ is called quasi-convex if

$$f(x^1) \geq f(x^2) \text{ implies } f(\alpha x^1 + (1 - \alpha)x^2) \leq f(x^1)$$

where x^1 and x^2 belong to the function domain, which should be a convex set, and $0 \leq \alpha \leq 1$.

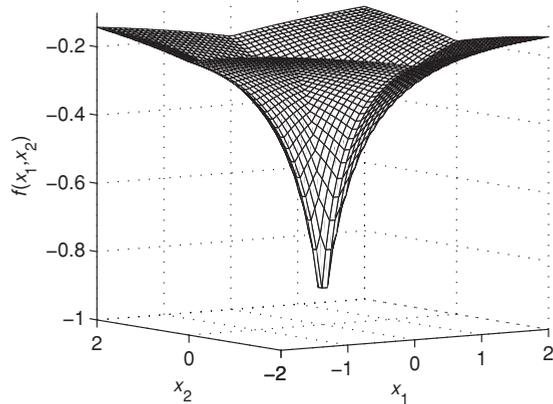


Figure 9 Example of a Two-Dimensional Quasi-Convex Function

A function f is called *quasi-concave* if $-f$ is quasi-convex.

An illustration of a two-dimensional quasi-convex function is given in Figure 9. It shows the graph of the function and Figure 10 illustrates the contour lines. A sublevel set is represented by all points inside some contour line. From a geometric viewpoint, the sublevel sets corresponding to the plotted contour lines are convex because any of them contains the straight line connecting any two points belonging to the set. Nevertheless, the function is not convex, which becomes evident from the surface in Figure 9. It is not guaranteed that a

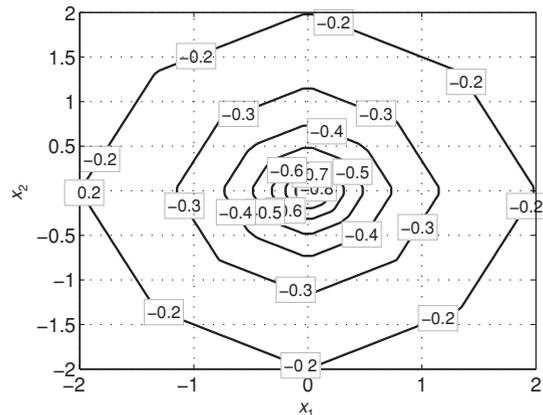


Figure 10 The Contour Lines of a Two-Dimensional Quasi-Convex Function

straight line connecting any two points on the surface will remain “above” the surface.

Properties of the quasi-convex functions include:

- Any convex function is also quasi-convex. The converse is not true, which is demonstrated in Figure 10.
- In contrast to the differentiable convex functions, the first-order condition is not necessary and sufficient for optimality in the case of differentiable quasi-convex functions. (There exists a class of functions larger than the class of convex functions but smaller than the class of quasi-convex functions, for which the first-order condition is necessary and sufficient for optimality. This is the class of pseudo-convex functions. Mangasarian [2006] provides more detail on the optimal properties of pseudo-convex functions.)
- It is possible to find a sequence of convex optimization problems yielding the global minimum of a quasi-convex function. Boyd and Vandenberghe (2004) provide further details. Its main idea is to find the smallest value of c for which the corresponding sublevel set L_c is nonempty. The minimal value of c is the global minimum, which is attained in the points belonging to the sublevel set L_c .
- Suppose that $g(x) > 0$ is a concave function and $f(x) > 0$ is a convex function. Then the ratio $g(x)/f(x)$ is a quasi-concave function and the ratio $f(x)/g(x)$ is a quasi-convex function.

Quasi-convex functions arise naturally in risk management when considering optimization of performance ratios. (See Chapter 10 in Rachev, Stoyanov, and Fabozzi [2008].)

CONSTRAINED OPTIMIZATION

In constructing optimization problems solving practical issues, it is very often the case that certain constraints need to be imposed in or-

der for the optimal solution to make practical sense. For example, long-only portfolio optimization problems require that the portfolio weights, which represent the variables in optimization, should be nonnegative and should sum up to one. According to the notation in this chapter, this corresponds to a problem of the type

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & x'e = 1 \\ & x \geq 0 \end{aligned} \quad (10)$$

where

$$\begin{aligned} f(x) &= \text{the objective function} \\ e \in \mathbb{R}^n &= \text{a vector of ones, } e = (1, \dots, 1) \\ x'e &= \text{the sum of all components of } x, \\ x'e &= \sum_i^n x_i \\ x \geq 0 &= \text{all components of the vector } x \in \mathbb{R}^n \\ &\text{are nonnegative} \end{aligned}$$

In problem (10), we are searching for the minimum of the objective function by varying x only in the set

$$\mathbf{X} = \left\{ x \in \mathbb{R}^n : \begin{array}{l} x'e = 1 \\ x \geq 0 \end{array} \right\} \quad (11)$$

which is also called the *set of feasible points* or the *constraint set*. A more compact notation, similar to the notation in the unconstrained problems, is sometimes used,

$$\min_{x \in \mathbf{X}} f(x)$$

where \mathbf{X} is defined in (11).

We distinguish between different types of optimization problems depending on the assumed properties for the objective function and the constraint set. If the constraint set contains only equalities, the problem is easier to handle analytically. In this case, the method of *Lagrange multipliers* is applied. For more general constraint sets, when they are formed by both equalities and inequalities, the method of Lagrange multipliers is generalized by the Karush-Kuhn-Tucker conditions (KKT conditions). Like the first-order conditions we considered in unconstrained optimization problems,

none of the two approaches lead to necessary and sufficient conditions for constrained optimization problems without further assumptions. One of the most general frameworks in which the KKT conditions are necessary and sufficient is that of *convex programming*. We have a convex programming problem if the objective function is a convex function and the *set of feasible points* is a convex set. As important subcases of convex optimization, *linear programming* and *convex quadratic programming* problems are considered.

In this section, we describe first the method of Lagrange multipliers, which is often applied to special types of mean-variance optimization problems in order to obtain closed-form solutions. Then we proceed with convex programming, which is the framework for reward-risk analysis.

Lagrange Multipliers

Consider the following optimization problem in which the set of feasible points is defined by a number of equality constraints:

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_1(x) = 0 \\ & h_2(x) = 0 \\ & \dots \\ & h_k(x) = 0 \end{aligned} \tag{12}$$

The functions $h_i(x)$, $i = 1, \dots, k$ build up the constraint set. Note that even though the right-hand side of the equality constraints is zero in the classical formulation of the problem given in (12), this is not restrictive. If in a practical problem the right-hand side happens to be different from zero, it can be equivalently transformed; for example:

$$\{x \in \mathbb{R}^n : v(x) = c\} \iff \{x \in \mathbb{R}^n : h_1(x) = v(x) - c = 0\}$$

In order to illustrate the necessary condition for optimality valid for (12), let us consider the

following two-dimensional example:

$$\begin{aligned} \min_{x \in \mathbb{R}^2} \quad & \frac{1}{2}x'Cx \\ \text{subject to} \quad & x'e = 1 \end{aligned} \tag{13}$$

where the matrix is

$$C = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$$

The objective function is a quadratic function and the constraint set contains one linear equality. A mean-variance optimization problem in which short positions are allowed is very similar to (13). (See Chapter 8 in Rachev, Stoyanov, and Fabozzi [2008].) The surface of the objective function and the constraint are shown in Figures 11 and 12. The black line on the surface shows the function values of the feasible points. Geometrically, solving problem (13) reduces to finding the lowest point of the black curve on the surface. The contour lines shown in Figure 12 imply that the feasible point yielding the minimum of the objective function is where a contour line is tangential to the line defined by the equality constraint. On the plot, the tangential contour line and the feasible points are in bold. The black dot indicates the position of the point in which the objective function attains its minimum subject to the constraints.

Even though the example is not general in the sense that the constraint set contains one linear

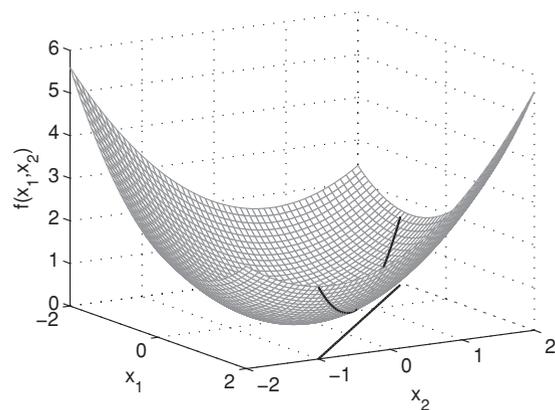


Figure 11 The Surface of a Two-Dimensional Quadratic Objective Function and the Linear Constraint $x_1 + x_2 = 1$

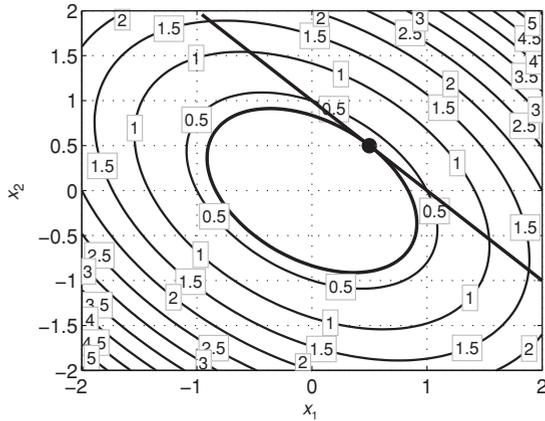


Figure 12 The Tangential Contour Line to the Linear Constraint $x_1 + x_2 = 1$

rather than a nonlinear equality, the same geometric intuition applies in the nonlinear case. The fact that the minimum is attained where a contour line is tangential to the curve defined by the nonlinear equality constraints in mathematical language is expressed in the following way: The gradient of the objective function at the point yielding the minimum is proportional to a linear combination of the gradients of the functions defining the constraint set. Formally, this is stated as:

$$\nabla f(x^0) - \mu_1 \nabla h_1(x^0) - \dots - \mu_k \nabla h_k(x^0) = 0 \quad (14)$$

where μ_i , $i = 1, \dots, k$ are some real numbers called *Lagrange multipliers* and the point x^0 is such that $f(x^0) \leq f(x)$ for all x which are feasible. Note that if there are no constraints in the problem, then (14) reduces to the first-order condition we considered in unconstrained optimization. Therefore, the system of equations behind (14) can be viewed as a generalization of the first-order condition in the unconstrained case.

The method of a Lagrange multipliers basically associates a function to the problem in (12) such that the first-order condition for unconstrained optimization for that function coincides with (14). The method of a Lagrange multiplier consists of the following steps.

1. Given the problem in (12), construct the following function:

$$L(x, \mu) = f(x) - \mu_1 h_1(x) - \dots - \mu_k h_k(x) \quad (15)$$

where $\mu = (\mu_1, \dots, \mu_k)$ is the vector of Lagrange multipliers. The function $L(x, \mu)$ is called the *Lagrangian* corresponding to problem (12).

2. Calculate the partial derivatives with respect to all components of x and μ and set them equal to zero:

$$\frac{\partial L(x, \mu)}{\partial x_i} = \frac{\partial f(x)}{\partial x_i} - \sum_{j=1}^k \mu_j \frac{\partial h_j(x)}{\partial x_i} = 0, \quad i = 1, \dots, n \quad (16)$$

$$\frac{\partial L(x, \mu)}{\partial \mu_m} = h_m(x) = 0, \quad m = 1, \dots, k$$

Basically, the system of equations (16) corresponds to the first-order conditions for unconstrained optimization written for the Lagrangian as a function of both x and μ , $L : \mathbb{R}^{n+k} \rightarrow \mathbb{R}$.

3. Solve the system of equalities in (16) for x and μ . Note that even though we are solving the first-order condition for unconstrained optimization of $L(x, \mu)$, the solution (x^0, μ^0) of (16) is not a point of local minimum or maximum of the Lagrangian. In fact, the solution (x^0, μ^0) is a saddle point of the Lagrangian.

The first n equations in (16) make sure that the relationship between the gradients given in (14) is satisfied. The following k equations in (16) make sure that the points are feasible. As a result, all vectors x solving (16) are feasible and the gradient condition is satisfied at them. Therefore, the points that solve the optimization problem (12) are among the solutions of the system of equations given in (16).

This analysis suggests that the method of Lagrange multipliers provides a necessary condition for optimality. Under certain assumptions for the objective function and the functions building up the constraint set, (16) turns out to be a necessary and sufficient condition. For

example, if $f(x)$ is a convex and differentiable function and $h_i(x)$, $i = 1, \dots, k$ are affine functions, then the method of Lagrange multipliers identifies the points solving (12). A function $h(x)$ is called affine if it has the form $h(x) = a + c'x$, where a is a constant and $c = (c_1, \dots, c_n)$ is a vector of coefficients. All linear functions are affine. Figure 12 illustrates a convex quadratic function subject to a linear constraint. In this case, the solution point is unique.

Convex Programming

The general form of convex programming problems is

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \quad i = 1, \dots, m \\ & h_j(x) = 0, \quad j = 1, \dots, k \end{aligned} \quad (17)$$

where

- $f(x)$ is a convex objective function
- $g_1(x), \dots, g_m(x)$ are convex functions defining the inequality constraints
- $h_1(x), \dots, h_k(x)$ are affine functions defining the equality constraints

Generally, without the assumptions of convexity, problem (17) is more involved than (12) because besides the equality constraints, there are inequality constraints. The KKT condition, generalizing the method of Lagrange multipliers, is only a necessary condition for optimality in this case. However, adding the assumption of convexity makes the KKT condition necessary and sufficient.

Note that, similar to problem (12), the fact that the right-hand side of all constraints is zero is nonrestrictive. The limits can be arbitrary real numbers.

Consider the following two-dimensional optimization problem

$$\begin{aligned} \min_{x \in \mathbb{R}^2} \quad & \frac{1}{2}x'Cx \\ \text{subject to} \quad & (x_1 + 2)^2 + (x_2 + 2)^2 \leq 3 \end{aligned} \quad (18)$$

in which

$$C = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$$

The objective function is a two-dimensional convex quadratic function and the function in the constraint set is also a convex quadratic function. In fact, the boundary of the feasible set is a circle with a radius of $\sqrt{3}$ centered at the point with coordinates $(-2, -2)$. Figures 13 and 14 show the surface of the objective function and the set of feasible points. The shaded part on the surface indicates the function values of all feasible points. In fact, solving problem (18) reduces to finding the lowest point on the shaded part of the surface. Figure 14 shows the contour lines of the objective function together with the feasible set, which is in gray. Geometrically, the point in the feasible set yielding the minimum of the objective function is positioned where a contour line only touches the constraint set. The position of this point is marked with a black dot and the tangential contour line is given in bold.

Note that the solution points of problems of the type (18) can happen to be not on the boundary of the feasible set but in the interior. For example, suppose that the radius of the circle defining the boundary of the feasible set in (18) is a larger number such that the point $(0, 0)$ is

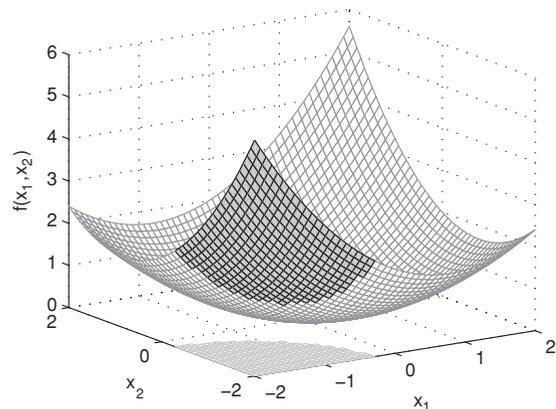


Figure 13 The Surface of a Two-Dimensional Convex Quadratic Function and a Convex Quadratic Constraint

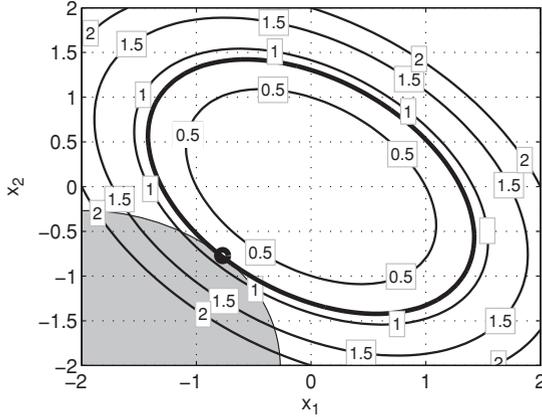


Figure 14 The Tangential Contour Line to the Feasible Set Defined by a Convex Quadratic Constraint

inside the feasible set. Then, the point $(0, 0)$ is the solution to problem (18) because at this point the objective function attains its global minimum.

In the two-dimensional case, when we can visualize the optimization problem, geometric reasoning guides us to finding the optimal solution point. In a higher dimensional space, plots cannot be produced and we rely on the analytic method behind the KKT conditions. The KKT conditions corresponding to the convex programming problem (17) are the following:

$$\begin{aligned} \nabla f(x) + \sum_{i=1}^m \lambda_i \nabla g_i(x) + \sum_{j=1}^k \mu_j \nabla h_j(x) &= 0 \\ g_i(x) &\leq 0 \quad i = 1, \dots, m \\ h_j(x) &= 0 \quad j = 1, \dots, k \\ \lambda_i g_i(x) &= 0, \quad i = 1, \dots, m \\ \lambda_i &\geq 0, \quad i = 1, \dots, m \end{aligned} \quad (19)$$

A point x^0 such that (x^0, λ^0, μ^0) satisfies (19) is the solution to problem (17). Note that if there are no inequality constraints, then the KKT conditions reduce to (16) in the method of Lagrange multipliers. Therefore, the KKT conditions generalize the method of Lagrange multipliers.

The gradient condition in (19) has the same interpretation as the gradient condition in the

method of Lagrange multipliers. The set of constraints

$$\begin{aligned} g_i(x) &\leq 0 \quad i = 1, \dots, m \\ h_j(x) &= 0 \quad j = 1, \dots, k \end{aligned}$$

guarantee that a point satisfying (19) is feasible. The next conditions

$$\lambda_i g_i(x) = 0, \quad i = 1, \dots, m$$

are called complementary slackness conditions. If an inequality constraint is satisfied as a strict inequality, then the corresponding multiplier λ_i turns into zero according to the complementary slackness conditions. In this case, the corresponding gradient $\nabla g_i(x)$ has no significance in the gradient condition. This reflects the fact that the gradient condition concerns only the constraints satisfied as equalities at the solution point.

Important special cases of convex programming problems include linear programming problems and convex quadratic programming problems, which we consider in the remaining part of this section.

Linear Programming

Optimization problems are said to be linear programming problems if the objective function is a linear function and the feasible set is defined by linear equalities and inequalities. Since all functions are linear, they are also convex, which means that linear programming problems are also convex problems. The definition of linear programming problems in standard form is the following:

$$\begin{aligned} \min_x \quad & c'x \\ \text{subject to} \quad & Ax \leq b \\ & x \geq 0 \end{aligned} \quad (20)$$

where A is an $m \times n$ matrix of coefficients, $c = (c_1, \dots, c_n)$ is a vector of objective function coefficients, and $b = (b_1, \dots, b_m)$ is a vector of real numbers. As a result, the constraint set contains m inequalities defined by linear functions. The feasible points defined by means of linear

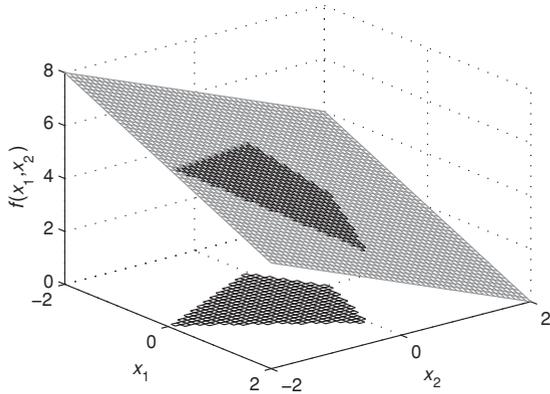


Figure 15 The Surface of a Linear Function and a Polyhedral Feasible Set

equalities and inequalities are also said to form a polyhedral set. In practice, before solving a linear programming problem, it is usually first reformulated in the standard form given in (20).

Figures 15 and 16 show an example of a two-dimensional linear programming problem which is not in standard form as the two variables may become negative. Figure 15 contains the surface of the objective function, which is a plane in this case, and the polyhedral set of feasible points. The shaded area on the surface corresponds to the points in the feasible set. Solving problem (20) reduces to finding the

lowest point in the shaded area on the surface. Figure 16 shows the feasible set together with the contour lines of the objective function. The contour lines are parallel straight lines because the objective function is linear. The point in which the objective function attains its minimum is marked with a black dot.

A general result in linear programming is that, on condition that the problem is bounded, the solution is always at the boundary of the feasible set and, more precisely, at a vertex of the polyhedron. Problem (20) may become unbounded if the polyhedral set is unbounded and there are feasible points such that the objective function can decrease indefinitely. We can summarize that, generally, due to the simple structure of (20), there are three possibilities:

1. The problem is not feasible, because the polyhedral set is empty.
2. The problem is unbounded.
3. The problem has a solution at a vertex of the polyhedral set.

From computational viewpoint, the polyhedral set has a finite number of vertices and an algorithm can be devised with the goal of finding a vertex solving the optimization problem in a finite number of steps. This is the basic idea behind the simplex method, which is an efficient numerical approach to solving linear programming problems. Besides the simplex algorithm, there are other, more contemporary methods, such as the interior point method.

Quadratic Programming

Besides linear programming, another class of problems with simple structure is the class of quadratic programming problems. It contains optimization problems with a quadratic objective function and linear equalities and inequalities in the constraint set:

$$\begin{aligned} \min_x \quad & c'x + \frac{1}{2}x'Hx \\ \text{subject to} \quad & Ax \leq b \end{aligned} \tag{21}$$

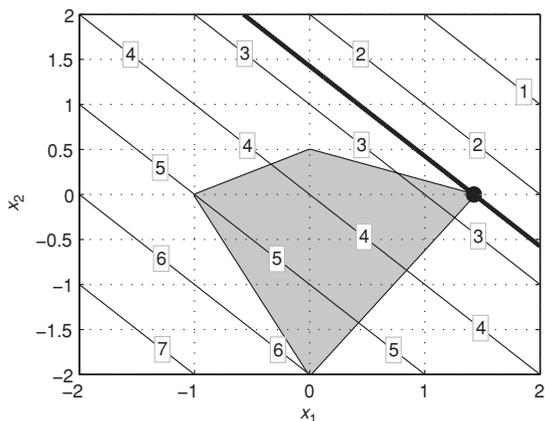


Figure 16 The Bottom Plot Shows the Tangential Contour Line to the Polyhedral Feasible Set

where

- $c = (c_1, \dots, c_n)$ is a vector of coefficients defining the linear part of the objective function
- $H = \{h_{ij}\}_{i,j=1}^n$ is an $n \times n$ matrix defining the quadratic part of the objective
- $A = \{a_{ij}\}$ is a $k \times n$ matrix defining k linear inequalities in the constraint set
- $b = (b_1, \dots, b_k)$ is a vector of real numbers defining the right-hand side of the linear inequalities

In optimal portfolio theory, mean-variance optimization problems in which portfolio variance is in the objective function are quadratic programming problems.

From the point of view of optimization theory, problem (21) is a convex optimization problem if the matrix defining the quadratic part of the objective function is positive semidefinite. In this case, the KKT conditions can be applied to solve it.

KEY POINTS

1. The mathematical theory of optimization concerns identifying the best alternative within a set of available, or feasible, alternatives and finds application in different areas of finance such as portfolio selection or, more generally, explaining behavior of economic agents in the face of uncertainty.
2. An optimization problem has two important components: an objective function defining the criterion to be optimized and a feasibility set described by means of equality or inequality constraints.
3. The properties of the objective function and the feasibility set are used to distinguish different classes of optimization problems with specific conditions for optimality and numerical solution methods. The most important classes include linear, quadratic, and convex programming problems.
4. In the theory of portfolio selection, the classical mean-variance analysis belongs to the class of quadratic optimization problems.
5. Employing more general reward and risk measures can result in a convex optimization problem but if scenarios for assets returns are available, the portfolio selection problem can be simplified to a linear programming problem in some cases. Optimization of performance ratios can be related to quasi-convex programs.

REFERENCES

- Boyd, S., and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge: Cambridge University Press.
- Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance* 7, 1: 77–91.
- Mangasarian, O. (2006). Pseudo-convex functions. In W. Ziemba and G. Vickson (eds.), *Stochastic Optimization Models in Finance* (pp. 23–32), Singapore: World Scientific Publishing.
- Rachev, Z. T., Stoyanov, S., and Fabozzi, F. J. (2008). *Advanced Stochastic Models, Risk Assessment and Portfolio Optimization: The Ideal Risk, Uncertainty and Performance Measures*. Hoboken, NJ: John Wiley & Sons.
- Rockafellar, R. T. (1997). *Convex Analysis*. Princeton, NJ: Princeton University Press, Landmarks in Mathematics.
- Ruszczynski, A. (2006). *Nonlinear Optimization*. Princeton, NJ: Princeton University Press.

Asset Allocation and Portfolio Construction Techniques in Designing the Performance-Seeking Portfolio

NOËL AMENC, PhD

Professor of Finance, EDHEC Business School, Director, EDHEC-Risk Institute

FELIX GOLTZ, PhD

Head of Applied Research, EDHEC-Risk Institute

LIONEL MARTELLINI, PhD

Professor of Finance, EDHEC Business School, Scientific Director, EDHEC-Risk Institute

PATRICE RETKOWSKY

Senior Research Engineer, EDHEC-Risk Institute

Abstract: Meeting the challenges of modern investment practice involves the design of novel forms of investment solutions, as opposed to investment products customized to meet investors' expectations. These new forms of investment solutions rely on the use of improved, more efficient performance-seeking portfolio and liability-hedging portfolio building blocks, as well as on the use of improved dynamic allocation strategies. Understanding the conceptual and technical challenges involved in the design of improved benchmarks for the performance-seeking portfolio is critical.

Management is justified as an industry by the capacity of adding value through the design of investment solutions that match investors' needs. For more than 50 years, the industry has in fact mostly focused on security selection decisions as a single source of added value.

This sole focus has somewhat distracted the industry from another key source of added value, namely portfolio construction and asset allocation decisions. In the face of recent crises, and given the intrinsic difficulty in delivering added value through security selection

decisions only, the relevance of the old paradigm has been questioned with heightened intensity, and a new paradigm is starting to emerge.

Academic research has provided very useful guidance with respect to how asset allocation and portfolio construction decisions should be analyzed so as to best improve investors' welfare. In a nutshell, the "fund separation theorems" that lie at the core of modern portfolio theory advocate a separate management of performance and risk control objectives. In the context of asset allocation decisions with consumption/liability objectives, it can be shown that the suitable expression of the fund separation theorem provides rational support for *liability-driven investment* (LDI) techniques that have recently been promoted by a number of investment banks and asset management firms. These solutions involve on the one hand the design of a customized *liability-hedging portfolio* (LHP), the sole purpose of which is to hedge away as effectively as possible the impact of unexpected changes in risk factors affecting liability values (most notably interest rate and inflation risks), and on the other hand the design of a *performance-seeking portfolio* (PSP), whose raison d'être is to provide investors with an optimal risk-return trade-off.

One of the implications of this LDI paradigm is that one should distinguish two different levels of asset allocation decisions: allocation decisions involved in the design of the performance-seeking or the liability-hedging portfolio (design of better building blocks, or BBBs), and asset allocation decisions involved in the optimal split between the PSP and the LHP (designed of advanced asset allocation decisions, or AAAs). We address the question of better building blocks in detail in this entry and provide some thoughts on integrating these building blocks in asset allocation. More specifically, we mainly focus here on how to construct efficient performance-seeking portfolios.

In this entry we provide an overview of the key conceptual challenges involved in asset al-

location and portfolio construction in designing the performance-seeking portfolio. We begin by presenting the fundamental principle of the maximization of risk/reward efficiency and then deal with estimation of risk parameters and expected return parameters. The empirical results of optimal portfolio construction modeling are presented. We also provide a brief discussion on integrating such properly designed building blocks in the overall PSP at the asset allocation level.

THE TANGENCY PORTFOLIO AS THE RATIONALE BEHIND SHARPE RATIO MAXIMIZATION

Modern portfolio theory provides some useful guidance with respect to the optimal design of a PSP that would best suit investors' needs. More precisely, the prescription is that the PSP should be obtained as the result of a portfolio optimization procedure aiming at generating the highest risk-reward ratio.

Portfolio optimization is a straightforward procedure, at least in principle. In a mean-variance setting, for example, the prescription consists of generating a maximum Sharpe ratio (MSR) portfolio based on expected return, volatility, and pairwise correlation parameters for all assets to be included in the portfolio, a procedure that can even be handled analytically in the absence of portfolio constraints.

More precisely, consider a simple mean-variance problem:

$$\max_w \mu_p - \frac{1}{2} \gamma \sigma_p^2$$

Here, the control variable is a vector w of optimal weight allocated to various risky assets, μ_p denotes the portfolio expected return, and σ_p denotes the portfolio volatility. We further assume that the investor is facing the following investment opportunity set: a riskless bond paying the risk-free rate r , and a set of N risky

assets with expected return vector μ (of size N) and covariance matrix Σ (of size $N \times N$), all assumed constant so far.

With these notations, the portfolio expected return and volatility are respectively given by:

$$\begin{aligned}\mu_p &= w'(\mu - re) + r \\ \sigma_p^2 &= w'\Sigma w\end{aligned}$$

In this context, it is straightforward to show by standard arguments that the only efficient portfolio composed with risky assets is the maximum Sharpe ratio portfolio, also known as the tangency portfolio.¹

Finally, the Sharpe ratio reads (where we further denote by e vector of ones of size N):

$$SR = \frac{w'(\mu - re)}{(w'\Sigma w)^{1/2}}$$

And the optimal portfolio is given by:

$$\begin{aligned}\max_w \left(\mu_p - \frac{1}{2}\gamma\sigma_p^2 \right) &\Rightarrow w_0^* = \frac{1}{\gamma}\Sigma^{-1}(\mu - re) \\ &= \frac{e'\Sigma^{-1}(\mu - re)}{\gamma} \underbrace{\frac{\Sigma^{-1}(\mu - re)}{e'\Sigma^{-1}(\mu - re)}}_{PSP}\end{aligned}\quad (1)$$

This is a two-fund separation theorem, which gives the allocation to the MSR performance-seeking portfolio (PSP), with the rest invested in cash, as well as the composition of the MSR performance-seeking portfolio.

In practice, investors end up holding more or less imperfect proxies for the truly optimal performance-seeking portfolio, if only because of the presence of parameter uncertainty, which makes it impossible to obtain a perfect estimate for the maximum Sharpe ratio portfolio. Denoting by λ the Sharpe ratio of the (generally inefficient) PSP actually held by the investor, and by σ its volatility, we obtain the following optimal allocation strategy:

$$w_0^* = \frac{\lambda}{\gamma\sigma}PSP \quad (2)$$

Hence the allocation to the performance-seeking portfolio is a function of two objective parameters, the PSP volatility and the PSP

Sharpe ratio, and one subjective parameter, the investor's risk aversion. The optimal allocation to the PSP is inversely proportional to the investor's risk aversion. If risk aversion goes to infinity, the investor holds the risk-free asset only, as should be expected. For finite risk-aversion levels, the allocation to the PSP is inversely proportional to the PSP volatility, and it is proportional to the PSP Sharpe ratio. As a result, if the Sharpe ratio of the PSP is increased, one can invest more in risky assets. Hence, portfolio construction modeling is not only about risk reduction; it is also about performance enhancement through a better spending of investors' risk budgets.

The expression (1) is useful because it provides in principle a straightforward expression for the optimal portfolio starting from a set of N risky assets. In the presence of a realistically large number N of securities, the curse of dimensionality, however, makes it practically impossible for investors to implement such direct one-step portfolio optimization decisions involving all individual components of the asset mix. The standard alternative approach widely adopted in investment practice consists instead in first grouping individual securities in various asset classes according to various dimensions, for example, country, sector, and/or style within the equity universe, or country, maturity, and credit rating within the bond universe, and subsequently generating the optimal portfolio through a two-stage process. On the one hand, investable proxies are generated for maximum Sharpe ratio (MSR) portfolios within each asset class in the investment universe. We call this step, which is typically delegated to professional money managers, the portfolio construction step. On the other hand, when the MSR proxies are obtained for each asset class, an optimal allocation to the various asset classes is eventually generated so as to generate the maximum Sharpe ratio at the global portfolio level. This step is called the asset allocation step, and it is typically handled by a centralized decision maker (e.g., a pension fund CIO) with or

without the help of specialized consultants, as opposed to being delegated to decentralized asset managers. In this entry, the discussion focuses on the first step, and we provide some concluding remarks on its relation to the second step at the end of this entry.

For the definition of *building blocks for asset allocation*, in the absence of active views, the default option consists of using market cap weighted indexes as proxies for the asset class MSR portfolio. Academic research, however, has found that such market cap indexes were likely to be severely inefficient portfolios.² In a nutshell, market cap weighted indexes are not good choices as investment benchmarks because they are poorly diversified portfolios. In fact, cap-weighting tends to lead to exceedingly high concentration in relatively few stocks. As a consequence of their lack of diversification, cap weighted indexes have been empirically found to be severely inefficient portfolios, which do not provide investors with the fair reward given the risk taken. As a result of their poor diversification, they have been found to be dominated by equally weighted benchmarks,³ which are naïvely diversified portfolios that are optimal if and only if all securities have identical expected return, volatilities, and all pairs of correlations are identical.

In what follows, we analyze in some detail a number of alternatives based on practical implementation of modern portfolio theory that have been suggested to generate more efficient proxies for the MSR portfolio in the equity investment universe. (See Figure 1.)

Modern portfolio theory was born with the efficient frontier analysis of Markowitz (1952). Unfortunately, early applications of the technique, based on naïve estimates of the input parameters, have been found of little use because they lead to nonsensible portfolio allocations.

In a first section, we explain how to help bridge the gap between portfolio theory and portfolio construction by showing how to generate enhanced parameter estimates so as to improve the quality of the portfolio optimiza-

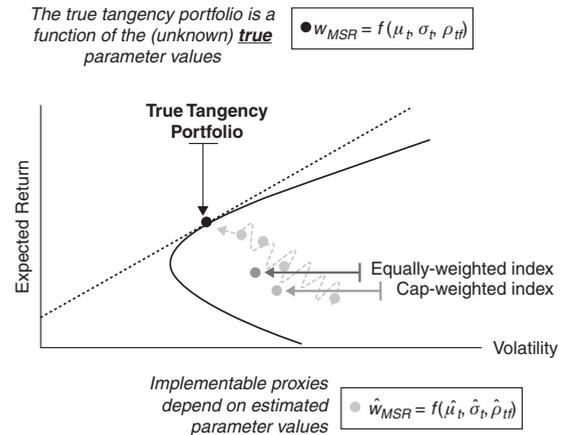


Figure 1 Inefficiency of Cap-Weighted Benchmarks, and the Quest for an Efficient Proxy for the True Tangency Portfolio

tion outputs (optimal portfolio weights). We first focus on enhanced covariance parameter estimates and explain how to meet the main challenge of sample risk reduction.⁴ Against this backdrop, we present the state-of-the-art methodologies for reducing the problem dimensionality and estimating the covariance matrix with multifactor models. We then turn to expected return estimation. We argue that statistical methodologies are not likely to generate any robust expected return estimates, which suggests that economic models such as the single-factor CAPM and the multifactor APT should instead be used for expected return estimation. Finally, we also present evidence that proxies for expected return estimates should not only include systematic risk measures, but they should also incorporate idiosyncratic risk measures as well as downside risk measures.

ROBUST ESTIMATORS FOR COVARIANCE PARAMETERS

In practice, success in the implementation of a theoretical model relies not only upon its conceptual grounds but also on the reliability of the inputs of the model. In the case of

mean-variance (MV) optimization the results will highly depend on the quality of the parameter estimates: the covariance matrix and the expected returns of assets.

Several improved estimates for the covariance matrix have been proposed, including most notably the factor-based approach suggested by Sharpe (1963), the constant correlation approach suggested by Elton and Gruber (1973), and the statistical shrinkage approach suggested by Ledoit and Wolf (2004). In addition, Jagannathan and Ma (2003) find that imposing (non-short selling) constraints on the weights in the optimization program improves the risk-adjusted out-of-sample performance in a manner that is similar to some of the aforementioned improved covariance matrix estimators.

In these papers, the authors have focused on testing the out-of-sample performance of global minimum variance (GMV) portfolios, as opposed to the MSR portfolios (also known as tangency portfolios), given that there is a consensus around the fact that purely statistical estimates of expected returns are not robust enough to be used. (This is discussed later in this entry.)

The key problem in covariance matrix estimation is the curse of dimensionality; when a large number of stocks are considered, the number of parameters to estimate grows exponentially, where the majority of them are pairwise correlations.

Therefore, at the estimation stage, the challenge is to reduce the number of factors that come into play. In general, a multifactor model decomposes the (excess) return (in excess to the risk-free asset) of an asset into its expected rewards for exposition to the “true” risk factors as follows:

$$r_{it} = \alpha_{it} + \sum_{j=1}^K \beta_{i,jt} \cdot F_{jt} + \varepsilon_{it}$$

or in matrix form for all N assets:

$$r_t = \alpha_t + \beta_t F_t + \varepsilon_t$$

where β_t is an $N \times K$ matrix containing the sensitivities of each asset i with respect to the corresponding j -th factor movements; r_t is the vector of the N assets' (excess) returns, F_t a vector containing the K risk factors' (excess) returns, and ε_t the $N \times 1$ vector containing the zero mean uncorrelated residuals ε_{it} . The covariance matrix for the asset returns implied by a factor model is given by:

$$\Omega = \beta \cdot \Sigma_F \cdot \beta^T + \Sigma_\varepsilon$$

where Σ_F is the $K \times K$ covariance matrix of the risk factors and Σ_ε an $N \times N$ covariance matrix of the residuals corresponding to each asset.

While the factor-based estimator is expected to allow for a reasonable trade-off between sample risk and model risk, there still remains, however, the problem of choosing the “right” factor model. One popular approach aims at relying as little as possible on strong theoretical assumptions by using principal components analysis (PCA) to determine the underlying risk factors from the data. The PCA method is based on a spectral decomposition of the sample covariance matrix, and its goal is to explain covariance structures using only a few linear combinations of the original stochastic variables, which will constitute the set of (unobservable) factors.

Bengtsson and Holst (2002) and Fujiwara et al. (2006) motivate the use of PCA in a similar way, extracting principal components in order to estimate expected correlation within MV portfolio optimization. Fujiwara et al. (2006) find that the realized risk-return of portfolios based on the PCA method outperforms the single-index-based one and that the optimization gives a practically reasonable asset allocation. Overall, the main strength of the PCA approach at this stage is that it allows “the data to talk” and has them tell the financial modeler what the underlying risk factors are that govern most of the variability of the assets at each point in time. This strongly contrasts with having to rely on the assumption that a particular factor model is the true pricing model and reduces the

specification risk embedded in the factor-based approach while keeping the sample risk reduction.

The question of determining the appropriate number of factors to structure the correlation matrix is critical for the risk estimation when using PCA as a factor model. Several options have been proposed to answer this question, some of them with more theoretical grounds than others.

As a final note, we need to recognize that the discussion is so far cast in a mean-variance setting, which can in principle only be rationalized for normally distributed asset returns. In the presence of non-normally distributed asset returns, optimal portfolio selection techniques require estimates for variance-covariance parameters, along with estimates for higher-order moments and comoments of the return distribution. This is a formidable challenge that severely exacerbates the dimensionality problem already present with mean-variance analysis. In a recent paper, Martellini and Ziemann (2010) extend the existing literature, which has mostly focused on the covariance matrix, by introducing improved estimators for the coskewness and cokurtosis parameters. On the one hand, they find that the use of these enhanced estimates generates a significant improvement in investors' welfare. On the other hand, they find that also that when the number of constituents in the portfolios is large (e.g., exceeding 20), the increase in sample risk related to the need to estimate higher-order comoments by far outweighs the benefits related to considering a more general portfolio optimization procedure. In the end, when portfolio optimization is performed on the basis of a large number of individual securities, it appears that maximizing the portfolio Sharpe ratio leads to a better out-of-sample return-to-VaR ratio or return-to-CVaR ratio compared to a procedure focusing on maximizing the return-to-VaR ratio or the return-to-CVaR ratio, a result that holds true even if improved estimators are used for higher-order comoments.

ROBUST ESTIMATORS FOR EXPECTED RETURNS

While it appears that risk parameters can be estimated with a fair degree of accuracy, it has been shown (Merton, 1980) that expected returns are difficult to obtain with a reasonable estimation error. What makes the problem worse is that optimization techniques are very sensitive to differences in expected returns, so that portfolio optimizers typically allocate the largest fraction of capital to the asset class for which estimation error in the expected returns is the largest.⁵

In the face of the difficulty of using sample-based expected return estimates in a portfolio optimization context, a reasonable alternative consists in using some risk estimate as a proxy for excess expected returns.⁶ This approach is based on the most basic principle in finance; that is, the natural relationship between risk and reward. In fact, standard asset pricing theories such as the arbitrage pricing theory as proposed by Ross (1976) imply that expected returns should be positively related to systematic volatility, such as measured through a factor model that summarizes individual stock return exposure with respect to a number of rewarded risk factors.

More recently, a series of papers have focused on the explanatory power of idiosyncratic, as opposed to systematic, risk for the cross section of expected returns. In particular, Malkiel and Xu (2006), extending an insight from Merton (1987), show that an inability to hold the market portfolio, whatever the cause, will force investors to care about total risk to some degree in addition to market risk so that firms with larger firm-specific variances require higher average returns to compensate investors for holding imperfectly diversified portfolios.⁷ That stocks with high idiosyncratic risk earn higher returns has also been confirmed in a number of recent empirical studies, including in particular Tinic and West (1986) as well as Malkiel and Xu (1997, 2006).

Taken together, these findings suggest that total risk, a model-free quantity given by the sum of systematic and specific risk, should be positively related to expected return. Most commonly, total risk is the volatility of a stock's returns. Martellini (2008) has investigated the portfolio implications of these findings and has found that tangency portfolios constructed on the assumption that the cross-section of excess expected returns could be approximated by the cross-section of volatility posted better out-of-sample risk-adjusted performance than their market-cap-weighted counterparts.

More generally, recent research suggests that the cross-section of expected returns might be best explained by risk indicators taking into account higher-order moments. Theoretical models have shown that, in exchange for higher skewness and lower kurtosis of returns, investors are willing to accept expected returns lower (and with volatility higher) than those of the mean-variance benchmark.⁸ More specifically, skewness and kurtosis in individual stock returns (as opposed to the skewness and kurtosis of aggregate portfolios) have been shown to matter in several papers. High skewness is associated with lower expected returns in Barberis and Huang (2004), Brunnermeier, Gollier, and Parker (2005), and Mitton and Vorkink (2007). The intuition behind this result is that investors like to hold positively skewed portfolios. The highest skewness is achieved by concentrating portfolios in a small number of stocks that themselves have positively skewed returns. Thus investors tend to be underdiversified and drive up the price of stocks with high positive skewness, which in turn reduces their future expected returns. Stocks with negative skewness are relatively unattractive and thus have low prices and high returns. The preference for kurtosis is in the sense that investors like low kurtosis and thus expected returns should be positively related to kurtosis. Boyer, Mitton, and Vorkink (2010) and Conrad, Dittmar and Ghysels (2008) provide empirical evidence that individual stocks' skewness and kurtosis is in-

deed related to future returns. An alternative to direct consideration of the higher moments of returns is to use a risk measure that aggregates the different dimensions of risk. In this line, Bali and Cakici (2004) show that future returns on stocks are positively related to their value-at-risk and Estrada (2000) and Chen, Chen, and Chen (2009) show that there is a relationship between downside risk and expected returns.

IMPLICATIONS FOR BENCHMARK PORTFOLIO CONSTRUCTION

Once careful estimates for risk and return parameters have been obtained, one may then design efficient proxies for an asset class benchmark with an attractive risk-return profile. For example, Amenc et al. (2011) find that efficient equity benchmarks designed on the basis of robust estimates for risk and expected return parameters substantially outperform in terms of risk-adjusted performance market cap weighted indexes that are often used as default options for investment benchmarks in spite of their well-documented lack of efficiency.⁹

Table 1, borrowed from Amenc et al. (2011), shows summary performance statistics for an efficient index constructed according to the aforementioned principles. For the average return, volatility, and the Sharpe ratio, we report differences with respect to cap-weighting and assess whether this difference is statistically significant.

Table 1 shows that the efficient weighting of index constituents leads to higher average returns, lower volatility, and a higher Sharpe ratio. All these differences are statistically significant at the 10% level, whereas the difference in Sharpe ratios is significant even at the 0.1% level. Given the data, it is highly unlikely that the unobservable true performance of efficient weighting was not different from that of capitalization weighting. Economically, the

Table 1 Risk and Return Characteristics for the Efficient Index

Index	Ann. Average Return (compounded)	Ann. Standard Deviation	Sharpe Ratio (compounded)	Information Ratio	Tracking Error
Efficient index	11.63%	14.65%	0.41	0.52	4.65%
Cap-weighted	9.23%	15.20%	0.24	0.00	0.00%
Difference (efficient minus cap-weighted)	2.40%	-0.55%	0.17	-	-
<i>p</i> -value for difference	0.14%	6.04%	0.04%	-	-

The table shows risk and return statistics portfolios constructed with the same set of constituents as the cap-weighted index. Rebalancing is quarterly subject to an optimal control of portfolio turnover (by setting the reoptimization threshold to 50%). Portfolios are constructed by maximizing the Sharpe ratio given an expected return estimate and a covariance estimate. The expected return estimate is set to the median total risk of stocks in the same decile when sorting by total risk. The covariance matrix is estimated using an implicit factor model for stock returns. Weight constraints are set so that each stock's weight is between $1/2N$ and $2/N$, where N is the number of index constituents. The *p*-values for differences are computed using the paired *t*-test for the average, the *F*-test for volatility, and a Jobson-Korkie test for the Sharpe ratio. The results are based on weekly return data from 01/1959 to 12/2008.

performance difference is pronounced, as the Sharpe ratio increases by about 70%.

ASSET ALLOCATION MODELING: PUTTING THE EFFICIENT BUILDING BLOCKS TOGETHER

After efficient benchmarks have been designed for various asset classes, these building blocks can be assembled in a second step, the asset allocation step, to build a well-designed multiclass performance-seeking portfolio.

While the methods we have discussed so far can in principle be applied in both contexts, a number of key differences should be emphasized.

In the asset allocation context, the number of constituents is small, and using time- and state-dependent covariance matrix estimates becomes reasonable, while they do not necessarily improve the situation in portfolio construction contexts when the number of constituents is large. Similarly, while it is not feasible in general, as explained above, to perform portfolio optimization with higher-order moments in a portfolio construction context where the number of constituents is typically

large, it is reasonable to go beyond mean-variance analysis in an asset allocation context where the number of constituents is limited.

Furthermore, in an asset allocation context, the universe is not homogenous, which has implications for expected returns and covariance estimation. In terms of a covariance matrix, it will not prove easy to obtain a universal factor model for the whole investment universe. In this context, it is arguably better to use statistical shrinkage toward, say, the constant correlation model, as opposed to using a factor model approach.¹⁰

KEY POINTS

- Modern portfolio theory advocates the separation of the management of performance and risk control objectives. In the context of asset allocation decisions, the fund separation theorem provides rational support for liability-driven investment techniques whose solutions involve the design of a customized liability-hedging portfolio and the design of a performance-seeking portfolio.
- The sole purpose of the liability-hedging portfolio is to hedge away as effectively as possible the impact of unexpected changes in risk

factors affecting liability values (most notably interest rate and inflation risks); the purpose of the performance-seeking portfolio is to provide investors with an optimal risk-return trade-off.

- An implication of the liability-driven investment paradigm is that one should distinguish two different levels of asset allocation decisions: (1) decisions involved in the design of the performance-seeking or the liability-hedging portfolio (design of better building blocks), and (2) decisions involved in the optimal split between the performance-seeking portfolio and liability-hedging portfolio (designed of advanced asset allocation decisions).
- Although modern portfolio theory provides some useful guidance with respect to the optimal design of performance-seeking portfolios that would best suit investors' needs, in practice, investors end up holding more or less imperfect proxies for the truly optimal performance-seeking portfolio, if only because of the presence of parameter uncertainty, which makes it impossible to obtain a perfect estimate for the maximum Sharpe ratio portfolio.
- The allocation to the performance-seeking portfolio is a function of two objective parameters, the PSP volatility and the PSP Sharpe ratio, and one subjective parameter, the investor's risk aversion. The optimal allocation to the PSP is inversely proportional to the investor's risk aversion.
- In practice, the success in the implementation of a theoretical model relies not only upon its conceptual grounds but also on the reliability of the inputs of the model. In the case of mean-variance optimization the results will highly depend on the quality of the parameter estimates: the covariance matrix and the expected returns of assets.
- Several improved estimates for the covariance matrix have been proposed: the factor-based approach, constant correlation approach, and statistical shrinkage approach.
- The key problem in covariance matrix estimation is the curse of dimensionality. Consequently, at the estimation stage, the challenge is to reduce the number of factors that come into play. In general, a multifactor model decomposes the (excess) return (in excess to the risk-free asset) of an asset into its expected rewards for exposition to the "true" risk factors.
- The problem of choosing the right factor model still remains. The statistical technique of principal components analysis is commonly used to determine the underlying risk factors from the data.
- While it appears that risk parameters can be estimated with a fair degree of accuracy, it has been shown that expected returns are difficult to obtain with a reasonable estimation error. What makes the problem worse is that optimization techniques are very sensitive to differences in expected returns, so that portfolio optimizers typically allocate the largest fraction of capital to the asset class for which estimation error in the expected returns is the largest. In the face of the difficulty of using sample-based expected return estimates in a portfolio optimization context, a reasonable alternative consists in using some risk estimate as a proxy for excess expected returns.
- Research suggests that the cross-section of expected returns might be best explained by risk indicators taking into account higher-order moments. Theoretical models have shown that, in exchange for higher skewness and lower kurtosis of returns, investors are willing to accept expected returns lower (and volatility higher) than those of the mean-variance benchmark.
- Once careful estimates for risk and return parameters have been obtained, one may then design efficient proxies for an asset class benchmark with an attractive risk-return profile. After efficient benchmarks have been designed for various asset classes, these building blocks can be assembled in a second step, the asset allocation step, to build

a well-designed multiclass performance-seeking portfolio.

NOTES

1. For more details, see Appendix A.1 in Amenc, Goltz, Martellini, and Mihau (2011).
2. See, for example, Haugen and Baker (1991), Grinold (1992), or Amenc, Goltz, and Le Sourd (2006).
3. De Miguel et al. (2009).
4. Another key challenge is the presence of nonstationary risk parameters, which can be accounted for with conditional factor models capturing time-dependencies (e.g., GARCH-type models) and state-dependencies (e.g., Markov regime switching models) in risk parameter estimates.
5. See Britten-Jones (1999) or Michaud (1998).
6. This discussion focuses on estimating the fair neutral reward for holding risky assets. If one has access to active views on expected returns, one may use a disciplined approach (e.g., the Black-Litterman model) to combine the active views with the neutral estimates.
7. See also Barberis and Huang (2001) for a similar conclusion from a behavioral perspective.
8. See Rubinstein (1973) and Kraus and Litzenberger (1976).
9. See, for example, Haugen and Baker (1991) and Grinold (1992).
10. See Ledoit and Wolf (2003, 2004).

REFERENCES

- Amenc, N., Goltz, F., and Le Sourd, V. (2006). Assessing the quality of stock market indices. EDHEC-Risk Institute Publication (September).
- Amenc, N., Goltz, F., Martellini, L., and Milhau, V. (2011). Asset allocation and portfolio construction. In F. J. Fabozzi and H. M. Markowitz (Eds.), *The Theory and Practice of Investment Management: Second Edition*. Hoboken, NJ: John Wiley & Sons.
- Amenc, N., Goltz, F., Martellini, L., and Retkowsky, P. Efficient indexation: An alternative to cap-weighted indices. *Journal of Investment Management*, forthcoming.
- Bali, T., and Cakici, N. (2004). Value at risk and expected stock returns. *Financial Analysts Journal* 60, 2 (March/April): 57–73.
- Barberis, N., and Huang, M. (2004). Stock as lotteries: The implication of probability weighting for security prices. Working paper (Stanford and Yale University).
- Barberis, N., and Huang, M. (2001). Mental accounting, loss aversion and individual stock returns. *Journal of Finance* 56, 4 (August): 1247–1292.
- Bengtsson, C., and Holst, J. (2002). On portfolio selection: Improved covariance matrix estimation for Swedish asset returns. Working paper (Lund University and Lund Institute of Technology).
- Boyer, B., Mitton, T., and Vorkink, K. (2010). Expected idiosyncratic skewness. *Review of Financial Studies* 23, 1 (January): 169–202.
- Britten-Jones, M. (1999). The sampling error in estimates of mean-variance efficient portfolio weights. *Journal of Finance* 54, 2 (April): 655–671.
- Brunnermeier, M. K., Gollier, C., and Parker, J. (2007). Optimal beliefs, asset prices, and the preference for skewed returns. *American Economic Review* 97, 2 (May): 159–165.
- Chen, D. H., Chen, C. D., and Chen, J. (2009). Downside risk measures and equity returns in the NYSE. *Applied Economics* 41, 8 (March): 1055–1070.
- Conrad, J., Dittmar, R. F., and Ghysels, E. (2008). Ex ante skewness and expected stock returns. Working paper (University of North Carolina at Chapel Hill).
- De Miguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *Review of Financial Studies* 22, 5: 1915–1953.
- Elton, E., and Gruber, M. (1973). Estimating the dependence structure of share prices: Implications for portfolio selection. *Journal of Finance* 28, 5 (December): 1203–1232.
- Estrada, J. (2000). The cost of equity in emerging markets: A downside risk approach. *Emerging Markets Quarterly* 4, 2 (fall): 19–30.
- Fujiwara, Y., Souma, W., Murasato, H., and Yoon, H. (2006). Application of PCA and random matrix theory to passive fund management. In Hideki Takayasu (ed.),

- Practical Fruits of Econophysics*. Tokyo: Springer, 226–230.
- Grinold, R. C. (1992). Are benchmark portfolios efficient? *Journal of Portfolio Management* 19, 1 (autumn): 34–40.
- Haugen, R. A., and Baker, N. L. (1991). The efficient market inefficiency of capitalisation-weighted stock portfolios. *Journal of Portfolio Management* 17, 3 (spring): 35–40.
- Jagannathan, R., and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance* 58, 4 (August): 1651–1684.
- Krauz, A., and Litzenberger, R. H. (1976). Skewness preference and the valuation of risk assets. *Journal of Finance* 31, 4 (September): 1085–1100.
- Ledoit, O., and Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management* 30, 4 (summer): 110–119.
- Ledoit, O., and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* 10, 5 (December): 603–621.
- Malkiel, B., and Xu, Y. (2006). Idiosyncratic risk and security returns. Working paper (University of Texas at Dallas).
- Malkiel, B., and Xu, Y. (1997). Risk and return revisited. *Journal of Portfolio Management* 23, 3 (spring): 9–14.
- Markowitz, H. (1952). Portfolio selection. *Journal of Finance* 7, 1 (March): 77–91.
- Martellini, L. (2008). Towards the design of better equity benchmarks: Rehabilitating the tangency portfolio from modern portfolio theory. *Journal of Portfolio Management* 34, 4 (summer): 34–41.
- Martellini, L., and Ziemann, V. (2010). Improved estimates of higher-order comoments and implications for portfolio selection. *Review of Financial Studies* 23, 4 (April): 1467–1502.
- Merton, R. C. (1987). A simple model of capital market equilibrium with incomplete information. *Journal of Finance* 42, 3 (July): 483–510.
- Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics* 8, 4 (December): 323–361.
- Michaud, R. (1998). *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation*. Boston, MA: Harvard Business School Press.
- Mitton, T., and Vorkink, K. (2007). Equilibrium underdiversification and the preference for skewness. *Review of Financial Studies* 20, 4 (July): 1255–1288.
- Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13, 3 (December): 341–360.
- Rubinstein, M. E. (1973). The fundamental theorem of parameter-preference security valuation. *Journal of Financial and Quantitative Analysis* 8, 1 (January): 61–69.
- Sharpe, W. F. (1963). A simplified model for portfolio analysis. *Management Science* 9, 2 (January): 277–293.
- Tinic, S., and West, R. (1986). Risk, return and equilibrium: A revisit. *Journal of Political Economy* 94, 1 (February): 126–147.

Asset Pricing Models

General Principles of Asset Pricing

GUOFU ZHOU, PhD

Frederick Bierman and James E. Spears Professor of Finance, Olin Business School,
Washington University in St. Louis

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Asset pricing is mainly about transforming asset payoffs into prices. The most important principles of valuation are no-arbitrage, law of one price, and linear positive state pricing. These principles imply asset prices are linearly related to their discounted payoffs in which the stochastic discount factor is a function of investors' risk tolerance and economy-wide risks. The arbitrage pricing theory, the capital asset pricing model, and the consumption asset pricing model, among others, are special cases of the discount factor models.

In this entry, we discuss the general principles of asset pricing. Our focus here is to analyze asset pricing in a more general setup. Due to its generality, this entry is inevitably more abstract and challenging, but important for understanding the foundations of modern asset pricing theory. First, by extending the state-dependent contingent claims with two possible states allowing for an arbitrary number of states, we introduce the economic notions of complete market, the law of one price, and arbitrage. Then, we provide the fundamental theorem of asset pricing that ties these concepts to asset pricing relations. Subsequently, we discuss stochastic discount factor models, which is the unified framework of various asset pricing theories that include the *capital asset pricing model* (CAPM) (see Sharpe, 1964; Lintner, 1965; Mossin, 1966) and *arbitrage pricing theory* (APT) (see Ross, 1976) as special cases.

ONE-PERIOD FINITE STATE ECONOMY

If a security has payoffs, denoted by x ,

$$\tilde{x} = \begin{cases} \$1, & \text{up} \\ 0, & \text{down} \end{cases}$$

it means that the economy will have two states next period, up or down, and the security will have a value of \$1 or 0 in the up and down states, respectively. Similarly, as a simple extension, we can think that the economy has three states next period: good, normal, and bad. Then, any security in this economy must have three payoffs corresponding to the three states. For example,

$$\tilde{x} = \begin{cases} \$3, & \text{good} \\ \$2, & \text{normal} \\ \$1, & \text{bad} \end{cases}$$

is a security in the states economy with values of \$3, \$2, and \$1, respectively, in the three states. For notational brevity, we sometimes use the transposed vector dropping the dollar sign, $(3, 2, 1)'$, to denote the payoff of this security, where the apostrophe (') is the symbol for transpose.

In general, we can consider an economy with an arbitrary number of s states and N securities. In this economy, the payoff of any security can be expressed as

$$\tilde{x} = \begin{cases} v_1, & \text{State 1} \\ v_2, & \text{State 2} \\ \vdots & \vdots \\ v_s, & \text{State } s \end{cases} \quad (1)$$

where the v 's are the values of the security in the m states. For example, suppose state $s = 4$, then a security with payoff $(1.10, 1.10, 1.10, 1.10)'$ is a well-defined security in our four-state economy. Suppose further that the price of this security is \$1, then this security earns \$0.10 or 10% ($\$0.1/\1) regardless of the state. Hence, this security is risk free with a rate of return of 10% regardless of the state of the economy.

Suppose now that there is a total of N securities, $\tilde{x}_1, \dots, \tilde{x}_N$, in an economy of s states. We can summarize the payoffs next period of all the N securities by using the following matrix,

$$X = \begin{pmatrix} v_{11} & \dots & v_{1N} \\ \vdots & \ddots & \vdots \\ v_{s1} & \dots & v_{sN} \end{pmatrix} \quad (2)$$

where each of the N columns represents the values of the securities. It is evident that matrix X summarizes payoffs of all the securities and determines their future values completely.

The asset pricing question is how to determine the price for each of the securities. Mathematically, the pricing mechanism can be viewed as a mapping from the j -th security (or the s vector, the payoff obtained from owning the security), to a price p that an investor is willing to pay today,

$$\rho(\tilde{x}_j) = p_j \quad (3)$$

As it turns out, simple economic principles imply many useful properties for the mapping, which comprises the *general principles of asset pricing* to be discussed below.

PORTFOLIOS AND MARKET COMPLETENESS

In evaluating securities, a key principle is to evaluate them as a whole, and not in isolation. To do so, consider a portfolio of the N securities

$$\tilde{x}_p = \varphi_1 \tilde{x}_1 + \varphi_2 \tilde{x}_2 + \dots + \varphi_N \tilde{x}_N \quad (4)$$

where the φ 's are portfolio weights that now represent the units of the securities we purchase in the portfolio, and \tilde{x}_p is the payoff of the portfolio, which simply adds up the individual values. Note that the weights can be either positive or negative. A negative weight on a security is a short position. In the case where no short sales are allowed, the weights are restricted to be positive.

Note that the portfolio weights are often the percentages of money we invest in the securities, where prices are given and we are interested in the return on a portfolio. In contrast, we focus here on the weights in terms of units because we are interested in determining the prices from payoffs. However, once the prices are given, the weights in terms of either units or percentages are equivalent. To see this, if we express a portfolio in term of returns, denoted by R , rather than payoffs as above, then the portfolio return is

$$R_p = w_1 R_1 + w_2 R_2 + \dots + w_N R_N \quad (5)$$

where

$$R_j = \frac{\tilde{x}_j}{p_j}$$

is the gross return on security j , which is one plus the usual percentage return. The relation between the φ 's and the w 's is

$$w_j = \frac{\varphi_j p_j}{\varphi_1 p_1 + \dots + \varphi_N p_N} \quad (6)$$

where the numerator is the amount of money allocated to security j , and the denominator is the total amount of money invested in the securities, so that the w 's are the percentage weights as before.

Consider the following two securities in a two-state economy:

$$\tilde{x}_1 = \begin{cases} 1, & \text{up} \\ 0, & \text{down} \end{cases}, \quad \tilde{x}_2 = \begin{cases} 0, & \text{up} \\ 1, & \text{down} \end{cases}$$

Suppose their prices today are \$1. Then, with an investment of \$1 that buys 0.5 unit each of the securities, one obtains a portfolio

$$\tilde{x} = \varphi_1 \tilde{x}_1 + \varphi_2 \tilde{x}_2 = 0.5\tilde{x}_1 + 0.5\tilde{x}_2$$

with payoff

$$\tilde{x} = \begin{cases} 0.5, & \text{up} \\ 0.5, & \text{down} \end{cases}$$

One can also buy 2 units of the first security, and short one unit of the second security; then the resulting portfolio is

$$\tilde{x} = 2\tilde{x}_1 + (-1)\tilde{x}_2$$

with payoff

$$\tilde{x} = \begin{cases} 2, & \text{up} \\ -1, & \text{down} \end{cases}$$

Note that the payoff of the portfolio is negative, $-\$1$, in the down state. This means that when the economy is down, one has to buy back the second security at a price of \$1 (its value in the down state) to cover the short position. The net cost is \$1, the payoff of the portfolio in the down state. In contrast to the portfolio where equal dollar amounts are invested in both securities, this portfolio with short sales permitted has a higher payoff of \$2 in the up state, which compensates for the loss in the down state.

Redundant Assets

A portfolio is uniquely determined by its portfolio weights, which can be summarized by the N -vector

$$\varphi = (\varphi_1, \varphi_2, \dots, \varphi_N)'$$

The portfolio's payoffs are then uniquely determined by the s -vector,

$$\text{Payoff} = X\varphi \quad (7)$$

For example, one can easily verify that this is true in our first illustration in which X is simply equal to the identity matrix.

A portfolio φ is said to be replicable if we can find another portfolio with different weights, ω , such that their payoffs are equal

$$X\omega = X\varphi, \quad \omega \neq \varphi \quad (8)$$

In particular, if one of the x 's can be replicated by a portfolio of others, it is called a *redundant asset* or *redundant security*. In any economy, redundant securities can be eliminated without affecting the properties of all the possible portfolios of the remaining assets. Sometimes, in order to distinguish the securities, the x 's that define the economy, and all their possible portfolios, we will refer to the x 's as *primitive securities* because all other portfolios are composed of them.

Consider the following two-state economy

$$\tilde{x}_1 = \begin{cases} 1, & \text{up} \\ 0, & \text{down} \end{cases}, \quad \tilde{x}_2 = \begin{cases} 2, & \text{up} \\ 0, & \text{down} \end{cases}$$

with prices for both securities being \$1 and \$2 today. The portfolio with weight vector $\varphi = (0.5, 0.5)'$ is

$$\tilde{x} = 0.5\tilde{x}_1 + 0.5\tilde{x}_2$$

This portfolio is replicable because it is also equal to

$$\tilde{x} = 1.5\tilde{x}_1$$

The primitive asset x_2 is redundant here because its payoff is simply double the payoff of the first asset.

Complete Market

In an economy with N risky securities and s states, a security market is formed if arbitrary buying and shorting are allowed, which creates infinitely many possible portfolios. We say the

market is *complete* and is hence referred to as a *complete market*, if, for any possible payoff, there is a portfolio of the primitive securities to replicate it. That is, for any desired payoff \tilde{x} , we can find portfolio weights such that

$$\varphi_1 \tilde{x}_1 + \varphi_2 \tilde{x}_2 + \cdots + \varphi_N \tilde{x}_N = \tilde{x} \quad (9)$$

A complete market not only allows investors to obtain any desired payoff in any state (with a price), but also permits unique security pricing, as will be clear later.

For example, the two securities in our first example will form a complete market. This is because for any possible payoff

$$\tilde{x} = \begin{cases} a, & \text{up} \\ b, & \text{down} \end{cases}$$

the portfolio

$$a \tilde{x}_1 + b \tilde{x}_2$$

yields the payoff. To see why, if one investor wants to get a \$2 payoff in the up state and \$3 in the down state, buying 2 units of the first security and 3 units of the second security will provide what is exactly desired. However, the two securities in our second example above form an incomplete market. This is because for any possible portfolios consisting of the two securities, it will be impossible to create a payoff of \$1 in the down state.

In terms of matrix and vector notation, a complete market requires that, for any payoff vector, we can find portfolio weights φ to solve the linear equation with φ as the unknown variable

$$X\varphi = y \quad (10)$$

Note that X is an s by N matrix and y is an s -vector. Recall from linear algebra that the number of independent columns of X is called the *rank* of the matrix X , denoted as $\text{rank}(X)$ below. If $\text{rank}(X) = s$, the linear combinations of these columns will generate all possible s -vectors. That is, a portfolio of those securities whose payoffs are those independent columns is capable of producing any possible payoffs, or the market must be complete. Conversely,

if the above linear equation has a solution to any y , it must do so for s independent y 's, say, the s columns of the s -dimensional identity matrix, which is an s by s matrix with diagonal elements 1 and zero elsewhere. For example, if $s = 2$, the y 's correspond to the payoffs of the two securities in our first example. This means that the linear combinations of the columns of X are capable of yielding s independent columns. So, the number of independent columns must be greater than or equal to s . Since X is an s by N matrix, its number of independent columns, $\text{rank}(X)$, cannot be greater than s . Then the only possibility is equal to s .

We can summarize our discussion in the following proposition:

Market Completeness Proposition: The market is complete if and only if the rank of the s by N payoff matrix X is s , that is,

$$\text{rank}(X) = s \quad (11)$$

Consequently, for s possible states, we should have at least $N \geq s$ primitive assets for the market to be complete. One can verify that the rank condition holds for the two securities in our first example, but not in our second example.

THE LAW OF ONE PRICE AND LINEAR PRICING

In this section, we first discuss the law of one price and its relation to the linear pricing rule, and then introduce the concept of state price and relate it to the law of one price.

Linear Pricing

The *law of one price* (LOP) says that two assets with identical payoffs must have the same price. In international trade, in the absence of tariffs and transportation costs, an apple sold in New York City must have the same price as an apple sold in London after converting the money into the same currency. This provides an economic

channel through which to tie the currencies together. In the financial markets, the LOP says that we should not be able to profit from buying the same security at a higher price and selling it at a lower one.

Mathematically, under LOP, if two portfolios have the same payoffs

$$X\varphi = X\omega \quad (12)$$

then their prices today must be the same

$$\rho(X\varphi) = \rho(X\omega) \quad (13)$$

where, as we recall from our earlier discussion, ρ is the mapping that maps the payoff of an asset or of a portfolio into its price.

A simple necessary and sufficient condition for the LOP to hold is that every portfolio with zero payoff must have zero price. To see the necessity, suppose that there is an asset with zero payoff that sells at a nonzero price, say, \$0.01. We can combine this asset with any other asset to form a new asset without changing the payoff, but the price of this new asset is \$0.01 higher than before packaging the two assets. The LOP says that the old one and the new one must have the same price, which is, of course, a contradiction. Conversely, if two portfolios with an identical price were sold at different prices, say \$2.01 and \$2, buying the one with the price of \$2.01 and shorting the one with a price of \$2 creates an asset with zero payoff, but a price of \$0.01. This is not possible from the zero price condition.

The LOP essentially prevents an asset from having multiple prices, which gives rise to its name. Only when it is true is it possible for there to be rational pricing with a unique price. An important theoretical implication of the LOP is that the price mapping, the ρ function, must be linear:

$$\rho[X(a\varphi + b\omega)] = a\rho(X\varphi) + b\rho(X\omega) \quad (14)$$

That is, the price of a portfolio must be equal to a portfolio of the component prices. Intuitively, the price of two burgers must be two times the price of one, and the price of a burger

and a Coke must be the same as the sum of the two individual prices. The linear pricing rule is fundamental in finance. It implies that, if the share price of a company is its future cash flows, then no matter how one slices the cash flows, the price will remain unchanged and is equal to the values of the slices added together.

The linear pricing rule clearly implies the LOP. The price mapping is uniquely determined by the payoffs only, and so it must be the case that the prices are identical if the payoffs are. Conversely, if the LOP is true, paying the price of the left-hand side of equation (14) will result in a portfolio with the identical payoff as the right-hand side, and hence their prices must be the same. A formal statement of this is as follows:

Linear Pricing Rule: The law of one price is valid if and only if the linear pricing rule is true.

State Price

In asset pricing, the concept of a state price is fundamental. In our states economy, there are s states. The *state price* in state i is the price investors are willing to pay today to obtain one unit of payoff in that state, and nothing in other states. The state price is also known as the *Arrow-Debreu price*, named in honor of the originators. A *state price vector* will then be an s -vector of all the prices in all the states. If there exists a state price vector $q = (q_1, q_2, \dots, q_s)$, then we can write the asset price for each primitive security as

$$p_j = q_1 v_{1j} + q_2 v_{2j} + \dots + q_s v_{sj} \quad (15)$$

In words, this equation says that the price of the j -th security is equal to its payoffs in each of the states times the price per unit value in that state.

The state price is not only useful for linking the payoffs of the primitive securities to their prices, but also useful to price any new assets, including any other contingent claims or derivatives in the economy. All we need to do is

to identify the payoffs of these assets and then sum the products of the payoffs with their state prices to obtain asset prices.

The question is whether the state price vector always exists. We rewrite the state pricing relation (15) in matrix form as

$$p = X'q \quad (16)$$

The existence of the state price vector q is the existence of solution q to the linear equation given by (16). In our states economy here, we can show that the LOP is necessary and sufficient for the existence of the state price, while in more complex economies, say those with an infinite number of assets and an infinite number of states, some auxiliary condition may be needed.

Existence of State Price Condition: The law of one price is valid if and only if the state price vector exists.

The proof of the above follows from linear algebra. If the state price vector exists, then

$$p'\varphi = q X'\varphi = q X'\omega = p'\omega$$

which says that the price of the portfolio with weights φ is the same as the price of another portfolio as long as their payoffs are identical. Conversely, if the LOP is true, then for any portfolio weights w with zero payoff or satisfying $X'\varphi = 0$, we must have zero price or $p'\varphi = 0$. This means that p is orthogonal to every vector that is orthogonal to X . Now projecting p on the entire N -dimensional space, p must then be a linear combination of the columns of X . The combination coefficients are exactly equal to q , which is what we are looking for. The proof is therefore complete.

As an example, consider the following two securities in a two-state economy,

$$\tilde{x}_1 = \begin{Bmatrix} 1, & up \\ 0, & down \end{Bmatrix}, \quad \tilde{x}_2 = \begin{Bmatrix} 2, & up \\ 0, & down \end{Bmatrix}$$

where the first security has a price of \$1 and the second of \$2. Clearly the prices are consistent with the LOP. In this case, a state price of $(1, 0)'$

can price all portfolios of the two securities:

$$1 = 1 \times 1 + 0 \times 0$$

and

$$2 = 1 \times 2 + 0 \times 0$$

Another state price $(1, 2)'$ can also do the same. A more subtle case is in an economy when

$$\tilde{x}_1 = \begin{Bmatrix} 1, & up \\ 1, & down \end{Bmatrix}, \quad \tilde{x}_2 = \begin{Bmatrix} 2, & up \\ 2, & down \end{Bmatrix}$$

with the same prices of \$1 and \$2. Then $(0.5, 0.5)'$ and $(0.2, 0.8)'$ both, among others, price the two primitive securities and all their portfolios correctly.

Under what conditions will the state price be unique? To find the conditions, recall the matrix form of the state pricing relation

$$p = X'q$$

The LOP is equivalent to the existence of the state price vector q . If the market is in addition complete, then q in the above equation can be uniquely solved as

$$q = (XX')^{-1}Xp \quad (17)$$

Note that X is s by N , so its inverse is undefined unless $s = N$. But the inverse of the s by s matrix, XX' , is well defined. Equation (17) leads to our next proposition.

Uniqueness of State Price Proposition: If the law of one price holds, and if the market is complete, the state price must exist and be unique.

For example, consider the following two securities in a two-state economy

$$\tilde{x}_1 = \begin{Bmatrix} 1, & up \\ 2, & down \end{Bmatrix}, \quad \tilde{x}_2 = \begin{Bmatrix} 3, & up \\ 4, & down \end{Bmatrix}$$

where the first security has a price of \$4 and the second of \$10. We can check that both the rank and LOP conditions are true. The unique state price vector is then given by equation (17),

$$\begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \begin{pmatrix} 5 & -3.5 \\ -3.5 & 2.5 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} \begin{bmatrix} 4 \\ 10 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

It can be verified that these prices indeed work for pricing the two primitive securities.

ARBITRAGE AND POSITIVE STATE PRICING

The assumption of the absence of arbitrage is the foundation upon which asset pricing theories rely. When there are any free lunches or what economists refer to as *arbitrage opportunities*, asset prices are not rational. Investors are likely to be able to correct the prices by exploiting the arbitrage opportunities, and eventually these opportunities will disappear, and the prices will reflect their true values. Asset pricing theory is largely concerned with these equilibrium true values.

In our states economy, the concept of arbitrage can be formally defined. There are two types of arbitrage. The first type exists if there is a portfolio strategy that requires no investment today (i.e., referred to earlier as a zero-investment strategy) and yet yields nonnegative payoffs in the future, and positive (or not identical to zero) at least in one of the states. Mathematically, this type of arbitrage can be expressed as

$$X\varphi \geq 0, \text{ and not equal to zero}$$

with

$$p_1\varphi_1 + p_2\varphi_2 + \cdots + p_N\varphi_N \leq 0$$

The second type of arbitrage is one in which a portfolio strategy earns money today, and yet has no future obligations. We can express this mathematically as follows:

$$X\varphi \geq 0$$

with

$$p_1\varphi_1 + p_2\varphi_2 + \cdots + p_N\varphi_N < 0$$

Consider as an example the following two securities in a two-state economy:

$$\tilde{x}_1 = \begin{cases} 1, & \text{up} \\ 2, & \text{down} \end{cases}, \quad \tilde{x}_2 = \begin{cases} 2, & \text{up} \\ 4.1, & \text{down} \end{cases}$$

with prices \$1 and \$2. If we follow a strategy that involves shorting two units of the first security and buying one unit of the second security, then our net investment will be zero, but the payoffs will be

$$-2 \times \tilde{x}_1 + 1 \times \tilde{x}_2 = \begin{bmatrix} 0 \\ 0.1 \end{bmatrix}$$

This is an arbitrage of the first type. However, there is no arbitrage of the second type. This is because for any weights φ_1 and φ_2 , if the cost is negative, that is,

$$\varphi_1 + 2\varphi_2 < 0$$

then the payoff in the up state of the portfolio,

$$\varphi_1 + 2\varphi_2$$

will be negative too.

To illustrate, consider the following two securities in a two-state economy,

$$\tilde{x}_1 = \begin{cases} 1, & \text{up} \\ -1, & \text{down} \end{cases}, \quad \tilde{x}_2 = \begin{cases} 2, & \text{up} \\ -4, & \text{down} \end{cases}$$

with prices \$1 and \$1.9. If we short two units of the first security and buy one unit of the second security, then our net investment will be

$$(-2) \times 1 + 1 \times 1.9 = -0.1$$

but the payoffs will be

$$-2 \times \tilde{x}_1 + 1 \times \tilde{x}_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

This is an arbitrage of the second type. However, there is no arbitrage of the first type. This is because for any weights φ_1 and φ_2 , the arbitrage requires the portfolio payoffs be nonnegative

$$\varphi_1 + 2\varphi_2 \geq 0$$

and

$$-\varphi_1 - 2\varphi_2 \geq 0$$

in the two states, respectively. The only non-negative payoffs for both the states is the zero payoff in this case. So, there cannot be an arbitrage of the first type.

Note that the pricing operator is to map the payoffs of an asset to its price, and it provides

that the state price of the payoff is \$1 in that state and nothing in other states. If the state price in one state is zero, this will be clearly an arbitrage opportunity as an investor can get future payoffs in this state for paying a zero price today. To rule out arbitrage opportunities in the economy, it is hence necessary to require the state prices be positive. When the pricing operator is both linear and implying positive state prices, we call it a positive linear pricing rule. As it turns out below, the existence of such a positive linear pricing rule is equivalent to the absence of arbitrage opportunities in the economy.

Arbitrage is also related to the LOP. If there is no arbitrage, the LOP must be true. This is because if two portfolios with two identical payoffs were sold at different prices, a “buy low and sell high” strategy will result in the construction of a portfolio with zero payoffs in the future, but with positive proceeds today. This is an arbitrage of the first type. Thus, the no-arbitrage condition is stronger than the LOP. In finance, the assumption of no arbitrage is crucial, as explained next by the fundamental theorem of asset pricing.

THE FUNDAMENTAL THEOREM OF ASSET PRICING

Consider now an investor’s utility maximization problem. Assume the investor prefers more to less, so that the utility function is monotonic in the consumption level. Given an initial wealth W_0 , and given the trading opportunities, the investor’s future consumption, as a vector in the s states, will be

$$C_1 = W_1 + (W_0 - C_0) \times R_p$$

where

C_0 = consumption (measured in dollars) today,

R_p = return on a portfolio of assets, which can be optimally chosen by the investor maximizing his or her utility, and
 W_1 = the investor’s income from other sources next period, such as labor income

The utility is a monotonic function of both C_0 and C_1 .

Then the following theorem ties together the no-arbitrage, positive linear pricing rule, and the utility maximization problem.

Fundamental Theorem of Asset Pricing: The following are equivalent:

1. Absence of arbitrage
2. Existence of a positive linear pricing rule
3. Existence of an investor with monotonic preference whose utility is maximized

We provide a simplified proof here. (A more rigorous proof is provided in Dybvig and Ross (1987).)

To see that the absence of arbitrage implies existence of a positive linear pricing rule, we note first that earlier we provided the argument for the existence of the linear pricing rule. The positivity of the state prices must be true in the absence of arbitrage. This is because, if there is a zero or negative state price in some state, then the payoffs in this state are free lunches, so arbitrage opportunities can arise. Conversely, if the state prices are positive, every single payoff in each state has a positive price, and there cannot be any free lunch.

Mathematically, this can also be easily demonstrated. If φ is an arbitrage portfolio so that its price is zero or negative, then

$$0 \geq p'\varphi = (X'q)'\varphi = q'(X\varphi)$$

where the first equality is the linear pricing rule, and the second equality holds by matrix multiplication rules. Because of positive state prices, all components of q are positive. If $p'\varphi$ is zero, $X\varphi$ must be all zeros, and if $p'\varphi$ is negative, $X\varphi$ must have strictly negative components. Both contradict the assumption that φ is an arbitrage portfolio. Hence, there are no arbitrage opportunities when the state prices are positive.

To see how the existence of a positive linear pricing rule implies the existence of an investor with monotonic preference whose utility is maximized, the consumption of the investor in each state must be finite since the investor has finite wealth, and since the investor faces a binding budget constraint due to positive state prices. Finally, the existence of an investor with monotonic preference whose utility is maximized clearly implies the absence of arbitrage. This is because adding an arbitrage portfolio (i.e., a free lunch) to the investor's portfolio will only strictly increase his or her utility without affecting the budget, contradicting the fact that the utility is maximized to begin with. This concludes our proof.

An important insight from the fundamental theorem is what we need for rational pricing. In deriving pricing formulas, many theoretical equilibrium asset pricing models assume all investors behave rationally and have identical information sets. The theorem says that, to rationally price assets or to ensure market pricing efficiency, we do not need to assume that all investors are smart. What we need is a few smart ones who can capitalize on any arbitrage opportunities. Then, the prices should be in line with their payoffs in the economy.

The Discount Factor

Related to the fundamental theorem is the concept of the discount factor. As it turns out, this is the common feature of almost all asset pricing models, a point that will become evident in the next section. Let $\theta_i > 0$ be the probability for state i to occur. The linear pricing rule given by equation (15) can be rewritten as

$$p_j = \theta_1(q_1/\theta_1)v_{1j} + \theta_2(q_2/\theta_2)v_{2j} + \cdots + \theta_s(q_s/\theta_s)v_{sj} = E(mv_j) \quad (18)$$

where m is a random variable whose value in state s is equal to

$$m_s = \frac{q_i}{\theta_s} \quad (19)$$

Equation (18) says that the price for asset j is given by the expected value of its payoff multiplied by a random variable m , where m is common for all assets.

Suppose now that there is a risk-free asset in the economy that can earn a risk-free interest rate r , and that the price of this risk-free asset today is \$1 (we can scale the asset unit if necessary). Then the payoff of this risk-free asset's price in the next period will be $1 + r$ in all the states. So, by equation (18), we have for the following expected payoff for this risk-free asset

$$1 = E[m(1 + r)]$$

and therefore

$$E[m] = \frac{1}{1 + r} \quad (20)$$

If there were no risks in the economy, and if there were no arbitrage, it is clear that all assets should earn the same risk-free rate of return. Hence, assets should be priced by their present values of the cash flows, or the prices are equal to the discounted cash flows with the discount factor $1/(1 + r)$. When there is risk as is the case now, the payoffs are multiplied by the random variable m whose mean is $1/(1 + r)$. This is why m is also known as a stochastic discount factor because (1) it is random, and (2) it extends the risk-free discounting to the risky asset case.

Consider, for example, three securities in a three-state economy with prices \$5, \$5, and \$6, and with the following payoff matrix:

$$X = \begin{pmatrix} 10 & 20 & 30 \\ 10 & 10 & 10 \\ 10 & 5 & 5 \end{pmatrix}$$

In this economy, the first asset is the risk-free asset since it has a constant payoff of \$10 regardless of the future state. Moreover, the risk-free rate is 100% because the asset is sold at a price of \$5. The state price vector can be solved using equation (17) and is $q = (0.1, 0.2, 0.2)'$. Assume the probability for each state is $1/3$. Then the

linear pricing rule can be expressed as

$$\begin{aligned} 5 = p_1 &= \frac{1}{3} \times (0.3 \times 10) + \frac{1}{3} \times (0.6 \times 10) \\ &\quad + \frac{1}{3} \times (0.6 \times 10), \\ 5 = p_2 &= \frac{1}{3} \times (0.3 \times 20) + \frac{1}{3} \times (0.6 \times 10) \\ &\quad + \frac{1}{3} \times (0.6 \times 5), \\ 6 = p_3 &= \frac{1}{3} \times (0.3 \times 30) + \frac{1}{3} \times (0.6 \times 10) \\ &\quad + \frac{1}{3} \times (0.6 \times 5) \end{aligned}$$

Let m be a random variable that has values 0.3, 0.6, and 0.6 in the three possible states. Then the above says that, for each asset, the price is the expected value of the discounted payoff. The mean of the discount factor is

$$\begin{aligned} E[m] &= \frac{1}{3} \times 0.3 + \frac{1}{3} \times 0.6 + \frac{1}{3} \times 0.6 = 0.5 \\ &= \frac{1}{1 + 100\%} \end{aligned}$$

This verifies equation (19).

The state price vector, or equivalently the discount factor, is not only useful for pricing primitive assets, but also useful to price any portfolio consisting of them, as well as derivatives. For example, consider a call option that grants the owner of the option the right to buy one unit of the second asset at a price of \$10. This option will have a value in state 1 equal to \$10 (the price of the second asset in state 1 reduced by the price that must be paid to acquire asset 1 as provided for by the option, \$10). The value of the option is therefore \$10, the difference between \$20 – \$10 in state 1). In the other two states, the value of the option is zero because the payoff (i.e., the price of the second asset) is no greater than \$10. Hence, it would not be economic for the owner of the option to exercise. Then the price of this call option is

$$\begin{aligned} \text{Price of Call} &= \frac{1}{3} \times (0.3 \times 10) + \frac{1}{3} \times 0 \\ &\quad + \frac{1}{3} \times 0 = 1 \end{aligned}$$

The discount factor prices the assets by taking the expectation under the true probabilities.

Pricing Using Risk-Neutral Probabilities

Alternatively, one can also price the assets under a probability measure known as the *risk-neutral probabilities*. The approach is especially useful for pricing derivatives. The reason is that the risk-neutralized payoffs are easier to determine, while the solution of the discount factor is more complex.

To see how the risk-neutral approach works here, we apply the linear pricing rule given by equation (18) to the risk-free asset. We have:

$$1 = q_1(1 + r) + q_2(1 + r) + \cdots + q_s(1 + r)$$

so that

$$q_1 + q_2 + \cdots + q_s = \frac{1}{1 + r} = q$$

which says the sum of state prices must be equal to the present value of \$1 today. Denote by q the sum of the individual q 's. Since now all the state prices are positive, the ratio of each to q can be considered a probability. Since the ratios sum to one, the probability is well defined. However, this is not the original true probability of the states, but rather some artificial probability, which will be useful in the future for pricing derivatives and other assets.

Suppose now, without loss of generality, that the risk-free asset is the first one. Then the pricing relations for the other assets are

$$\begin{aligned} p_j &= q_1 v_{1j} + q_2 v_{2j} + \cdots + q_s v_{sj} \\ &= \frac{1}{1 + r} \left(\frac{q_1}{q} v_{1j} + \frac{q_2}{q} v_{2j} + \cdots + \frac{q_s}{q} v_{sj} \right) \\ &= \frac{1}{1 + r} E^Q[v_j] \end{aligned} \quad (21)$$

that is, the price is the present value discounted at the risk-free rate of the risk-adjusted expected payoff of the asset, where E^Q denotes the expectation taken under the artificial probability. In other words, for any risky asset, we compute

its value in two steps. In the first step, the risk-neutralized payoff is calculated. In the second step, treating this payoff as riskless, the payoff is discounted at the risk-free rate to obtain the price. Consequently, the artificial probability is also often referred to as the *risk-neutral probability measure*.

For example, for the assets in our previous example, the sum of the state prices is

$$0.1 + 0.2 + 0.2 = \frac{1}{1 + 100\%} = 0.5$$

Moreover, the risk-neutral probabilities are $1/5$, $2/5$, and $2/5$. So the expected payoff of the earlier call option is

$$E^Q(\text{call}) = \frac{1}{5} \times 10 + \frac{2}{5} \times 0 + \frac{1}{5} \times 0 = 2$$

Discounting the \$2 at the risk-free rate (100% in our example), we get the price of \$1 (= \$2/(1 + 1)). This price is, of course, the same as computed above using the discount factor to price the call option.

DISCOUNT FACTOR MODELS

In this section, we provide the discount factor models in a more general setup by allowing the asset returns to be arbitrarily distributed, not necessarily finite states as in the previous section. Then we derive a lower bound on the variance of all possible discount factors, known as the Hansen-Jagannathan bound, and apply it to analyze the implications of some important theories in financial economics.

STOCHASTIC DISCOUNT FACTORS

Consider now a more general problem of an investor who is interested in maximizing utility over the current and future values of consumption,

$$U(C_t, C_{t+1}) = u(C_t) + \delta E[u(C_{t+1})]$$

where the first term is the utility of consumption today, the second term is the utility of fu-

ture consumption, and δ is the subjective time-discount factor of the investor that captures the investor's trade-off between current and future consumption. Note that the second term has an expectation operation since future consumption is unknown today, and the investor can only maximize the expected utility with the expectation taken over all possible random realizations of the future consumption.

Besides the quadratic utility, another popular form of utility function is the power utility

$$u(C_t) = \frac{C_t^{1-\gamma}}{1-\gamma}$$

where γ is the risk-aversion coefficient. The higher the γ , the more risk averse the investor. Typically, a value of γ of about 3 is believed to be reasonable.

For notational brevity, we assume there is only one risky asset, which the pricing relation developed holds for an arbitrary number of assets by adding them into the model. Unlike earlier sections in this entry where finite payoffs were assumed, we now assume the payoff of the risky asset can have an arbitrary probability distribution, so long as the expectation is well defined. The budget constraints for maximizing the utility can be written as

$$\begin{aligned} C_t &= W_t - p_t w \\ C_{t+1} &= W_{t+1} + X_{t+1} w \end{aligned}$$

where W_t and W_{t+1} are the investor's wealth from other sources, w is the number of units of the risky asset the investor purchases today at time t , p_t is the security price, and X_{t+1} is the payoff.

Plugging the budget constraints into the utility function, and taking the derivative with respect to w , we obtain the first-order condition (FOC):

$$p_t u'(C_t) = E_t[\delta u'(C_{t+1}) X_{t+1}]$$

or

$$p_t = E_t[m X_{t+1}], \quad m = \delta \frac{u'(C_{t+1})}{u'(C_t)} \quad (22)$$

This equation says that the price today is the expected value of the discounted payoff, and m

is the discount factor. In the case of the power utility,

$$m = \delta \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} \quad (23)$$

which is a power function of the consumptions.

What we have derived in equation (22) is called a *consumption-based asset pricing model*, so named because the theory is motivated from the perspective of consumption. This motivation is different from the earlier no-arbitrage arguments that yield equation (18). However, the pricing equations have the same form, except that the discount factor now takes a new specification. Indeed, most, if not all, asset pricing models are of the discount factor form, and different theories may specify the m differently. For the particular specification of m given by equation (22), it is also known as the marginal rate of substitution because it is the ratio of the marginal utilities.

Intuitively, when the marginal rate of substitution is high, the value of future consumption will be high, and an investor is willing to pay more for the asset if the asset's payoff is high in this case. This is why the price, as given by equation (22), is high.

The discount factor representation of asset prices is often also expressed in terms of returns. Let R_t be the gross return on the asset where the gross return is equal to one plus the return. That is, $R_t = X_{t+1}/p_t$. Then the pricing relation in equation (22) is equivalent to

$$1 = E_t[mR_{t+1}] \quad (24)$$

If an asset price is scaled to be equal to \$1, the payoff will be its return, and then the expected discounted return must be equal to \$1, its price today. When there are N risky assets, we can write the discount factor model as

$$1 = E_t[mR_{j,t+1}] \quad (25)$$

where $R_{j,t+1}$ is the return on the asset j .

Note that the expectation in equation (25) is conditional on all available information and therefore the pricing relation is known as the

conditional form of the discount factor model. Taking expectation on both sides of equation (25), we obtain

$$1 = E[mR_{j,t+1}] \quad (26)$$

which is known as the *unconditional form of the discount factor model*. Since conditional implies unconditional, and the reverse is not necessarily true, equation (26) is a weaker form of the model.

Application to CAPM and APT

To see the generality of the discount factor model, consider now its relation to the two dominant equilibrium asset pricing models: the CAPM and APT. As explained shortly, one can write these two asset pricing models as follows:

$$E[R_j] = \tau + \lambda_1 \beta_{j1} + \cdots + \lambda_K \beta_{jK} \quad (27)$$

where R_j is the gross return on asset j , β_{jk} is the beta or risk exposure on the k -th factor f_k , λ_k is the factor risk premium, for $k = 1, 2, \dots, K$, and τ is a constant.

Although equation (27) is now written out in terms of the gross returns to conform with discount factor notations, it can be reduced to have exactly the same expression in terms of returns. For example, the CAPM specifies $K = 1$, τ as the gross risk-free rate $1 + r$, $\lambda_1 = E[R_m] - 1 - r$, and R_m is the gross return on the market portfolio. In this case, λ_1 is same as the usual market return in excess of the risk-free rate since the ones in their difference will be canceled out.

We claim that if, and only if, the stochastic discount factor is a linear function of the factors

$$m = a + b_1 f_1 + \cdots + b_K f_K \quad (28)$$

we will obtain equation (27). Conversely, if equation (27) is true, the discount factor must be a linear function of the factors. Therefore, the CAPM and APT are special cases of the discount factors models.

To see why, it is sufficient to analyze the case of $K = 1$. For simplicity, we drop the subscripts

so that we want to show

$$m = a + bf \quad (29)$$

and

$$E[R_j] = \tau + \lambda\beta_j \quad (30)$$

are equivalent. The latter is often referred to as a *beta pricing model*. In the proof below, we can assume $E[f] = 0$ since we can always move the mean of f into a . Recall the simple statistical formula that the covariance between any two random variables can be written as a sum of the expectation of their product and the product of their expectations

$$\text{Cov}(x, y) = E[xy] + E[x]E[y] \quad (31)$$

Using this formula and $E[f] = 0$, we have, if equation (29) is true,

$$\begin{aligned} 1 &= E[mR_j] = aE[R_j] + bE[fR_j] \\ &= aE[R_j] + b\text{Cov}(R_j, f) - bE[R_j]E[f] \\ &= aE[R_j] + b\text{Cov}(R_j, f) \end{aligned}$$

Solving for $E[R_j]$, we obtain

$$E[R_j] = \frac{1}{a} - \frac{b}{a}\text{Cov}(R_j, f) \quad (32)$$

Comparing this equation with equation (30), it follows that

$$\tau = \frac{1}{a}, \quad \lambda = -\frac{b}{a}\sigma^2(f) \quad (33)$$

where $\sigma^2(f)$ is the variance of the factor. Hence, if the discount factor model is true, it must imply the beta pricing model. Conversely, if the beta pricing model is true, we can solve a and b from equation (33) to get the discount factor model.

Hansen-Jagannathan Bound

As we discussed, an asset pricing model is a specification of the discount factor. The question is what properties all the possible discount factors m must have. Hansen and Jagannathan (1991) show that the variance of the discount factors has to be bounded below. In other

words, m must be volatile enough with respect to the asset returns to be priced.

The discount factor relation, equation (26), ties the return R_t of an asset to its price via the expectation of its product with m . It will be useful to separate R_t out to understand further the relation between m and R_t . Again using the covariance formula, equation (31), we have

$$1 = \text{Cov}[m, R_{t+1}] + E[m]E[R_{t+1}] \quad (34)$$

Suppose that a risk-free asset with gross return $R_f = 1 + r$ is available, where r is the usual risk-free rate. Applying equation (34) to the risk-free asset, the first term will be zero, and hence

$$E[m] = \frac{1}{1+r} \quad (35)$$

Note that this equation is true for all possible discount factors and is an extension of earlier equation (20). In other words, for all possible stochastic discount factors, their mean must be equal to $1/(1+r)$ to price the risk-free asset.

Now we multiply equation (34) by R_f on both sides, and obtain

$$E[R_{t+1}] - R_f = -R_f\text{Cov}[m, R_{t+1}]$$

This says that an asset's return in excess of the risk-free rate will be higher if it has a larger negative covariance with m . Recall that the covariance is related to correlation and standard deviations by

$$\text{Cov}[x, y] = \sigma(x) \times \sigma(y) \times \text{Corr}(x, y)$$

where $\sigma(\cdot)$ denotes the standard deviation function. Since the correlation is always between -1 and 1 , we have from the earlier equation that

$$\begin{aligned} |E[R_{t+1}] - R_f| &= R_f|\text{Cov}[m, R_{t+1}]| \\ &\leq R_f \times \sigma(m) \times \sigma(R_{t+1}) \end{aligned}$$

Separating terms on m from those on R_{t+1} , we have a lower bound on the standard deviation of m as denoted by $\sigma(m)$

$$\frac{\sigma(m)}{E[m]} \geq \frac{|E[R_{t+1}] - R_f|}{\sigma(R_{t+1})} \quad (36)$$

The right-hand side, the ratio of the expected return on a risky asset to its standard deviation, is the Sharpe ratio that measures the extra return beyond the risk-free rate per unit of asset risk. The relationship given by equation (36) says that, for any discount factor that prices the assets, it must have enough variability so that its standard deviation divided by its mean must be greater than the Sharpe ratio of any risky asset in the economy.

The above lower bound on $\sigma(m)$ is known as the *Hansen-Jagannathan bound*. It is an important result since if an asset pricing model fails to pass this bound, then the proposed asset pricing model can be rejected. For example, to test the validity of either the discount factor model given by equation (18) for a finite state economy, or the consumption-based asset pricing given by equation (22), or the CAPM and the APT, one can test first whether it passes the bound given by (36). No further testing will be necessary if it fails the Hansen-Jagannathan bound. Theoretically, Kan and Zhou (2006) show that the Hansen-Jagannathan bound can be tightened substantially with the use of information on the state variables of the stochastic discount factor.

KEY POINTS

- A complete market is one in which any desired payoff in the future can be generated by a suitable portfolio of the existing assets in the economy.
- In a world where the number of states (future scenarios) is finite, a market is complete if and only if this number is equal to the rank of the asset payoff matrix. In particular, it is necessary for the number of assets to be greater than the number of states.
- The law of one price states that any two assets with identical payoffs in the future must have the same price today.
- A linear pricing rule means that the price of a basket of assets is equal to the sum of the prices of those assets in the basket. The law of one price is true if and only if the linear pricing rule is true.
- The state price is the price one has to pay today to obtain a one dollar payoff in a particular future state and nothing in other states. The existence of the state price is equivalent to the validity of the law of one price. It will be unique if the market is in addition complete.
- There are two types of arbitrage opportunities. The first is paying nothing today and obtaining something in the future, and the second is obtaining something today with no future obligations.
- The fundamental theorem of asset pricing asserts the equivalence of three key issues in finance: (1) absence of arbitrage; (2) existence of a positive linear pricing rule; and (3) existence of an investor who prefers more to less and who has maximized utility (no more free lunches to pick up from the economy).
- Due to risk, a rational investor will not pay a price equal to the expected value of an asset and will instead discount it by a suitable factor for compensation for taking on the risk. A stochastic discount factor is a random variable such that the expected value of its product with the asset payoffs is the rational price of the asset. The stochastic discount extends the risk-free discounting (time value of money) to the risky asset case and is the same for pricing all the assets in the economy.
- The CAPM and APT are special cases of stochastic discount factor models in which the discount factor is a linear function of the market factor or APT factors. Moreover, almost all asset models can be formulated as stochastic discount factor models.
- The Hansen-Jagannathan bound provides a simple bound on the variance of a stochastic discount factor, so that one can examine whether the stochastic discount factor satisfies some basic restrictions on the data.

If not, we can reject it without further analysis.

REFERENCES

- Dybvig, P. H., and Ross, S. A. (1987). Arbitrage. In J. Eatwell, M. Milgate, and P. Newman (eds.), *A Dictionary of Economics: Vol. 1*. Macmillan Press, London, 100–106.
- Hansen, L. P., and Jagannathan, R. (1991). Implications of security market data for models of dynamic economies. *Journal of Political Economy* 99: 225–262.
- Kan, R., and Zhou, G. (2006). A new variance bound on the stochastic discount factor. *Journal of Business* 79: 941–961.
- Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolio and capital budgets. *Review of Economics and Statistics* 47, 1: 13–37.
- Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica* 34, October: 768–783.
- Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13, December: 343–362.
- Sharpe, W. F. (1964). Capital asset prices. *Journal of Finance* 19, 3: 425–442.

Capital Asset Pricing Models

FRANK J. FABOZZI, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

HARRY M. MARKOWITZ, PhD
Consultant

Abstract: Risk-return analysis in finance is a “normative” theory: It does not purport to describe, rather it offers advice. Specifically, it offers advice to an investor regarding how to manage a portfolio of securities. The investor may be an institution, such as a pension fund or endowment; or it may be an asset management firm with multiple portfolios to manage (e.g., managing various mutual funds and funds for institutional clients). The focus of risk-return analysis is on advice for each individual portfolio. This contrasts with capital asset pricing models, which are hypotheses concerning capital markets as a whole. They are “positive” models, that is, they are hypotheses about that which is—as opposed to “normative” models, which advise on what should be or, more precisely, advise on what an investor should do.

INTRODUCTION

Asset pricing theory seeks to explain how the price or value of a claim from ownership of a financial asset is determined. The pricing or valuation of an asset must take into account the timing of the payments expected to be received and the risk associated with receiving the expected payments. The major challenge in asset pricing theory is often not the timing issue but the treatment of risk. The formulation of an asset pricing theory that has empirically proven to have good predictive value offers investors the opportunity to capitalize on mispriced assets. Moreover, the theory provides investors with a tool for pricing new financial instruments and nonpublicly traded assets.

Cochrane (2001) suggests two popular approaches to asset pricing: absolute pricing and

relative pricing. The absolute pricing approach seeks to price an asset by reference to its exposure to fundamental macroeconomic risk. An example of an absolute pricing approach is the consumption-based capital asset pricing model (CAPM) formulated by Breeden (1979). In contrast, the relative pricing approach seeks to value an asset based only on the prices of other assets without reliance on the exposure of the asset to the various sources of macroeconomic factors. The well-known option pricing model formulated by Black and Scholes (1973) is an example of an asset pricing model that employs the relative pricing approach.

Most asset pricing models used in practice today are the result of a blend of both approaches. *Capital asset pricing models* (CAPM), the subject of this entry, are an example. The CAPM starts

as an absolute pricing model but then, as will be explained, prices assets relative to the market. There is no attempt in the CAPM to determine how the market risk premium or the risk factor is determined in an economy.

In this entry, we focus on the basic CAPM first formulated in the 1960s by several academicians. There have been numerous extensions of the basic CAPM that have been proposed in the decades that followed but these will not be covered in this entry. However, because there is considerable confusion regarding certain aspects of the theory, in addition to describing the basic CAPM in this entry we explain the sources of the confusion and their implications.

SHARPE-LINTNER CAPM

The first CAPM was that of Sharpe (1964) and Lintner (1965). The *Sharpe-Lintner CAPM* (SL-CAPM) assumes the following:

- All investors have the same beliefs concerning security returns.
- All investors have mean-variance efficient portfolios.
- All investors can lend all they have or can borrow all they want at the same risk-free interest rate that the U.S. federal government pays to borrow money.

By the mean it is meant the expected value of the return of a security or portfolio. Thus, throughout this entry, we use the terms “mean return” and “expected return” interchangeably. By variance, we mean the variance of the returns of a security or portfolio. This is the square of the standard deviation, the most commonly used measure in statistics to quantify the dispersion of the possible outcomes of some random variable. Standard deviation is the more intuitively meaningful measure: Most of any probability distribution is between its mean mi-

nus two standard deviations and mean plus two distributions. It is not true that most of a distribution is between the mean and plus or minus two variances, or any other number of variances. While standard deviation is the more intuitive measure, formulas are more conveniently expressed in terms of variance. One can most easily compute the variance of a portfolio and then take its square root to obtain its standard deviation.

By *mean-variance efficient portfolios*, we mean that of all the possible portfolios that can be created from all of the securities in the market, the ones that have highest mean for a given variance.

The two major conclusions of the SL-CAPM are:

CAPM Conclusion 1. The market portfolio is a mean-variance efficient portfolio.

CAPM Conclusion 2. The difference between the expected return and the risk-free interest rate, referred to as the *excess return*, of each security is proportional to its *beta*.

The “market portfolio” includes all securities in the market. The composition of the portfolio is such that the sum of the weights allocated to all the securities is equal to one. That is, denoting X_i^M as the percentage of security i in the market portfolio (denoted by M), then

$$\sum_{i=1}^n X_i^M = 1 \quad (1)$$

Each holding of a security is proportional to its part of the total market capitalization. That is,

$$X_i^M = \frac{\text{Market value of } i\text{-th security}}{\text{Total market value of all securities}} \quad (2)$$

CAPM Conclusion 1 is that this “market portfolio” is on the mean-variance efficient frontier.

Let r_i stand for the return on the i -th security during some period. The return on the market

portfolio then is

$$r^M = \sum_{i=1}^n X_i^M r_i \quad (3)$$

The beta (β) referred to in CAPM Conclusion 2 can be estimated using regression analysis from historical data on observed returns for a security and observed returns for the market. In this regression analysis, security return is the “dependent variable” and market return is the “independent variable.” However, the beta produced by this analysis should be interpreted as a measure of association rather than causation. That is, it is a measure of the extent that the two quantities move up and down together, not as the so-called “independent variable” causing the level of the “dependent variable.” Below we examine why there is this association (not causation) in CAPM between security returns and market return.

The excess return, denoted by e_i , is the difference between the security’s expected return, $E(r_i)$, and the risk-free interest rate, r_f , at which all investors are assumed to lend or borrow:

$$e_i = E(r_i) - r_f \quad (4)$$

CAPM Conclusion 2 is that the excess return for security i is proportional to its β . That is, letting k be a constant then

$$e_i = k\beta_i \quad i = 1, \dots, n \quad (5)$$

It can also be shown that equation (5) applies to portfolios as well as individual securities. Thus in an SL-CAPM world, each security and portfolio has an excess return that is proportional to the regression of the security or portfolio’s return against the return of the market portfolio.

ROY CAPM

A second CAPM, which appeared shortly after that of the writings of Sharpe and Lintner, differs from the SL-CAPM only in its assumption

concerning the investment constraint imposed by investors. More specifically, it assumes that each investor (I) can choose any portfolio that satisfies

$$\sum_{i=1}^n X_i^I = 1 \quad (6)$$

without regard to the sign of the variables. Positive X_i^I is interpreted as a long position in a security while a negative X_i^I is interpreted as a short position in a security.

However, a negative X_i^I is far from a realistic model of real-world constraints on shorting. For example, equation (6) would consider feasible a portfolio with

$$\begin{aligned} X_1 &= -1,000 \\ X_2 &= 1,001 \\ X_i &= 0 \quad i = 3, \dots, n \end{aligned}$$

since the above sums to one. This would correspond to an investor depositing \$1,000 with a broker; shorting \$1,000,000 of stock 1; then using the proceeds of the sale, plus the \$1,000 deposited with the broker to buy \$1,001,000 worth of stock 2. In fact, in this example, Treasury Regulation T (Reg T) would require that the sum of long positions, plus the value of the stocks sold short, not exceed \$2,000.

Equation (6), as the only constraint on portfolio choice, was first proposed by Roy (1952), albeit not in a CAPM context. Since it is difficult to pin down who first used this constraint set in a CAPM (more than one did so almost simultaneously), we refer to this as the *Roy CAPM* as distinguished from the SL-CAPM.

CONFUSIONS REGARDING THE CAPM

Probably no other part of financial theory has been subject to more confusion, by

professionals and amateurs alike, than the CAPM. Major areas of confusion include the following:

Confusion 1. Failure to distinguish between the following two statements:

The market is efficient in that each participant has correct beliefs and uses them to their advantage.

and

The market portfolio is a mean-variance efficient portfolio.

Confusion 2. Belief that equation (5) shows that CAPM investors get paid for bearing “market risk.” That this view—held almost universally until quite recently—is in error is easily demonstrated by examples in which securities have the same covariance structure but different excess returns.

Confusion 3. Failure to distinguish between the beta in Sharpe’s one-factor model of covariance (see Sharpe, 1963) and that in Sharpe’s CAPM.

The following sections present the assumptions and conclusions of the SL-CAPM and the Roy CAPM, and discuss the nature of these three historic sources of confusion, and their practical implications.

TWO MEANINGS OF MARKET EFFICIENCY

CAPM is an elegant theory. With the aid of some simplifying assumptions, it reaches dramatic conclusions about practical matters. For example:

- How can an investor choose an efficient portfolio? The answer: Just buy the market.
- How can an investor forecast expected returns? The answer: Just forecast betas.
- How should an investor price a new security? The answer is once again: Forecast its beta.

CAPM’s simplifying assumptions make it easier to deduce properties of market equilib-

ria, which is like computing falling body trajectories while assuming there is no air. But, before betting the ranch that the feather and the brick will hit the ground at the same time, it is best to consider the implications of some of the omitted complexities. The present section mostly explores the implications of generalizing one of the CAPMs’ simplifying assumptions.

Note the difference between the statement “The market is efficient,” in the sense that market participants have accurate information and use it correctly to their benefit, and the statement “The market portfolio is a mean-variance efficient portfolio.” Under some assumptions the two statements are equivalent. Specifically, if we assume:

Assumption 1. Transaction costs and other illiquidities can be ignored.

Assumption 2. All investors hold mean-variance efficient portfolios.

Assumption 3. All investors hold the same (correct) beliefs about means, variances, and covariances of securities.

Assumption 4. Every investor can lend all she or he has or can borrow all she or he wants at the risk-free interest rate.

Then based on these four assumptions we get CAPM Conclusion 1: The market portfolio is a mean-variance efficient portfolio. This CAPM conclusion also follows if Assumption 4 is replaced by the following assumption:

Assumption 4’. Equation (6) is the only constraint on the investor’s choice of portfolio.

As noted earlier, a negative X_i is interpreted as a short position; but this is clearly a quite unrealistic model of real-world short constraints. Equation (6) would permit any investor to deposit \$1,000 with a broker, sell short \$1,000,000 worth of one security, and buy long \$1,001,000 worth of another security.

In addition to CAPM Conclusion 1, Assumptions 1 through 4 imply CAPM Conclusion 2: In equilibrium, excess returns are proportional to betas, as in equation (5). This CAPM

Table 1 Expected Returns and Standard Deviations for Three Hypothetical Securities^a

Security	Expected Return	Standard Deviation
1	0.15%	0.18%
2	0.10%	0.12%
3	0.20%	0.30%

^aSecurity returns are uncorrelated.

conclusion is the basis for the CAPM’s prescriptions for risk adjustment and asset valuation.

Since a Roy CAPM world may or may not have a risk-free asset, Assumptions 1–3 plus Assumption 4’ cannot imply CAPM Conclusion 2. These assumptions do, however, imply the following:

CAPM Conclusion 2’. Expected returns are a linear function of betas, that is, there are constants, a and b , such that

$$E(r_i) = a + b\beta_i \quad i = 1, \dots, n \quad (7)$$

Equation (5) of the SL-CAPM is the same as equation (7) of the Roy CAPM with $a = r_f$.

CAPM Conclusions 1 and 2 (or 2’) do not follow from Assumptions 1, 2, and 3 if 4 (or Assumption 4’) is replaced by a more realistic description of the investor’s investment constraints. This is illustrated by an example with the expected returns and standard deviations given in Table 1. In this example, it is assumed that the returns are uncorrelated (but similar results occur with correlated returns). The example assumes that investors cannot sell short or borrow. The same results hold if investors can borrow limited amounts or can sell short but are subject to Reg T or a similar constraint.

Assumptions 1 through 3 are assumed in this example. Rather than Assumption 4 or Assumption 4’, the example assumes that the investor can choose any portfolio that meets the following constraints:

$$X_1 + X_2 + X_3 = 1.0 \quad (8a)$$

and

$$X_1 \geq 0, X_2 \geq 0, X_3 \geq 0 \quad (8b)$$

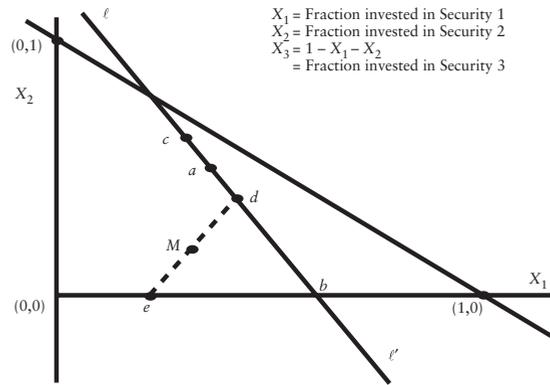


Figure 1 Example Illustrating That When Short Sales Are Not Allowed, the Market Portfolios Are Typically Not Mean-Variance Efficient

This is the “standard” portfolio selection constraint set presented in Markowitz (1952). It differs from the Roy constraint set in the inclusion of nonnegativity constraints, the inequalities given by (8b).

In Figure 1, X_1 —the fraction invested in Security 1—is plotted on the horizontal axis; X_2 —the fraction invested in Security 2—is plotted on the vertical axis; and X_3 —the fraction invested in the third security—is given implicitly by the relationship $X_3 = 1 - X_1 - X_2$. In the figure, the portfolio labeled “c” has smaller variance than any other portfolio that satisfies the equation (8a) constraint. In general, such a minimum-overall-variance portfolio may or may not satisfy the inequalities given by (8b) constraints. In other words, the minimum-overall-variance portfolio may or may not be feasible for the original Markowitz constraint set (Markowitz, 1952). In the present example it is. Results similar to those we illustrate here also typically hold when c is not feasible for the standard model.¹

The line ll' connects all points (portfolios) that minimize variance, on the portfolio-as-a-whole, for various levels of portfolio expected return, subject to equation (8a), ignoring non-negativity inequalities (8b). Using differential calculus, one can minimize a function such as

$$V = \sum_{i=1}^3 X_i^2 V_i \quad (9a)$$

subject to constraints

$$\sum_{i=1}^3 X_i = 1 \quad (9b)$$

$$E_0 = \sum_{i=1}^3 X_i E(r_i) \quad (9c)$$

One can do so with the expected returns and standard deviations from Table 1, letting E_0 vary, and thereby obtain the line in Figure 1. Moving downward and to the right on $\ell\ell'$, the portfolio expected return increases. This downward direction for increasing expected return does not always hold: It depends on the choice of security expected returns.

In the Roy model, every point in the figure is feasible since they all satisfy equation (6) or, equivalently, equation (8a). It follows that, in the Roy CAPM, *all* points on $\ell\ell'$, from “ c ” downward in the direction of increasing E , are efficient. But in the standard model, including nonnegativity inequalities (8b), all points on $\ell\ell'$ below the point “ b ” are not feasible (since they have negative X_2) and therefore cannot be efficient. In this example, when portfolio choice is subject to the standard constraint set, the set of efficient portfolios is the same as that of the Roy constraint set from portfolio c to portfolio b . After that, the set of efficient portfolios moves horizontally along the X_1 axis, ending at point $(0, 0)$. This represents the portfolio with everything invested in Security 3, which has maximum expected return in the example.

Suppose that some investors select the cautious portfolio d , while the remainder selects the more aggressive portfolio e . The market portfolio M lies on the straight line that connects d and e (e.g., halfway between if both groups have equal amounts invested).

But M is not an efficient portfolio, either for the standard constraint set or for the Roy constraint set. Thus, even though all investors hold mean-variance efficient portfolios, the market portfolio is not mean-variance efficient!

A Simple Market

Figure 1 demonstrates that if the expected returns and variances for our three hypothetical securities in Table 1 reflect equilibrium beliefs, then the market portfolio would not be a mean-variance efficient portfolio. But can these be equilibrium beliefs? Consider the following simple market: Inhabitants of an island live on coconuts and produce them from their own gardens. The island has three enterprises, namely, three coconut farms. Once a year, a stock market convenes to trade the shares of the three farms. Each year the resulting share prices turn out to be the same as those of preceding years. Thus the only source of uncertainty of return is the dividend each stock pays during the year, which is the stock’s pro rata share of the farm’s production. Markowitz (2005) shows that means, variances, and covariances of coconut production exist that imply the efficient set in Figure 1, or in any of the other three security-efficient sets presented in Markowitz (1952 and Chapter 7 in 1959) initial works.

With such probability distributions of returns, the market is rational in the sense that each participant knows the true probability distribution of returns, and each seeks and achieves mean-variance efficiency. Nevertheless, in contrast to the usual CAPM conclusion, the market portfolio is not an efficient portfolio. It also follows that there is no representative investor since no one wants to hold the market portfolio.

Arbitrage

Suppose that most investors are subject to the nonnegativity requirement of inequalities (8b), but one investor can short in the CAPM sense. (Perhaps the CAPM investor has surreptitious access to a vault containing stock certificates that he or she can “borrow” temporarily without posting collateral.) Would this CAPM investor, with unlimited power to short and use the proceeds to buy long, arbitrage away the inefficiency in the market portfolio?

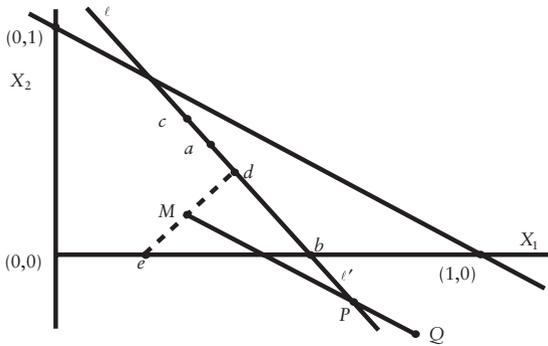


Figure 2 Illustration That an Investor Who Can Sell Short and Use the Proceeds to Buy Long Should Not Short an Inefficient Market

Figure 2 shows an investor would not do so. Suppose that portfolio P is the one most preferred by the Roy CAPM investor. If this investor shorts M and uses the proceeds to buy more P , then the resulting portfolio will be on the straight line connecting M and P —but this time on the far side of P (e.g., at Q) rather than between M and P . But Q is not efficient for the Roy CAPM investor since it does not lie on the ll' line. The Roy CAPM investor is better off just holding P rather than shorting M to buy more P .

With market participants holding portfolios d , e , and P and with the weighted average of the d and e investors being at M , the new market portfolio will be on the straight line between M and P , such as at M^a , M^b , or M^c in Figure 3.

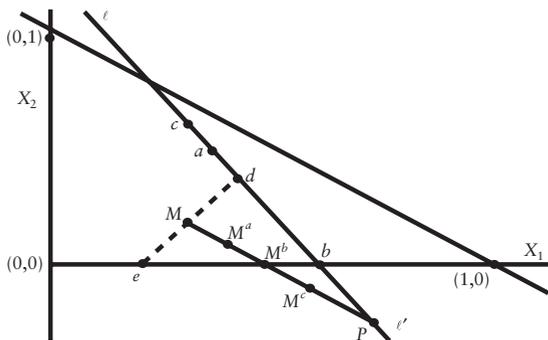


Figure 3 Illustration That the Presence of a CAPM Short Seller Does Not Make the Market Portfolio Efficient

M^c cannot be the market equilibrium since this would imply a negative market value for Security 2. Similarly, M^b implies a zero market value for Security 2, therefore a zero price.

Thus the only points (portfolios) between M and P that are consistent with positive prices for all securities lie strictly between M and M^b , such as M^a ; but M^a is not efficient for the investors with either a standard or a Roy constraint set.

Expected Returns and Betas

If Assumptions 1 through 4 (or Assumption 4') are true, then CAPM Conclusion 2' follows: Expected returns are linearly related to the betas of each security as in equation (7), that is,

$$E_1 = a + b\beta_1$$

$$E_2 = a + b\beta_2$$

$$E_3 = a + b\beta_3$$

where β_i is the coefficient of the regression of the return on the i th security against the return on the market portfolio. In other words, all (E_i, β_i) combinations lie on the straight line

$$Y = a + bX$$

But equation (7) does not typically hold if Assumptions 1 through 3 are true but neither Assumption 4 nor Assumption 4' is also true, as illustrated using the data in Tables 2 and 3, and Figure 4. Table 2 shows the β_i for portfolio P ; Table 3 shows them for portfolio M . These betas are computed using the fact that the regression coefficient $\beta_{s,r}$ of random variable s against a random variable r is

$$\beta_{s,r} = \frac{\text{Covariance}(r, s)}{\text{Variances}(s)} \tag{10}$$

Table 2 Betas versus Portfolio P

Security	Percent in P	cov _{i,P} = P _i V _i	beta _{i,P}
1	0.70%	0.0227	0.52
2	-0.25	-0.0036	-0.08
3	0.55	0.0495	1.12

Note: var(P) = 0.0440; beta_{i,P} = cov_{i,P} / var(P).

Table 3 Betas versus Portfolio M

Security	Percent in M	$cov_{i,M} = M_i V_i$	$\beta_{i,M}$
1	0.30	0.0097	0.36
2	0.19	0.0027	0.10
3	0.51	0.0459	1.71

Note: $\text{var}(M) = 0.0268$; $\beta_{i,M} = \text{cov}_{i,M} / \text{var}(M)$.

Figure 4 shows the plot of these betas against the expected returns given in Table 1. The relationship between beta and expected return is linear for regressions against P , as implied by equation (7), but not against M . In general, expected returns are a linear function of betas if and only if the regressions are against a portfolio on the $\ell\ell'$ line. (See Chapter 12 in Markowitz and Todd [2000].)

Limited Borrowing

Thus far we have seen that the market portfolio is not necessarily an efficient portfolio, and there is usually no linear relationship between expected returns and betas (regressions against the market portfolio) if the SL-CAPM or Roy CAPM is replaced by the standard, Markowitz constraint set, constraints given by (8). Figure 5 illustrates that the same conclusions hold if borrowing and lending at a risk-free interest rate are permitted, but borrowing is limited, for example, to 100% of the equity in the portfolio. In Figure 5, Security 3 is the risk-free asset.

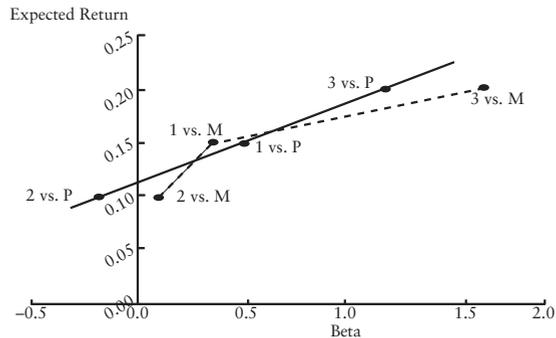


Figure 4 Linear Relationship between Expected Returns and Betas If and Only If the Regression Is Against a Portfolio on the Line $\ell\ell'$ in Figure 1.

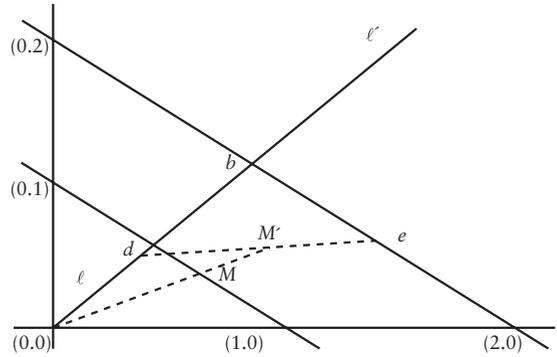


Figure 5 Illustration That If Borrowing Is Permitted but Limited, the Market Portfolio Is Still Typically Not an Efficient Portfolio

With 100% borrowing permitted, the set of feasible portfolios is no longer on and in the triangle with $(0, 0)$, $(1, 0)$, and $(0, 1)$ as its vertices. Rather, the feasible region is on and in the triangle whose vertices are $(0, 0)$, $(2, 0)$, and $(0, 2)$. For example, the $(2, 0)$ point represents the portfolio with 200% invested in Security 1.

In the SL-CAPM, the efficient set starts at the portfolio $(0, 0)$, which holds only the risk-free asset. From there, the efficient set moves along a straight line in the first quadrant of Figure 5.² In the SL-CAPM, this efficient line would continue in the same direction without limit. In the model with borrowing limited to at most 100% of equity, the ray extending from $(0, 0)$ is no longer feasible (therefore no longer efficient) when it crosses the line connecting $(0, 2)$ and $(2, 0)$ —at b in the figure. The efficient set then moves towards the leveraged portfolio with highest expected return: $(2, 0)$ in the present case. Thus in Figure 5 the set of efficient portfolios is the line segment connecting $(0, 0)$ to b , followed by the segment connecting b to $(2, 0)$. As in our analysis using the standard constraint set, if some investors hold portfolio d and the remainder hold portfolio e , then the “market portfolio” will be between them (e.g., at M') and will not be an efficient portfolio.

We put “market portfolio” in quotes above because M' is a leveraged portfolio. In order to meet the definition of market portfolio in

equation (1), so that the holdings in the market portfolio sum to one, we must rescale M' . This gives us the market portfolio (no quotation marks) M , which is also not an efficient portfolio.

Finally, as in the analysis of the standard case since M is not on the $\ell\ell'$ line, there does not exist a linear relationship between expected returns and betas. Also, there is no “representative investor,” since no investor wants to hold the market portfolio.

Further Generalizations

Suppose that there are n securities (for $n = 3$ or 30 or $3,000$). Suppose that one security has the highest expected return, and that the n securities have a “nonsingular covariance matrix.” This means that there is no riskless combination of risky securities. If the only constraint on the choice of portfolio is equation (6), then the portfolios that minimize portfolio variance V_p for various values of portfolio expected return E_p lie on a single straight line in n -dimensional portfolio space. This is not true for an investor also subject to nonnegativity constraints such as in the inequalities given by (8b).

The *critical line algorithm* (CLA) for tracing out all efficient portfolios begins with the portfolio that is 100% invested in the security with highest expected return (see Markowitz and Todd, 2000). It traces out the set of efficient portfolios in a series of iterations. Each iteration computes one piece (one linear segment) of the piecewise linear efficient set. Each successive segment has either one more or one less security than the preceding segment. If the universe consists of, say 10,000 securities, and if all securities are to be demanded by someone, then this universal efficient frontier must contain at least 10,000 segments. If investors have sufficiently diverse risk tolerances, they will choose portfolios on many different segments. The market portfolio is a weighted average of individual portfolios and typically will not be on any efficient segment.

This characterization of efficient sets remains true if limited borrowing is allowed, as we saw.

It also remains true when short selling is permitted but is subject to Reg T or a similar constraint (see Jacobs, Levy, and Markowitz, 2005).

CAPM INVESTORS DO NOT GET PAID FOR BEARING RISK

Recall that if the SL-CAPM assumptions are made, then a stock’s beta (regression against the market portfolio) is proportional to its excess return, as shown in equation (5). Markowitz shows that this does *not* imply that CAPM investors are paid to bear risk (see Markowitz, 2008).

This is most easily seen if we assume that risks are uncorrelated. (CAPM should cover this case, too.) In this case, we show that two securities can have the same variance but different expected returns, or the same expected returns and different variances. Therefore, it cannot be true that the investor is paid for bearing risk!

According to equation (10), the beta of r_i against r_M is

$$\beta_i = \frac{\text{Covariance}(r_i, r^M)}{\text{Variance}(r^M)}$$

Therefore, equation (5) holds if and only if we also have

$$e_i = \tilde{b} \text{covariance}(r_i, r^M) \quad (11)$$

where

$$\tilde{b} = b/\text{Var}(r^M)$$

In other words, excess return is proportional to β_i if and only if it is proportional to the covariance between r_i and r^M .

As a calculus exercise one can show that, in the uncorrelated case, the SL-CAPM investor minimizes portfolio variance for given portfolio mean if and only if the investor chooses a portfolio such that

$$VX_i^l = k^l e_i \quad (12a)$$

where V_i is the variance of r_i and k^l depends on the investor's risk aversion.

Equation (12a) implies a similar relationship for the market portfolio:³

$$V_i X_i^M = k^M e_i \quad (12b)$$

Therefore,

$$X_i^M = k^M \left(\frac{e_i}{V_i} \right) \quad \text{for } i = 1, \dots, n \quad (12c)$$

Thus if two securities have the same positive excess return but different variances, the market portfolio will contain a larger dollar value of the one with the lower variance. Conversely, if two securities have the same variance but different positive excess returns, the market portfolio will contain a larger dollar value of the one with the higher excess return.

Now let us consider where the linear relationship in equation (5), or (11), comes from in this case of uncorrelated returns. It can be shown that in equation (12b), $V_i X_i^M$ is the covariance of the r_i with the market. Therefore, covariance with the market is proportional to excess return (and vice versa) because the security with the higher ratio of excess return to variance is a larger part of the market portfolio.

Thus, in the uncorrelated case, the relationship between beta and excess return in equation (5) results from the security with higher excess return (per unit variance) being a larger part of the market portfolio. The beta in equation (5) is the regression of r_i against the market portfolio and, in the uncorrelated case, the only security in the market portfolio with which it is correlated is itself.

When returns are correlated, the formula for the covariance between security return and market portfolio return is more complicated, but the basic principle is the same. For example, if two securities have the same covariance structure, the one with the higher expected return will constitute a larger share of the market portfolio—despite the presence in the market portfolio of securities with which it is correlated—and hence have its own re-

turns more correlated with returns on the market portfolio.

THE “TWO BETA” TRAP

Two distinct meanings of the word “beta” are used in modern financial theory. These meanings are sufficiently alike for people to converse—some with one meaning in mind, some with the other—without realizing they are talking about two different things. The meanings are sufficiently different, however, that one can validly derive diametrically opposite conclusions depending on which one is used. The net result of all this can be like an Abbott and Costello vaudeville comedy routine with portfolio theory rather than baseball as its setting. This is what Markowitz (1984) calls the *two beta trap*. Below we first review the background of the two betas and then tabulate propositions that are true for one concept and false for the other.

Beta₁₉₆₃

Sharpe's *single-index* (or one-factor) model of covariance introduced in 1963 assumes that the returns of different securities are correlated with each other because each is dependent on some underlying systematic factor (see Sharpe, 1963). This can be written as

$$r_i = \alpha_i + \beta_i F + u_i \quad (13)$$

where the expected value of u_i is zero, and u_i is uncorrelated with F and every other u_j .

Originally F was denoted by I and described as an “underlying factor, the general prosperity of the market as expressed by some index.” We have changed the notation from I to F to emphasize that r_i depends on the underlying factor rather than the index used to estimate the factor. The index never measures the factor exactly, no matter how many securities are used in the index, provided that each security has positive

variance of u_i , since the index I equals:

$$\begin{aligned} I &= \sum w_i r_i \\ &= \sum \alpha_i w_i + F(w_i \beta_i) + \sum u_i w_i \\ &= A + BF + U \end{aligned} \quad (14)$$

where w_i is the weight of return r_i in the index, and

$$\begin{aligned} A &= \sum \alpha_i w_i \\ B &= \sum w_i \beta_i \\ U &= \sum u_i w_i \end{aligned}$$

U is the error in the observation of F . Under the conditions stated, the variance of U is

$$V_U = \sum_{i=1}^N w_i^2 V_{u_i} > 0 \quad (15)$$

Sharpe (1963) tested equation (13) as an explanation of how security returns tend to go up and down together. He concluded that equation (13) was as complex a model of covariance as seemed to be needed. This conclusion was supported by research of Cohen and Pogue (1967). King (1966) found strong evidence for industry factors in addition to the market-wide factor. Rosenberg (1974) found other sources of systematic risk beyond the market-wide factor and industry factors.

We refer to the beta coefficient in equation (13) as “beta₁₉₆₃” since it is the subject of Sharpe’s 1963 article. We contrast the properties of this beta with that of the beta that arises from the Sharpe-Lintner CAPM. The latter we will refer to as “beta₁₉₆₄” since it is the subject of Sharpe (1964).

Beta₁₉₆₄

We noted that the SL-CAPM makes various assumptions about the world, including that all investors are mean-variance efficient, have the same beliefs, and can lend or borrow all they want at the same “risk-free” interest rate. Note, however, one assumption that the SL-CAPM

does *not* make is that the covariances among securities satisfy equation (13). On the contrary, the assumptions it makes concerning covariances are quite general.⁴ They are consistent with equation (13) but do not require it. They are also consistent with the existence of industry factors as noted by King, or other sources of systematic risk such as those identified by Rosenberg.

As previously noted, the beta that appears in the CAPM relationship of equation (5) (which we now refer to as beta₁₉₆₄) is the regression of the i th security’s return against the return on the market portfolio. This is defined whether or not the covariance structure is generated by the single-factor model of equation (13). Equation (5) is an assertion about the expected return of a security and how it relates to the regression of the security’s return against the market-portfolio return. Unlike equation (13), it is not an assertion about how security returns covary.

One source of confusion between beta₁₉₆₃ and beta₁₉₆₄ is that William Sharpe presented each of them. Sharpe, however, has never been confused on this point. In particular, when explaining beta₁₉₆₄ he emphasizes that he derived it without assuming equation (13).

Propositions about Betas

Table 4 lists various propositions about betas and indicates whether they are true or false for beta₁₉₆₃ or beta₁₉₆₄. The first column presents each proposition, the second indicates whether the proposition is true or false for beta₁₉₆₃, and the third column indicates the same for beta₁₉₆₄. Most of the propositions in Table 4 are true for one of the betas and false for the other.

Proposition 1

Because of the definition of a regression beta in general, both beta₁₉₆₃ and beta₁₉₆₄ equal

$$\beta_i = \text{cov}(r_i, R) / V(R)$$

for some random variable R . In the case of beta₁₉₆₃, R is F for equation (13); in the case of beta₁₉₆₄, R is the M in equations (1) and (2).

Table 4 Propostions about Beta

	β_{1963}	β_{1964}
1. The β_i of the i th security equals $\text{cov}(r_i, R)/V(R)$ for some random variable R .	T	T
2. R is “observable”; specifically, it may be computed exactly from security returns (r_i) and market values (X_i).	F	T
3. R is a <i>value</i> -weighted average of the (r_i).	F	T
4. An index I that estimates R should ideally be weighted by a combination of ($1/V_{u_i}$) and (β_i/V_i). Unfortunately, the β_i and V_{u_i} needed to determine these weights are unobservable.	T	F
5. If ideal weights are not used, then equal weights are “not bad” in computing I ; specifically, nonoptimum weights can be compensated for by increased sample size.	T	F
6. Essentially, all that is important in computing I is to have a large number of securities; it is not necessary to have a large fraction of all securities.	T	F
7. The ideally weighted index is an efficient portfolio.	F	T

Proposition 2

Equation (15) implies that F cannot be observed exactly no matter how many securities are used to estimate it, provided that no security has a zero variance of u_i . In contrast, portfolio M in equation (2) is observable, at least in principle, if only we are diligent enough to measure each X_i^M in the market. Thus, the assertion that R is observable is true in principle for β_{1964} and false for β_{1963} .

Propositions 3 and 4

One source of confusion about the two betas concerns whether an index estimating R should be “value weighted”; that is, should the w_i used in computing an estimate of R from the r_i equal the X_i^M ? We have seen that in the case of β_{1964} :

$$R = \sum X_i^M r_i$$

In this case $W_i = X_i^M$ = market-value weights.

The answer is different in the case of β_{1963} . Ideally, we would like to eliminate the error term U from equation (14). Our index would be perfect if $V_U = 0$, provided of course $B \neq 0$. Nevertheless, as long as no security has $V_{u_i} = 0$, the perfect index cannot be achieved with a finite number of securities. Short of this, it might seem that the best to be wished is that V_U be a minimum. In this case, w_i would equal $1/V_{u_i}$. The optimum choice of weights for estimating

the underlying factor F is more complicated, depending also on β_i/V_i (see Markowitz, 1983) and more complicated still, since V_{u_i} and β_i are not known.

Proposition 5

The fifth proposition in Table 4 asserts that if ideal weights cannot be obtained, equal weights are good enough. In particular, an increase in the number of securities can compensate for nonoptimum weights. We have already seen that this proposition is false for β_{1964} . It is easily seen to be true for β_{1963} under mild restrictions on how fast the V_{u_i} increases as i increases.

Proposition 6

The next proposition asserts that all that is important in designing a good index is to have many securities, as opposed to having a large percentage of the population represented in the index. This proposition is true for I_{1963} and false for I_{1964} , as may be illustrated by two extreme examples.

First, suppose that there are only a few securities in the entire population, and all of them are used in computing a value-weighted index. Then I_{1964} would, in fact, be M and would be precisely correct. In the case of I_{1963} , on the other hand, equation (15) implies that if $n = 6$, for example, the error term V_U is the same

regardless of whether the six securities are 100% or 1% of the universe.

At the other extreme, imagine that the sample is large but is a small percentage of the total population. For example, suppose $N = 1,000$ out of 100,000 securities. Then I_{1963} will give a good reading for F , and therefore β_{1963} , but I_{1964} may lead to serious misestimates of β_{1964} . First, the covariance with I_{1964} of an asset not in this index will tend to be too low. Second, if the index contains more of certain kinds of assets than is characteristic of the entire population, then assets of this sort will tend to have a higher correlation with the index than with the true M , and assets of other sorts will tend to have lower correlations. More precisely, the covariance between return r_i and the market is a weighted average of the covariances σ_{ij} (including $V_i = \sigma_{ii}$) weighted by market values. If the index chosen does not have approximately the same average σ_{ij} for a given i , the estimates of $\beta_{i,1964}$ will be in error.

Proposition 7

This proposition asserts that the ideal index is an efficient portfolio. This is true for I_{1964} and false for I_{1963} since one of the conclusions of the SL-CAPM assumptions is that the market portfolio is efficient. In fact, the market portfolio is the only combination of risky assets that is efficient in this CAPM. All other efficient portfolios consist of either investment in the market portfolio plus lending at the risk-free rate, or of investment in the market portfolio financed in part by borrowing at the risk-free rate. On the other hand, β_{1963} has nothing to do with expected returns or market efficiency.

KEY POINTS

- The two major conclusions of the Sharpe-Lintner CAPM are that (1) the market portfolio is a mean-variance efficient portfolio; and (2) the excess return of each security is proportional to its beta.
- The “market portfolio” includes all securities in the market.
- The beta (β) in the CAPM is estimated using regression analysis using historical data on observed returns for a security (response variable) and observed returns for the market (explanatory variable).
- The Roy CAPM differs from the Sharpe-Lintner CAPM only in its assumption concerning the investment constraint imposed by investors. More specifically, it assumes that each investor can short securities.
- Confusion regarding the CAPM involves (1) the failure to distinguish between the following two statements: The market is efficient in that each participant has correct beliefs and uses them to their advantage on the one hand, and the market portfolio is a mean-variance efficient portfolio on the other hand; (2) belief that CAPM investors get paid for bearing nondiversifiable risk; and (3) failure to distinguish between the beta in Sharpe’s one-factor model of covariance (1963 beta) and that in Sharpe’s CAPM (1964 beta).

NOTES

1. Markowitz presents examples of three-security standard analyses in which “ c ” is feasible in some cases and not feasible in others. It is possible in the latter case for the set of mean-variance efficient portfolios to be a single line segment or even a single point. But typically, when “ c ” is outside of the feasible triangle, as well as when it is within it, the set of efficient portfolios consists of two or more line segments (the “efficient segments”), which meet at “corner portfolios.” Thus the construction in Figure 1 can typically be carried out in cases in which “ c ” is not feasible. (See Markowitz, 1952.)
2. The SL-CAPM requires nonnegative investments. Thus if the parameters of an example were such that the straight line would move into, say, the fourth quadrant, X_2 would equal zero on the line and would, in

effect, drop out of the market, and out of the analysis.

3. If we multiply both sides of equation (12a) by w^l , the l -th investor's equity as a fraction of total market equity, and sum we get

$$V_i \left(\sum_l w^l X_i^l \right) = \left(\sum_l w^l k^l \right) e_i$$

If we sum the above over all securities, the second factor on the left, namely

$$S = \sum_i \left(\sum_l w^l X_i^l \right)$$

will not necessarily sum to one since nothing in the SL-CAPM assumptions prevents market participants from being either net borrowers or net lenders. However, if we divide both sides of equation (12c) by S , we get equation (12b) for the market portfolio as defined in equations (1) and (2).

4. Mossin (1966) provides a precise statement of the assumptions behind the S-L CAPM. Specifically all that Mossin assumes about covariances is that the covariance matrix is nonsingular (i.e., that no portfolio of risky securities is riskless).

REFERENCES

- Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 3: 637–654.
- Breeden, D. T. (1979). An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics* 7: 265–296.
- Cochrane, J. H. (2001). *Asset Pricing*. Princeton, NJ: Princeton University Press.
- Cohen, K. J., and Pogue, J. A. (1967). An empirical evaluation of alternative portfolio-selection models. *Journal of Business* 40, 2: 166–193.
- Jacobs, B. I., Levy, K. N., and Markowitz, H. M. (2005). Portfolio optimization with factors, scenarios, and realistic short positions. *Operations Research* 53, 4: 586–599.
- King, B. F. (1966). Market and industry factors in stock price behavior. *Journal of Business* 39, 1, Part II: 139–190.
- Lintner, J. (1995). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47, 1: 13–37.
- Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance* 7, 1: 77–91.
- Markowitz, H. M. (1959). *Portfolio Selection: Efficient Diversification of Investments*, 2nd Edition. New York: John Wiley & Sons.
- Markowitz, H. M. (1984). The two beta trap. *Journal of Portfolio Management* 11, 1: 12–20.
- Markowitz, H. M. (2005). Market efficiency: A theoretical distinction and so what? *Financial Analysts Journal* 61, 5: 17–30.
- Markowitz, H. M. (2008). CAPM investors do not get paid for bearing risk. *Journal of Portfolio Management* 34, 2: 91–94.
- Markowitz, H. M., and Todd, P. (2000). *Mean-Variance Analysis in Portfolio Choice and Capital Markets*. Hoboken, NJ: John Wiley.
- Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica* 34, 4: 768–783.
- Rosenberg, B. (1974). Extra-market components of covariance in security returns. *Journal of Financial and Quantitative Analysis* 9, 2: 263–273.
- Roy, A. D. (1952). Safety first and the holding of assets. *Econometrica* 20: 431–449.
- Sharpe, W. F. (1963). A simplified model for portfolio analysis. *Management Science* 9, 2: 277–293.
- Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19, 3: 425–442.

Modeling Asset Price Dynamics

DESSISLAVA A. PACHAMANOVA, PhD

Associate Professor of Operations Research, Babson College

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: The dynamics of asset price processes in discrete time increments are typically described by two kinds of models: trees (lattices) and random walks. Arithmetic, geometric, and mean reverting random walks are examples of the latter type of models. When the time increment used to model the asset price dynamics becomes infinitely small, we talk about stochastic processes in continuous time. Models for asset price dynamics can incorporate different observed characteristics of an asset price process, such as a drift or a reversion to a mean, and are important building blocks for risk management and financial derivative pricing models.

Many classical asset pricing models, such as the capital asset pricing theory and the arbitrage pricing theory, take a myopic view of investing: They consider events that happen one time period ahead, where the length of the time period is determined by the investor. This entry presents apparatus that can handle asset dynamics and volatility over time. The dynamics of price processes in discrete time increments are typically described by two kinds of models: *trees* (such as *binomial trees*) and *random walks*. When the time increment used to model the asset price dynamics becomes infinitely small, we talk about *stochastic processes* in continuous time.

In this entry, we introduce the fundamentals of binomial tree and random walk models, providing examples for how they can be used in practice. We briefly discuss the special nota-

tion and terminology associated with stochastic processes at the end of the entry; however, our focus is on interpretation and simulation of processes in discrete time. The roots for the techniques we describe are in physics and the other natural sciences. They were first applied in finance at the beginning of the 20th century and have represented the foundations of asset pricing ever since.

FINANCIAL TIME SERIES

Let us first introduce some definitions and notation. A financial time series is a sequence of observations of the values of a financial variable, such as an asset price (index level) or asset (index) returns, over time. Figure 1 shows an example of a time series, consisting of weekly observations of the S&P 500 price level over a

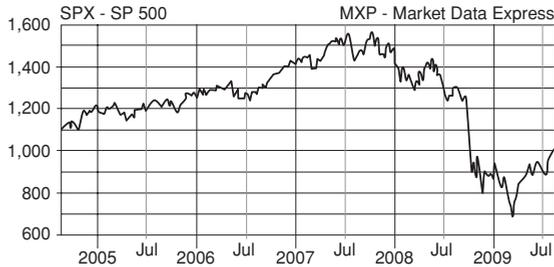


Figure 1 S&P 500 Index Level between August 19, 2005 and August 19, 2009

period of five years (August 19, 2005 to August 19, 2009).

When we describe a time series, we talk about its drift and volatility. The term “drift” is used to indicate the direction of any observable trend in the time series. In the example shown in Figure 1, it appears that the S&P 500 time series has a positive drift up from August 2005 until about the middle of 2007, as the level of prices appears to have been generally increasing over that time period. From the middle of 2007 until the beginning of 2009, there is a negative drift. The volatility is smaller (the time series is less “squiggly”) from August 2005 until about the middle of 2007, but increases dramatically between the middle of 2007 and the beginning of 2009.

We are usually interested also in whether the volatility increases when the price level increases, decreases when the price level decreases, or remains constant independently of the current price level. In this example, the volatility is lower when the price level is increasing, and is higher when the price level is decreasing.

Finally, we talk about the *continuity* of the time series—is the time series smooth, or are there jumps whose magnitude appears to be large relative to the price movements the rest of the time? From August 2005 until about the middle of 2007, the time series is quite smooth. However, some dramatic drops in price levels can be observed between the middle of 2007 and the beginning of 2009—notably in the fall of 2008.

For the remainder of this entry, we will use the following notation:

- S_t : value of underlying variable (price, interest rate, index level, etc.) at time t .
- S_{t+1} : value of underlying variable (price, interest rate, etc.) at time $t + 1$.
- ω_t : a random error term observed at time t . (For the applications in this entry, it will follow a normal distribution with mean equal to 0 and standard deviation equal to σ .)
- ε_t : a realization of a normal random variable with mean equal to 0 and standard deviation equal to 1 at time t .

BINOMIAL TREES

Binomial trees (also called *binomial lattices*) provide a natural way to model the dynamics of a random process over time. The initial value of the security S_0 (at time 0) is known. The length of a time period, Δt , is specified before the tree is built. (The symbol Δ is often used to denote *difference*. The notation Δt therefore means time difference, i.e., length of one time period.)

The binomial tree model assumes that at the next time period, only two values are possible for the price, that is, the price may go up with probability p or down with probability $(1 - p)$. Usually, these values are represented as multiples of the price at the beginning of the period. The factor u is used for an up movement, and d is used for a down movement. For example, the two prices at the end of the first time period are $u \cdot S_0$ and $d \cdot S_0$. If the tree is recombining, there will be three possible prices at the end of the second time period: $u^2 \cdot S_0$, $u \cdot d \cdot S_0$, and $d^2 \cdot S_0$. Proceeding in a similar manner, we can build the tree in Figure 2.

The binomial tree model may appear simple, because, given a current price, it only allows for two possibilities for the price at each time period. However, if the length of the time period is small, it is possible to represent a wide range of values for the price after only a few

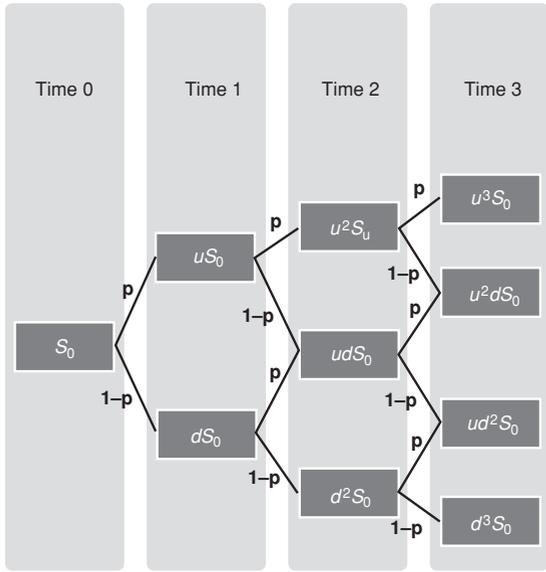


Figure 2 Example of a Binomial Tree

steps. To see this, notice that each step in the tree can be thought of as a Bernoulli trial—it is a “success” with probability p and a “failure” with probability $(1 - p)$. (One can think of the Bernoulli random variable as the numerical coding of the outcome of a coin toss, where one outcome is considered a “success” and one outcome is considered a “failure.” The Bernoulli random variable takes the value 1 (“success”) with probability p and the value of 0 (“failure”) with probability $1 - p$. Note that the definition of success and failure here is arbitrary, because

an increase in price is not always desirable, but we define them in this way for the example’s sake.)

After n steps, each particular value for the price will be reached by realizing k successes and $(n - k)$ failures, where k is a number between 0 and n . The probability of reaching each value for the price after n steps will be

$$P(k \text{ successes}) = \frac{n!}{k!(n - k)!} p^k (1 - p)^{n-k}$$

For large values of n , the shape of the binomial distribution becomes more and more symmetric and looks like a continuum. (See Figure 3(A)–(C).) In fact, the binomial distribution approximates a normal distribution with specific mean and standard deviation related to the probability of success and the number of trials. (The normal distribution is a continuous probability distribution. It is represented by a bell-shaped curve, and the shape of the curve is entirely described by the distribution mean and variance. Figure 4 shows a graph of the standard normal distribution, which has a mean of zero and a standard deviation of 1.) One can therefore represent a large range of values for the price as long as the number of time periods used in the binomial tree is large. Practitioners often use also trinomial trees, that is, trees with three branches emanating from each node, in order to obtain a better representation of the range of possible prices in the future.

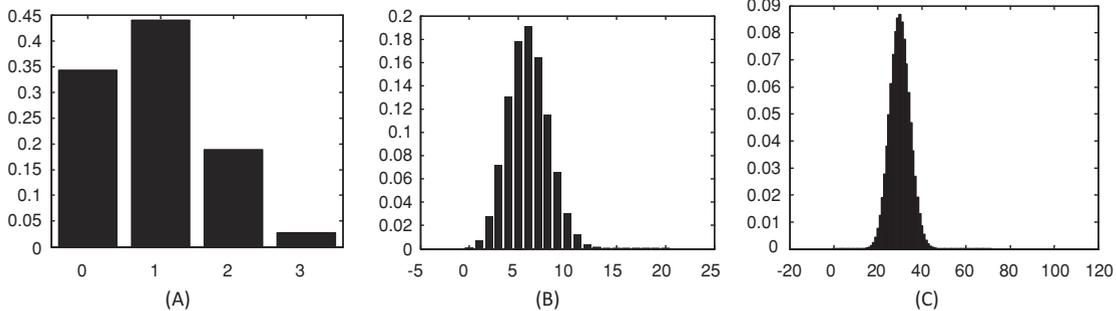


Figure 3 Binomial Distribution

Note: Probability of success (p) assumed to be 0.3. Number of trials (A) $n = 3$; (B) $n = 20$; (C) $n = 100$.

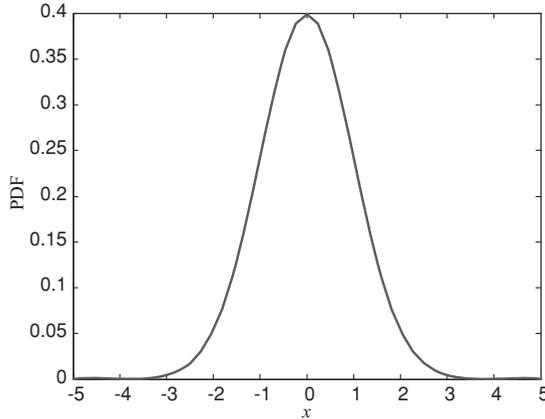


Figure 4 Standard Normal Distribution

ARITHMETIC RANDOM WALKS

Instead of assuming that at each step the asset price can only move up or down by a certain multiple with a given probability, we could assume that the price moves by an amount that follows a normal distribution with mean μ and standard deviation σ . In other words, the price for each period is determined from the price of the previous period by the equation

$$S_{t+1} = S_t + \mu + \tilde{\omega}_t$$

where $\tilde{\omega}_t$ is a normal random variable with mean 0 and standard deviation σ . We will also assume that the random variable $\tilde{\omega}_t$ describing the change in the price in one time period is independent of the random variables describing the change in the price in any other time period. (This is known as the Markov property. It implies that past prices are irrelevant for forecasting the future, and only the current value of the price is relevant for predicting the price in the next time period.) A sequence of independent and identically distributed (IID) random variables $\tilde{\omega}_0, \dots, \tilde{\omega}_t, \dots$ with zero mean and finite variance σ^2 is sometimes referred to as *white noise*.

The movement of the price expressed through the equation above is called an *arithmetic random walk with drift*. The drift term, μ , represents the

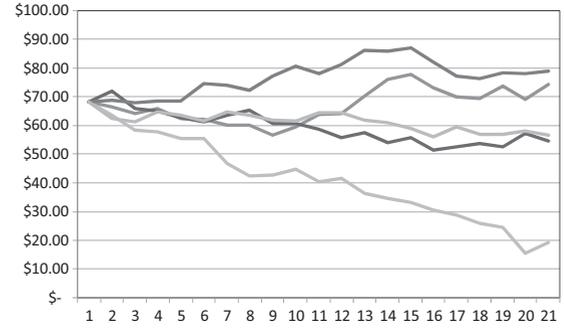


Figure 5 Five Paths of an Arithmetic Random Walk Assuming $\mu = -0.1697$ and $\sigma = 3.1166$

average change in price over a single time period. Note that for every time period t , we can write the equation for the arithmetic random walk as

$$\begin{aligned} S_t &= S_{t-1} + \mu + \tilde{\omega}_{t-1} \\ &= (S_{t-2} + \mu + \tilde{\omega}_{t-2}) + \mu + \tilde{\omega}_{t-1} \\ &= (S_{t-3} + \mu + \tilde{\omega}_{t-3}) + 2 \cdot \mu + \tilde{\omega}_{t-1} + \tilde{\omega}_{t-2} \\ &= S_0 + \mu \cdot t + \sum_{i=0}^{t-1} \tilde{\omega}_i \end{aligned}$$

Therefore, an arithmetic random walk can be thought of as a sum of two terms: a deterministic straight line $S_t = S_0 + \mu \cdot t$ and a sum of all past noise terms. (See Figure 5.)

Simulation

The equation for the arithmetic random walk can be expressed also as

$$S_{t+1} = S_t + \mu + \sigma \cdot \tilde{\varepsilon}_t$$

where $\tilde{\varepsilon}_t$ is a standard normal random variable. To show this, we need to mention that every normal distribution can be expressed in terms of the standard normal distribution, and the latter has mean of 0 and standard deviation of 1. Namely, if $\tilde{\varepsilon}$ is a standard normal variable with mean 0 and standard deviation 1, and \tilde{x} is a normal random variable with mean μ and standard deviation σ , we have

$$\tilde{\varepsilon} = \frac{\tilde{x} - \mu}{\sigma} \text{ (equivalently, } \tilde{x} = \sigma \cdot \tilde{\varepsilon} + \mu \text{)}$$

This is a property unique to the normal distribution—no other family of probability distributions can be transformed in the same way. In the context of the equation for the arithmetic random walk, we have a normal random variable $\tilde{\omega}_t$ with mean 0 and standard deviation σ . It can be expressed through a standard normal variable $\tilde{\varepsilon}_t$ as $\sigma \cdot \tilde{\varepsilon}_t + 0$.

The equation for S_{t+1} above makes it easy to generate paths for the arithmetic random walk by simulation. All we need is a way of generating the standard normal random variables $\tilde{\varepsilon}_t$. We start with an initial price S_0 , which is known. We also know the values of the drift μ and the volatility σ over one period. To generate the price at the next time period, S_1 , we add μ to S_0 , simulate a normal random variable from a standard normal distribution, multiply it by σ , and add it to $S_0 + \mu$. At the next step (time period 2), we use the price at time period 1 we already generated, S_1 , add to it μ , simulate a new random variable from a standard normal distribution, multiply it by σ , and add it to $S_1 + \mu$. We proceed in the same way until we generate the desired number of steps of the random walk. For example, given a current price S , in Excel the price for the next time period can be generated with the formula

$$S + \mu + \sigma * \text{NORMINV}(\text{RAND}(), 0, 1)$$

Parameter Estimation

In order to simulate paths of the arithmetic random walk, we need estimates of the parameters (μ and σ). We need to assume that these parameters remain constant over the time period of estimation. Note that the equation for the arithmetic random walk can be written as

$$S_{t+1} - S_t = \mu + \sigma \cdot \tilde{\varepsilon}_t$$

Given a historical series of T prices for an asset, we can therefore do the following to estimate μ and σ :

1. Compute the price changes $S_{t+1} - S_t$ for each time period t , $t = 0, \dots, T-1$.

2. Estimate the drift of the arithmetic random walk, μ , as the average of all the price changes.
3. Estimate the volatility of the arithmetic random walk, σ , as the standard deviation of all the price changes.

An important point to keep in mind is the units in which the parameters are estimated. If we are given time series in monthly increments, then the estimates of μ and σ we will obtain through steps 1–3 will be for monthly drift and monthly volatility. If we then need to simulate future paths for monthly observations, we can use the same μ and σ . However, if, for example, we need to simulate weekly observations, we will need to adjust μ and σ to account for the difference in the length of the time period. In general, the parameters should be stated as annual estimates. The annual estimates can then be adjusted for daily, weekly, monthly, and so on increments.

For example, suppose that we have estimated the weekly drift and the weekly volatility. To convert the weekly drift to an annual drift, we multiply the number we found for the weekly drift by 52, the number of weeks in a year. To convert the weekly volatility to annual volatility, we multiply the number we found for the weekly volatility by the square root of the number of weeks in a year, that is, by $\sqrt{52}$. Conversely, if we are given annualized values for the drift and the volatility, we can obtain weekly values by dividing the annual drift and the volatility by 52 and $\sqrt{52}$, respectively.

Arithmetic Random Walks: Some Additional Facts

If we use the arithmetic random walk model, any price in the future, S_t , can be expressed through the initial (known) price S_0 as

$$S_t = S_0 + \mu \cdot t + \sigma \cdot \sum_{i=0}^{t-1} \tilde{\varepsilon}_i$$

The random variable corresponding to the sum of t independent normal random variables $\tilde{\varepsilon}_0, \dots, \tilde{\varepsilon}_{t-1}$ is a normal random variable with mean equal to the sum of the means and standard deviation equal to the square root of the sum of variances. Since $\tilde{\varepsilon}_0, \dots, \tilde{\varepsilon}_{t-1}$ are independent standard normal variables, their sum is a normal variable with mean 0 and standard deviation equal to

$$\underbrace{\sqrt{1 + \dots + 1}}_{t \text{ times}} = \sqrt{t}$$

Therefore, we can have a closed-form expression for computing the asset price at time t given the asset price at time 0:

$$S_t = S_0 + \mu \cdot t + \sigma \cdot \sqrt{t} \cdot \tilde{\varepsilon}$$

where $\tilde{\varepsilon}$ is a standard normal random variable.

Based on the discussion so far in this section, we can state the following observations about the arithmetic random walk:

- The arithmetic random walk has a constant drift μ and volatility σ , that is, at every time period, the change in price is normally distributed, on average equal to μ , with a standard deviation of σ .
- The overall noise in a random walk never decays. The price change over t time periods is distributed as a normal distribution with mean equal to $\mu \cdot t$ and standard deviation equal to $\sigma \sqrt{t}$. That is why in industry one often encounters the phrase “The uncertainty grows with the square root of time.”
- Prices that follow an arithmetic random walk meander around a straight line $S_t = S_0 + \mu \cdot t$. They may depart from the line, and then cross it again.
- Because the distribution of future prices is normal, we can theoretically find the probability that the future price at any time will be within a given range.
- Because the distribution of future prices is normal, future prices can theoretically take infinitely large or infinitely small values. Thus, they can be negative, which is an undesirable consequence of using the model.

Asset prices, of course, cannot be negative. In practice, the probability of the price becoming negative can be made quite small as long as the drift and the volatility parameters are selected carefully. However, the possibility of generating negative prices with the arithmetic random walk model is real.

Another problem with the assumptions underlying the arithmetic random walk is that the change in the asset price is drawn from the same random probability distribution, independently of the current level of the prices. A more natural model is to assume that the parameters of the random probability distribution for the change in the asset price vary depending on the current price level. For example, a \$1 change in a stock price is more likely when the stock price is \$100 than when it is \$4. Empirical studies confirm that over time, asset prices tend to grow, and so do fluctuations. Only returns appear to remain stationary, that is, to follow the same probability distribution over time. A more realistic model for asset prices may therefore be that *returns* are an IID sequence. We describe such a model in the next section.

GEOMETRIC RANDOM WALKS

Consider the following model:

$$r_t = \mu + \sigma \cdot \tilde{\varepsilon}_t$$

where $\tilde{\varepsilon}_0, \dots, \tilde{\varepsilon}_t$ is a sequence of independent normal variables, and r_t , the return, is computed as

$$r_t = \frac{S_{t+1} - S_t}{S_t}$$

Returns are therefore normally distributed, and the return over each interval of length 1 has mean μ and standard deviation σ . How can we express future prices if returns are determined by the equations above?

Suppose we know the price at time t , S_t . The price at time $t+1$ can be written as

$$\begin{aligned} S_{t+1} &= S_t \cdot \frac{S_{t+1}}{S_t} \\ &= S_t \cdot \left(\frac{S_t}{S_t} + \frac{S_{t+1} - S_t}{S_t} \right) \\ &= S_t \cdot \left(1 + \frac{S_{t+1} - S_t}{S_t} \right) \\ &= S_t \cdot (1 + \tilde{r}_t) \\ &= S_t + \mu \cdot S_t + \sigma \cdot S_t \cdot \tilde{\varepsilon}_t \end{aligned}$$

This last equation is very similar to the equation for the arithmetic random walk, except that the price from the previous time period appears as a factor in all of the terms.

The equation for the *geometric random walk* makes it clear how paths for the geometric random walk can be generated. As in the case of the arithmetic random walk, all we need is a way of generating the normal random variables $\tilde{\varepsilon}_t$. We start with an initial price S_0 , which is known. We also know the values of the drift μ and the volatility σ over one period. To generate the price at the next time period, S_1 , we add $\mu \cdot S_0$ to S_0 , simulate a normal random variable from a standard normal distribution, multiply it by σ and S_0 , and add it to $S_0 + \mu \cdot S_0$. At the next step (time period 2), we use the price at time period 1 we already generated, S_1 , add to it $\mu \cdot S_1$, simulate a new random variable from a standard normal distribution, multiply it by σ and S_1 , and add it to $S_1 + \mu \cdot S_1$. We proceed in the same way until we generate the desired number of steps of the geometric random walk. For example, given a current price S , in Excel the price for the next time period can be generated with the formula

$$S + \mu * S + \sigma * S * \text{NORMINV}(\text{RAND}(), 0, 1)$$

Using similar logic to the derivation of the price equation earlier, we can express the price at any time t in terms of the known initial price S_0 . Note that we can write the price at time t as

$$S_t = S_0 \cdot \frac{S_1}{S_0} \cdot \dots \cdot \frac{S_{t-1}}{S_{t-2}} \cdot \frac{S_t}{S_{t-1}}$$

Therefore,

$$S_t = S_0 \cdot (1 + \tilde{r}_0) \cdot \dots \cdot (1 + \tilde{r}_{t-1})$$

In the case of the arithmetic random walk, we determined that the price at any time period follows a normal distribution. This was because if we know the starting price S_0 , the price at any time period could be obtained by adding a sum of independent normal random variables to a constant term and S_0 . The sum of independent normal random variables is a normal random variable itself. In the equation for the geometric random walk, each of the terms $(1 + \tilde{r}_0), \dots, (1 + \tilde{r}_{t-1})$ is a normal random variable as well. (It is the sum of a normal random variable and a constant.) However, they are multiplied together. The product of normal random variables is not a normal random variable, which means that we cannot have a nice closed-form expression for computing the price S_t based on S_0 .

To avoid this problem, let us consider the natural logarithm of prices. (The natural logarithm is the function \ln so that $e^{\ln(x)} = x$, where e is the number 2.7182...) Unless otherwise specified, we will use "logarithm" to refer to the natural logarithm, that is, the logarithm of base e .

If we take logarithms of both sides of the equation for S_t , we get

$$\begin{aligned} \ln(S_t) &= \ln(S_0 \cdot (1 + \tilde{r}_0) \cdot \dots \cdot (1 + \tilde{r}_{t-1})) \\ &= \ln(S_0) + \ln(1 + \tilde{r}_0) + \dots + \ln(1 + \tilde{r}_{t-1}) \end{aligned}$$

Log returns are in fact differences of log prices. To see this, note that

$$\begin{aligned} \ln(1 + r_t) &= \ln\left(1 + \frac{S_{t+1} - S_t}{S_t}\right) \\ &= \ln\left(\frac{S_{t+1}}{S_t}\right) \\ &= \ln(S_{t+1}) - \ln(S_t) \end{aligned}$$

Now assume that log returns (not returns) are independent and follow a normal distribution with mean μ and standard deviation σ :

$$\ln(1 + \tilde{r}_t) = \ln(S_{t+1}) - \ln(S_t) = \mu + \sigma \cdot \tilde{\varepsilon}_t$$

As a sum of independent normal variables, the expression

$$\ln(S_0) + \ln(1 + \tilde{r}_0) + \dots + \ln(1 + \tilde{r}_{t-1})$$

is also normally distributed. This means that $\ln(S_t)$ (rather than S_t) is normally distributed, that is, S_t is a lognormal random variable. Similarly to the case of an arithmetic random walk, we can compute a closed-form expression for the price S_t given S_0 :

$$\ln(S_t) = \ln(S_0) + \left(\mu - \frac{1}{2} \cdot \sigma^2 \right) \cdot t + \sigma \cdot \sqrt{t} \cdot \tilde{\varepsilon}$$

or, equivalently,

$$S_t = S_0 \cdot e^{(\mu - \frac{1}{2} \cdot \sigma^2) \cdot t + \sigma \cdot \sqrt{t} \cdot \tilde{\varepsilon}}$$

where $\tilde{\varepsilon}$ is a standard normal variable.

Notice that the only inconsistency with the formula for the arithmetic random walk is the presence of the extra term

$$\left(-\frac{1}{2} \cdot \sigma^2 \right) \cdot t$$

in the drift term

$$\left(\mu - \frac{1}{2} \cdot \sigma^2 \right) \cdot t$$

Why is there an adjustment of one half of the variance in the expected drift? In general, if \tilde{Y} is a normal random variable with mean μ and variance σ^2 , then the random variable, which is an exponential of the normal random variable \tilde{Y} , $\tilde{X} = e^{\tilde{Y}}$, has mean

$$E[\tilde{X}] = e^{\mu + \frac{1}{2} \cdot \sigma^2}$$

At first, this seems unintuitive—why is the expected value of \tilde{X} not

$$E[\tilde{X}] = e^{\mu}?$$

The expected value of a linear function of a random variable is a linear function of the expected value of the random variable. For example, if a is a constant, then

$$E[a \cdot \tilde{Y}] = a \cdot E[\tilde{Y}]$$

However, determining the expected value of a nonlinear function of a random variable (in par-

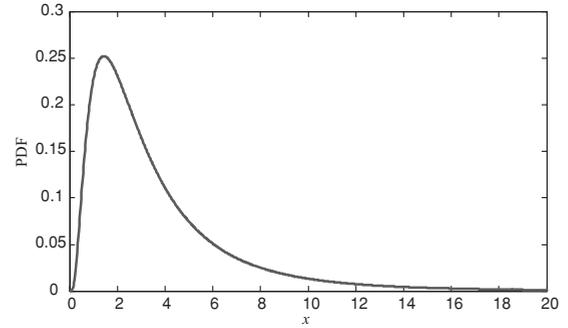


Figure 6 Example of a Lognormal Distribution with Mean of 1 and Standard Deviation of 0.8

ticular, the exponential function, which is the function we are using here) is not as trivial. For example, there is a well-known relationship, the Jensen inequality, which states that the expected value of a convex function of a random variable is less than the value of the function at the expected value of the random variable.

In our example, \tilde{X} is a lognormal random variable, so its probability distribution has the shape shown in Figure 6. The random variable \tilde{X} cannot take values less than 0. Since its variance is related to the variance of the normal random variable \tilde{Y} , as the variance σ^2 of \tilde{Y} increases, the distribution of \tilde{X} will spread out in the upward direction. This means that the mean of the lognormal variable \tilde{X} will increase not only as the mean of the normal variable \tilde{Y} , μ , increases, but also as \tilde{Y} 's variance, σ^2 , increases. In the context of the geometric random walk, \tilde{Y} represents the normally distributed log returns, and \tilde{X} is in fact the factor by which the asset price from the previous period is multiplied in order to generate the asset price in the next time period. In order to make sure that the geometric random process grows exponentially at average rate μ , we need to subtract a term (that term turns out to be $\sigma^2/2$), which will correct the bias.

Specifically, suppose that we know the price at time t , S_t . We have

$$\ln(S_{t+1}) = \ln(S_t) + \ln(1 + \tilde{r}_t)$$

that is,

$$S_{t+1} = S_t \cdot e^{\ln(1+\tilde{r}_t)}$$

Note that we are explicitly assuming a *multiplicative model* for asset prices here—the price in the next time period is obtained by multiplying the price from the previous time period by a random factor. In the case of an arithmetic random walk, we had an additive model—a random shock was added to the asset price from the previous time period.

If the log return $\ln(1 + \tilde{r}_t)$ is normally distributed with mean μ and standard deviation σ , then the expected value of

$$e^{\ln(1+\tilde{r}_t)}$$

is

$$e^{\mu + \frac{1}{2} \cdot \sigma^2}$$

and hence

$$E[S_{t+1}] = S_t \cdot e^{\mu + \frac{1}{2} \cdot \sigma^2}$$

In order to make sure that the geometric random walk process grows exponentially at an average rate μ (rather than $(\mu + 0.5 \cdot \sigma^2)$), we need to subtract the term $0.5 \cdot \sigma^2$ when we generate the future price from this process. This argument can be extended to determining prices for more than one time period ahead.

We will understand better why this formula holds when we review stochastic processes at the end of this entry.

Simulation

It is easy to see how future prices can be generated based on the initial price S_0 . First, we compute the term in the power of e : We simulate a value for a standard normal random variable, multiply it by the standard deviation and the square root of the number of time periods between the initial point and the point we are trying to compute, and subtract the product from the drift term adjusted for the volatility and the number of time periods. We then raise e to the exponent we just computed and multiply

the resulting value by the value of the initial price. For example, given a current price S , in Excel we use the formula

$$S * \exp((\mu - 0.5 * \sigma^2) * t - \sigma * \sqrt{t}) * \text{NORMINV}(\text{RAND}(), 0, 1)$$

One might wonder whether this approach for simulating realizations of an asset price following a geometric random walk is equivalent to the simulation approach mentioned earlier when we introduced geometric random walks, which is based on the discrete version of the equation for a random walk. The two approaches are different (for example, the approach based on the discrete version of the equation for the geometric random walk does not produce the expected lognormal price distribution), but it can be shown that the differences in the two simulation approaches tend to cancel over many steps.

Parameter Estimation

In order to simulate paths of the geometric random walk, we need to have estimates of the parameters (μ and σ). The implicit assumption here, of course, is that these parameters remain constant over the time period of estimation. (We will discuss how to incorporate considerations for changes in volatility later in this entry.) Note that the equation for the geometric random walk can be written as

$$\ln(S_{t+1}) - \ln(S_t) = \ln(1 + \tilde{r}_t)$$

Equivalently,

$$\ln\left(\frac{S_{t+1}}{S_t}\right) = \mu + \sigma \cdot \tilde{\epsilon}_t$$

Given a historical series of T prices of an asset, we can therefore do the following to estimate μ and σ :

1. Compute $\ln(S_{t+1}/S_t)$ for each time period t , $t = 0, \dots, T-1$.
2. Estimate the volatility of the geometric random walk, σ , as the standard deviation of all $\ln(S_{t+1}/S_t)$.

- Estimate for the drift of the arithmetic random walk, μ , as the average of all $\ln(S_{t+1}/S_t)$, plus one half of the standard deviation squared.

If we are given data on the returns r_t of an asset rather than the prices of the asset, we can compute $\ln(1 + r_t)$, and use it to replace $\ln(S_{t+1}/S_t)$ in steps 1–3 above. This is because

$$\log\left(\frac{S_{t+1}}{S_t}\right) = \log\left(1 + \frac{S_{t+1} - S_t}{S_t}\right) = \log(1 + \tilde{r}_t)$$

Geometric Random Walk: Some Additional Facts

To summarize, the geometric random walk has several important characteristics:

- It is a multiplicative model, that is, the price at the next time period is a multiple of a random term and the price from the previous time period.
- It has a constant drift μ and volatility σ . At every time period, the percentage change in price is normally distributed, on average equal to μ , with a standard deviation of σ .
- The overall noise in a geometric random walk never decays. The percentage price change over t time periods is distributed as a normal distribution with mean equal to $\mu \cdot t$ and standard deviation equal to $\sigma \sqrt{t}$.
- The exact distribution of the future price knowing the initial price can be found. The price at time t is lognormally distributed with specific probability distribution parameters.
- Prices that follow a geometric random walk in continuous time never become negative.

The geometric random walk model is not perfect. However, its computational simplicity makes the geometric random walk and its variations the most widely used processes for modeling asset prices. The geometric random walk defined with log returns never becomes negative, because future prices are always a multiple of the initial stock price and a positive term. (See

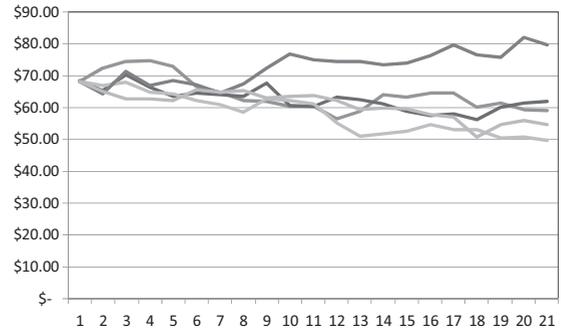


Figure 7 Five Paths of a Geometric Random Walk with $\mu = -0.0014$ and $\sigma = 0.0411$

Note: Although the drift is slightly negative, it is still possible to generate paths that generally increase over time.

Figure 7.) In addition, observed historical stock prices can actually be quite close to lognormal.

It is important to note that, actually, the assumption that log returns are normal is not required to justify the lognormal model for prices. If the distribution of log returns is non-normal, but the log returns are IID with finite variance, the sum of the log returns is asymptotically normal. (This is based on a version of the central limit theorem.) Stated differently, the log return process is approximately normal if we consider changes over sufficiently long intervals of time.

Price processes, however, are not always geometric random walks, even asymptotically. A very important assumption for the geometric random walk is that price increments are independently distributed; if the time series exhibits autocorrelation, the geometric random walk is not a good representation. We will see some models that incorporate considerations for autocorrelation and other factors later in this entry.

MEAN REVERSION

The geometric random walk provides the foundation for modeling the dynamics for asset prices of many different securities, including stock prices. However, in some cases it is not

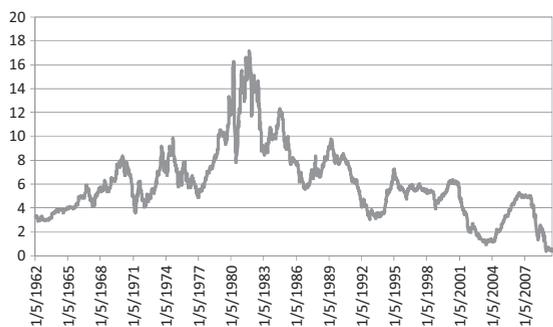


Figure 8 Weekly Data for One-Year Treasury Bill Rates: January 5, 1962–July 31, 2009

justified to assume that asset prices evolve with a particular drift, or can deviate arbitrarily far from some kind of a representative value. Interest rates, exchange rates, and the prices of some commodities are examples for which the geometric random walk does not provide a good representation over the long term. For instance, if the price of copper becomes high, copper mines would increase production in order to maximize profits. This would increase the supply of copper in the market, therefore decreasing the price of copper back to some equilibrium level. Consumer demand plays a role as well—if the price of copper becomes too high, consumers may look for substitutes, which would reduce the price of copper back to its equilibrium level.

Figure 8 illustrates the behavior of the one-year Treasury bill yield from the beginning of January 1962 through the end of July 2009. It can be observed that, even though the variability of Treasury bill rates has changed over time, there is some kind of a long-term average level of interest rates to which they return after deviating up or down. This behavior is known as *mean reversion*.

The simplest mean reversion (MR) model is similar to an arithmetic random walk, but the means of the increments change depending on the current price level. The price dynamics are represented by the equation

$$S_{t+1} = S_t + \kappa \cdot (\mu - S_t) + \sigma \cdot \tilde{\varepsilon}_t$$

where $\tilde{\varepsilon}_t$ is a standard normal random variable. The parameter κ is a nonnegative number that represents the speed of adjustment of the mean-reverting process—the larger its magnitude, the faster the process returns to its long-term mean. The parameter μ is the long-term mean of the process. When the current price S_t is lower than the long-term mean μ , the term $(\mu - S_t)$ is positive. Hence, on average there will be an upward adjustment to obtain the value of the price in the next time period, S_{t+1} . (We add a positive number, $\kappa \cdot (\mu - S_t)$, to the current price S_t .) By contrast, if the current price S_t is higher than the long-term mean μ , the term $(\mu - S_t)$ is negative. Hence, on average there will be a downward adjustment to obtain the value of the price in the next time period, S_{t+1} . (We add a negative number, $\kappa \cdot (\mu - S_t)$, to the current price S_t .) Thus, the mean-reverting process will behave in the way we desire—if the price becomes lower or higher than the long-term mean, it will be drawn back to the long-term mean.

In the case of the arithmetic and the geometric random walks, the cumulative volatility of the process increases over time. By contrast, in the case of mean reversion, as the number of steps increases, the variance peaks at

$$\frac{\sigma^2}{\kappa \cdot (2 - \kappa)}$$

In continuous time, this basic mean-reversion process is called the *Ornstein-Uhlenbeck process*. (See the last section of this entry.) It is widely used when modeling interest rates and exchange rates in the context of computing bond prices and prices of more complex fixed-income securities. When used in the context of modeling interest rates, this simple mean-reversion process is also referred to as the Vasicek model (see Vasicek, 1977).

The mean-reversion process suffers from some of the disadvantages of the arithmetic random walk—for example, it can technically become negative. However, if the long-run mean is positive, and the speed of mean reversion is large relative to the volatility, the price will be

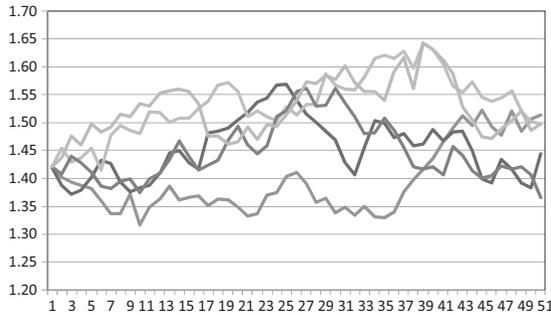


Figure 9 Five Paths with 50 Steps Each of a Mean-Reverting Process with $\mu = 1.4404$, $\kappa = 0.0347$, and $\sigma = 0.0248$

pulled back to the mean quickly when it becomes negative. Figure 9 contains an example of five paths generated from a mean-reverting process.

Simulation

The formula for the mean-reverting process makes it clear how paths for the mean-reverting random walk can be generated. As in the case of the arithmetic and the geometric random walks, all we need is a way of simulating the standard normal random variables $\tilde{\epsilon}_t$. We start with an initial price S_0 , which is known. We know the values of the drift μ , the speed of adjustment κ , and the volatility σ over one period. To generate the price at the next time period, S_1 , we add $\kappa \cdot (\mu - S_0)$ to S_0 , simulate a normal random variable from a standard normal distribution, multiply it by σ , and add it to $S_0 + \kappa \cdot (\mu - S_0)$. At the next step (time period 2), we use the price at time period 1 we already generated, S_1 , add to it $\kappa \cdot (\mu - S_1)$, simulate a new random variable from a standard normal distribution, multiply it by σ , and add it to $S_1 + \kappa \cdot (\mu - S_1)$. We proceed in the same way until we generate the desired number of steps of the random walk. For example, given a current price S , in Excel the price for the next time period can be generated with the formula

$$S + \kappa \cdot (\mu - S) + \sigma \cdot \text{NORMINV}(\text{RAND}(), 0, 1)$$

Parameter Estimation

In order to simulate paths of the mean-reverting random walk, we need estimates of the parameters (κ , μ , and σ). Again, we assume that these parameters remain constant over the time period of estimation. The equation for the mean-reverting process can be written as

$$S_{t+1} - S_t = \kappa \cdot (\mu - S_t) + \sigma \cdot \tilde{\epsilon}_t$$

or, equivalently,

$$S_{t+1} - S_t = \kappa \cdot \mu - \kappa \cdot S_t + \sigma \cdot \tilde{\epsilon}_t$$

This equation has the characteristics of a linear regression model, with the absolute price change ($S_{t+1} - S_t$) as the response variable and S_t as the explanatory variable. Given a historical series of T prices for an asset, we can therefore do the following to estimate κ , μ , and σ :

1. Compute the price changes ($S_{t+1} - S_t$) for each time period t , $t = 0, \dots, T-1$.
2. Run a linear regression with ($S_{t+1} - S_t$) as the response variable and S_t as the explanatory variable.
3. Verify that the estimates from the linear regression model are valid:
 - a. Plot the values of S_t versus ($S_{t+1} - S_t$). The points in the scatter plot should approximately vary around a straight line with no visible cyclical or other patterns.
 - b. The p -value for the coefficient in front of the explanatory variable S_t should be small, preferably less than 0.05. (The p -values of the regression coefficients are part of standard regression output for most software packages. Most generally, they measure the degree of significance of the regression coefficient for explaining the response variable in the regression.)
4. An estimate for the speed of adjustment of the mean-reversion process, κ , can be obtained as the negative of the coefficient in front of S_t . Since the speed of adjustment cannot be a negative number, if the coefficient in front of S_t is positive, the regression model

cannot be used for estimating the parameters of the mean reverting process.

5. An estimate for the long-term mean of the mean-reverting process, μ , can be obtained as the ratio of the intercept term estimated from the regression and the slope coefficient in front of S_t (if that slope coefficient is valid, i.e., negative and with low p -value).
6. An estimate for the volatility of the mean-reverting process, σ , can be obtained as the standard error of the regression. (The standard error of the regression is also part of standard regression output for statistical software packages and spreadsheet programs like Excel. It measures the standard deviation of the points around the regression line.)

Geometric Mean Reversion

A more advanced mean-reversion model that bears some similarity to the geometric random walk is the geometric mean reversion (GMR) model

$$S_{t+1} = S_t + \kappa \cdot (\mu - S_t) \cdot S_t + \sigma \cdot S_t \cdot \tilde{\epsilon}_t$$

(Note that this is a special case of the mean reversion model $S_{t+1} = S_t + \kappa \cdot (\mu - S_t) \cdot S_t + \sigma \cdot S_t^\gamma \cdot \tilde{\epsilon}_t$, where γ is a parameter selected in advance. The most commonly used models have $\gamma = 1$ or $\gamma = 1/2$.) The intuition behind this model is similar to the intuition behind the discrete version of the geometric random walk—the variability of the process changes with the current level of the price. However, the GMR model allows for incorporating mean reversion. Even though it is difficult to estimate the future price analytically from this model, it is easy to simulate. For example, given a current price S , in Excel the price for the next time period can be generated with the formula

$$S + \kappa * (\mu - S) * S + \sigma * S$$

*NORMINV(RAND(), 0, 1)

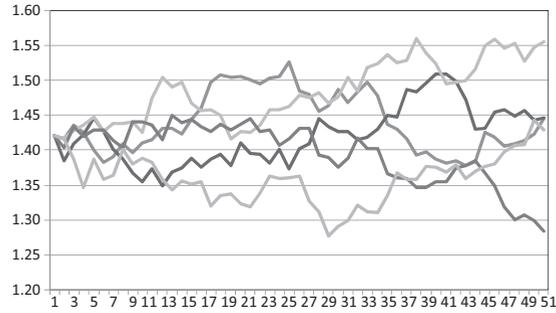


Figure 10 Five Paths with 50 Steps Each of a Geometric Mean Reversion Process with $\mu = 1.4464$, $\kappa = 0.0253$, and $\sigma = 0.0177$

Figure 10 contains an example of five paths generated from a geometric mean reversion model.

To estimate the parameters κ , μ , and σ to use in the simulation, we can use a series of T historical observations for the price of an asset. Assume that the parameters of the geometric mean reversion remain constant during the time period of estimation.

Note that the equation for the geometric mean-reverting random walk can be written as

$$\frac{S_{t+1} - S_t}{S_t} = \kappa \cdot (\mu - S_t) + \sigma \cdot \tilde{\epsilon}_t$$

or, equivalently, as

$$\frac{S_{t+1} - S_t}{S_t} = \kappa \cdot \mu - \kappa \cdot S_t + \sigma \cdot \tilde{\epsilon}_t$$

Again, this equation bears characteristics of a linear regression model, with the percentage price change $(S_{t+1} - S_t)/S_t$ as the response variable and S_t as the explanatory variable. Given a historical series of T prices of an asset, we can therefore do the following to estimate κ , μ , and σ :

1. Compute the percentage price changes $(S_{t+1} - S_t)/S_t$ for each time period t , $t = 0, \dots, T-1$.
2. Run a linear regression with $(S_{t+1} - S_t)/S_t$ as the response variable and S_t as the explanatory variable.

3. Verify that the estimates from the linear regression model are valid:
 - a. Plot the values of S_t versus $(S_{t+1} - S_t)/S_t$. The points in the scatter plot should approximately vary around a straight line with no visible cyclical or other patterns.
 - b. The p -value for the coefficient in front of the explanatory variable S_t should be small, preferably less than 0.05.
4. An estimate for the speed of adjustment of the mean-reverting process, κ , can be obtained as the negative of the coefficient in front of S_t . Since the speed of adjustment cannot be a negative number, if the coefficient in front of S_t is positive, the regression model cannot be used for estimating the parameters of the geometric mean-reverting process.
5. An estimate for the long-term mean of the mean-reverting process, μ , can be obtained as the ratio of the intercept term estimated from the regression and the slope coefficient in front of S_t (if that slope coefficient is valid, i.e., negative and with low p -value).
6. An estimate for the volatility of the mean-reverting process, σ , can be obtained as the standard error of the regression.

ADVANCED RANDOM WALK MODELS

The models we described so far provide building blocks for representing the asset price dynamics. However, observed real-world asset price dynamics has features that cannot be incorporated in these basic models. For example, asset prices exhibit correlation—both with each other and with themselves over time. Their volatility typically cannot be assumed to be constant. This section reviews several techniques for making asset price models more realistic depending on observed price behavior.

Correlated Random Walks

So far, we have discussed models for asset prices that assume that the dynamic processes for the prices of different assets evolve inde-

pendently of each other. This is an unrealistic assumption—it is expected that market conditions and other factors will have an impact on the prices of groups of assets simultaneously. For example, it is likely that stock prices for companies in the oil industry will generally move together, as will stock prices for companies in the telecommunications industry.

The argument that asset prices are codependent has theoretical and empirical foundations as well. If asset prices were independent random walks, then large portfolios would be fully diversified, have no variability, and therefore be completely deterministic. Empirically, this is not the case. Even large aggregates of stock prices, such as the S&P 500, exhibit random behavior.

If we make the assumption that log returns are jointly normally distributed, then their dependencies can be represented through the covariance matrix (equivalently, through the correlation matrix). It is worth noting that in general, covariance and correlation are not equivalent with dependence of random variables. Covariance and correlation measure only the strength of linear dependence between two random variables. However, in the case of a multivariate normal distribution, covariance and correlation are sufficient to represent dependence.

Let us give an example of how one can model two correlated stock prices assumed to follow geometric random walks. Suppose we are given two historical series of T observations each of observed asset prices for Stock 1 and Stock 2. We follow the steps described in the previous sections of this entry to estimate the drifts and the volatilities of the two processes. To estimate the correlation structure, we find the correlation between

$$\ln \left(\frac{S_{t+1}^{(1)}}{S_t^{(1)}} \right) \quad \text{and} \quad \ln \left(\frac{S_{t+1}^{(2)}}{S_t^{(2)}} \right)$$

where the indices (1) and (2) correspond to Stock 1 and Stock 2, respectively. For example, in Excel the correlation between two data series

stored in Array1 and Array2 can be computed with the function CORREL(Array1, Array2). This correlation can then be incorporated in the simulation. (Excel cannot simulate correlated normal random variables. A number of Excel add-ins for simulation are available, however, and they have the capability to do so. Such add-ins include @RISK (sold by Palisade Corporation, <http://www.palisade.com>), Crystal Ball (sold by Oracle, <http://www.oracle.com>), and Risk Solver (from Frontline Systems, the developers of the original Excel Solver, <http://www.solver.com>.) Basically, at every step, we generate correlated normal random variables, $\varepsilon_t^{(1)}$ and $\varepsilon_t^{(2)}$, with means of zero and with a given covariance structure. Those realizations of the correlated normal random variables are then used to compute the next period's Stock 1 price and the next period's Stock 2 price.

When we consider many different assets, the covariance matrix becomes very large and cannot be estimated accurately. Factor models can be used to reduce the dimension of the covariance structure. Multivariate random walks are in fact dynamic factor models for asset prices. A multifactor model for the return of asset i can be written in the following general form:

$$r_t^{(i)} = \mu^{(i)} + \sum_{k=1}^K \beta^{(i,k)} \cdot f_t^{(k)} + \varepsilon_t^{(i)}$$

where the K factors $f^{(k)}$ follow random walks, $\beta^{(i,k)}$ are the factor loadings, and $\varepsilon_t^{(i)}$ are normal random variables with zero means.

It is important to note that the covariance matrix cannot capture correlations at lagged times (i.e., correlations of dynamic nature). Furthermore, the assumptions that log returns behave as multivariate normal variables is not always applicable—some assets exhibit dependency of a nonlinear kind, which cannot be captured by the covariance or correlation matrix. Alternative tools for modeling covariability include copula functions and transfer entropies. (See, for example, Chapter 17 and Appendix B in Fabozzi, Focardi, and Kolm, 2006.)

Incorporating Jumps

Many of the dynamic asset price processes used in industry assume continuous sample paths, as was the case with the arithmetic, geometric, and the different mean-reverting random walks we considered earlier in this entry. However, there is empirical evidence that the prices of many securities incorporate jumps. The prices of some commodities, such as electricity and oil, are notorious for exhibiting “spikes.” The logarithm of a price process with jumps is not normally distributed, but is instead characterized by a high peak and heavy tails, which are more typical of market data than the normal distribution. Thus, more advanced models are needed to incorporate realistic price behavior.

A classical way to include jumps in models for asset price dynamics is to add a Poisson process to the process (geometric random walk or mean reversion) used to model the asset price. A Poisson process is a discrete process in which arrivals occur at random discrete points in time, and the times between arrivals follow an exponential distribution with average time between arrivals equal to $1/\lambda$. This means that the number of arrivals in a specific time interval follows a Poisson distribution with mean rate of arrival λ . The “jump” Poisson process is assumed to be independent of the underlying “smooth” random walk.

The Poisson process is typically used to figure out the times at which the jumps occur. The magnitude of the jumps itself could come from any distribution, although the lognormal distribution is often used for tractability.

Let us explain in more detail how one would model and simulate a geometric random walk with jumps. At every point in time, the process moves as a geometric random walk and updates the price S_t to S_{t+1} . If a jump happens, the size of the jump is added to S_t as well to obtain S_{t+1} . In order to avoid confusion about whether we have included the jump in the calculation, let us denote the price right before we find out whether a jump has occurred $S_{t+1}^{(-)}$, and keep the

total price for the next time period as S_{t+1} . We therefore have

$$S_{t+1}^{(-)} = S_t + \mu \cdot S_t + \sigma \cdot S_t \cdot \tilde{\varepsilon}_t$$

that is, $S_{t+1}^{(-)}$ is computed according to the normal geometric random walk rule. Now suppose that a jump of magnitude \tilde{J}_t occurs between time t and time $t+1$. Let us express the jump magnitude as a percentage of the asset price, that is, let

$$S_{t+1} = S_{t+1}^{(-)} \cdot \tilde{J}_t$$

If we restrict the magnitude of the jumps \tilde{J}_t to be nonnegative, we will make sure that the asset price itself does not become negative.

Let us now express the changes in price in terms of the jump size. Based on the relationship between S_{t+1} , $S_{t+1}^{(-)}$, and \tilde{J}_t , we can write

$$S_{t+1} - S_{t+1}^{(-)} = S_{t+1}^{(-)} \cdot (\tilde{J}_t - 1)$$

and hence

$$S_{t+1}^{(-)} = S_{t+1} - S_{t+1}^{(-)} \cdot (\tilde{J}_t - 1)$$

Thus, we can substitute this expression for $S_{t+1}^{(-)}$ and write the geometric random walk with jumps model as

$$S_{t+1} = S_t + \mu \cdot S_t + \sigma \cdot S_t \cdot \tilde{\varepsilon}_t + S_{t+1}^{(-)} \cdot (\tilde{J}_t - 1)$$

How would we simulate a path for the jump-geometric random walk process? Note that given the relationship between S_{t+1} , $S_{t+1}^{(-)}$, and \tilde{J}_t , we can write

$$\ln(S_{t+1}) = \ln(S_{t+1}^{(-)}) + \ln(\tilde{J}_t)$$

Since $S_{t+1}^{(-)}$ is the price resulting only from the geometric random walk at time t , we already know what $\ln(S_{t+1}^{(-)})$ is. Recall based on our discussion of the geometric random walk that

$$\ln(S_{t+1}^{(-)}) = \ln(S_t) + (\mu - 0.5 \cdot \sigma^2) + \sigma \tilde{\varepsilon}_t$$

Therefore, the overall equation will be

$$\begin{aligned} \ln(S_{t+1}) &= \ln(S_t) + (\mu - 0.5 \cdot \sigma^2) + \sigma \cdot \tilde{\varepsilon}_t \\ &\quad + \sum_i \ln(J_t^{(i)}) \end{aligned}$$

where $J_t^{(i)}$ are all the jumps that occur during the time period between t and $t+1$. This means that

$$S_{t+1} = S_t \cdot e^{\mu - 0.5 \cdot \sigma^2 + \sigma \cdot \tilde{\varepsilon}_t} \cdot \prod_i J_t^{(i)}$$

where the symbol Π denotes product. (If no jumps occurred between t and $t+1$, we set the product to 1.)

Hence, to simulate the price at time $t+1$, we need to simulate

- A standard normal random variable $\tilde{\varepsilon}_t$, as in the case of a geometric random walk.
- How many jumps occur between t and $t+1$.
- The magnitude of each jump.

For more details, see Pachamanova and Fabozzi (2010) and Glasserman (2004).

As Merton (1976) pointed out, if we assume that the jumps follow a lognormal distribution, then $\ln(\tilde{J}_t)$ is normal, and the simulation is even easier. See Glasserman (2004) for more advanced examples.

Stochastic Volatility

The models we considered so far all assumed that the volatility of the stochastic process remains constant over time. Empirical evidence suggests that the volatility changes over time, and more advanced models recognize that fact. Such models assume that the volatility parameter σ itself follows a random walk of some kind. Since there is some evidence that volatility tends to be mean-reverting, often different versions of mean-reversion models are used. For more details on stochastic volatility models and their simulation see, for example, Glasserman (2004) and Hull (2008).

STOCHASTIC PROCESSES

In this section, we provide an introduction to what is known as stochastic calculus. Our goal is not to achieve a working knowledge in the

subject, but rather to provide context for some of the terminology and the formulas encountered in the literature on modeling asset prices with random walks.

So far, we discussed random walks for which every step is taken at a specific discrete point in time. When the time increments are very small, almost zero in length, the equation of a random walk describes a stochastic process in continuous time. In this context, the arithmetic random walk model is known as a *generalized Wiener process* or *Brownian motion* (BM). The geometric random walk is referred to as *geometric Brownian motion* (GBM), and the arithmetic mean-reverting walk is the Ornstein-Uhlenbeck process described earlier.

Special notation is used to denote stochastic processes in continuous time. Increments are denoted by d or Δ . (For example, $(S_{t+1} - S_t)$ is denoted dS_t , meaning a change in S_t over an infinitely small interval.) The equations describing the process, however, have a very similar form to the equations we introduced earlier in this section:

$$dS_t = \mu dt + \sigma dW$$

Equations involving small changes (“differences”) in variables are referred to as *differential equations*. In words, the equation above reads “The change in the price S_t over a small time period dt equals the average drift μ multiplied by the small time change plus a random term equal to the volatility σ multiplied by dW , where dW is the increment of a Wiener process.” The Wiener process, or Brownian motion, is the fundamental building block for many of the classical asset price processes.

A standard Wiener process $W(t)$ has the following properties:

1. For any time $s < t$, the difference $W(t) - W(s)$ is a normal random variable with mean zero and variance $(t - s)$. It can be expressed as $\sqrt{t - s} \cdot \tilde{\epsilon}$, where $\tilde{\epsilon}$ is a standard normal random variable.
2. For any times $0 \leq t_1 < t_2 \leq t_3 < t_4$, the differences $(W(t_2) - W(t_1))$ and $(W(t_4) - W(t_3))$ (which are random variables) are independent. (These differences are the actual increments of the process at different points in time.) Note that independent implies uncorrelated.
3. The value of the Wiener process at the beginning is zero, $W(t_0) = 0$.

Using the new notation, the first two properties can be restated as

Property 1. The change dW during a small period of time dt is normally distributed with mean 0 and variance dt and can be expressed as $\sqrt{dt} \cdot \tilde{\epsilon}$.

Property 2. The values of dW for any two nonoverlapping time intervals are independent.

The arithmetic random walk can be obtained as a *generalized Wiener process*, which has the form

$$dS_t = a dt + b dW$$

The appeal of the generalized Wiener process is that we can find a closed-form expression for the price at any time period. Namely,

$$S_t = S_0 + a \cdot t + b \cdot W(t)$$

The generalized Wiener process is a special case of the more general class of *Ito processes*, in which both the drift term and the coefficient in front of the random term are allowed to be nonconstant. The equation for an Ito process is

$$dS_t = a(S, t) dt + b(S, t) dW$$

GBM and the Ornstein-Uhlenbeck process are both special cases of Ito processes.

In contrast to the generalized Wiener process, the equation for the Ito process does not allow us to write a general expression for the price at time t in closed form. However, an expression can be found for some special cases, such as GBM. We now show how this can be derived.

The main relevant result from stochastic calculus is the so-called Ito’s lemma, which states

the following. Suppose that a variable x follows an Ito process

$$dx_t = a(x, t) dt + b(x, t) dW$$

and let y be a function of x , that is,

$$y_t = f(x, t)$$

Then, y evolves according to the following differential equation:

$$dy_t = \left(\frac{\partial f}{\partial x} \cdot a + \frac{\partial f}{\partial t} + \frac{1}{2} \cdot \frac{\partial^2 f}{\partial x^2} \cdot b^2 \right) dt + \frac{\partial f}{\partial x} \cdot b \cdot dW$$

where the symbol ∂ is standard notation for the partial derivative of the function f with respect to the variable in the denominator. For example, $\partial f / \partial t$ is the derivative of the function f with respect to t assuming that all terms in the expression for f that do not involve t are constant. Respectively, ∂^2 denotes the second derivative of the function f with respect to the variable in the denominator, that is, the derivative of the derivative.

This expression shows that a function of a variable that follows an Ito process also follows an Ito process.

Although a rigorous proof of Ito's lemma is beyond the scope of this entry, we will provide some intuition. Let us see how we would go about computing the expression for y in Ito's lemma.

In ordinary calculus, we could obtain an expression for a function of a variable in terms of that variable by writing the Taylor series extension:

$$dy = \frac{\partial f}{\partial x} \cdot dx + \frac{\partial f}{\partial t} \cdot dt + \frac{1}{2} \cdot \frac{\partial^2 f}{\partial x^2} \cdot dx^2 + \frac{1}{2} \cdot \frac{\partial^2 f}{\partial t^2} \cdot dt^2 + \frac{\partial^2 f}{\partial x \partial t} \cdot dx dt + \dots$$

We will get rid of all terms of order dt^2 or higher, deeming them too small. We need to expand the terms that contain dx , however, because they will contain terms of order dt . We

have

$$dy = \frac{\partial f}{\partial x} \cdot (a(x, t) dt + b(x, t) dW) + \frac{\partial f}{\partial t} \cdot dt + \frac{1}{2} \cdot \frac{\partial^2 f}{\partial x^2} \cdot (a(x, t) dt + b(x, t) dW)^2$$

The last expression in parentheses, when expanded, becomes (dropping the arguments of a and b for notational convenience)

$$(a dt + b dW)^2 = a^2(dt)^2 + b^2(dW)^2 + 2ab \cdot dt \cdot dW = b^2 dt$$

To obtain this expression, we dropped the first and the last term in the expanded expression, because they are of order higher than dt . The middle term, $b^2(dW)^2$, in fact equals $b^2 \cdot dt$ as dt goes to 0. The latter is not an obvious fact, but it follows from the properties of the standard Wiener process. The intuition behind it is that the variance of $(dW)^2$ is of order dt^2 , so we can ignore it and treat the expression as deterministic and equal to its expected value. The expected value of $(dW)^2$ is in fact dt .

Substituting this expression back into the expression for dy , we obtain the expression in Ito's lemma.

Using Ito's lemma, let us derive the equation for the price at time t , S_t that was the basis for the exact simulation method for the geometric random walk. Suppose that S_t follows the GBM

$$dS_t = (\mu \cdot S_t) dt + (\sigma \cdot S_t) dW$$

We will use Ito's lemma to compute the equation for the process followed by the logarithm of the stock price. In other words, in the notation we used in the definition of Ito's lemma, we have

$$y_t = f(x, t) = \ln S_t$$

We also have

$$a = \mu \cdot S \quad \text{and} \quad b = \sigma \cdot S$$

Finally, we have

$$\frac{\partial f}{\partial x} = \frac{\partial(\ln S)}{\partial S} = \frac{1}{S} \quad \text{and} \quad \frac{\partial^2 f}{\partial x^2} = \frac{\partial(1/S)}{\partial S} = -\frac{1}{S^2}$$

Plugging into the equation for y in Ito's lemma, we obtain

$$\begin{aligned} d \ln S &= \left(\frac{1}{S} \cdot a + 0 + \frac{1}{2} \cdot \left(-\frac{1}{S^2} \right) \cdot b^2 \right) dt \\ &\quad + \frac{1}{S} \cdot b \cdot dW \\ &= \left(\mu - \frac{1}{2} \cdot \sigma^2 \right) dt + \sigma \cdot dW \end{aligned}$$

which is the equation we presented earlier. This also explains the presence of the

$$-\frac{1}{2} \cdot \sigma^2$$

term in the expression for the drift of the GBM.

KEY POINTS

- Models of asset dynamics include trees (such as binomial trees) and random walks (such as arithmetic, geometric, and mean-reverting random walks). Such models are called discrete when the changes in the asset price are assumed to happen at discrete time increments. When the length of the time increment is assumed to be infinitely small, we refer to them as stochastic processes in continuous time.
- The arithmetic random walk is an additive model for asset prices—at every time period, the new price is determined by the price at the previous time period plus a deterministic drift term and a random shock that is distributed as a normal random variable with mean equal to zero and a standard deviation proportional to the square root of the length of the time period. The probability distribution of future asset prices conditional on a known current price is normal.
- The arithmetic random walk model is analytically tractable and convenient; however, it has some undesirable features such as a nonzero probability that the asset price will become negative.
- The geometric random walk is a multiplicative model for asset prices—at every time period, the new price is determined by the price at the previous time period multiplied by a deterministic drift term and a random shock that is distributed as a lognormal random variable. The volatility of the process grows with the square root of the elapsed amount of time. The probability distribution of future asset prices conditional on a known current price is lognormal.
- The geometric random walk is not only analytically tractable, but is more realistic than the arithmetic random walk, because the asset price cannot become negative. It is widely used in practice, particularly for modeling stock prices.
- Mean reversion models assume that the asset price will meander, but will tend to return to a long-term mean at a speed called the speed of adjustment. They are particularly useful for modeling prices of some commodities, interest rates, and exchange rates.
- The codependence structure between the price processes for different assets can be incorporated directly (by computing the correlation between the random terms in their random walks), by using dynamic multifactor models, or by more advanced means such as copula functions and transfer entropies.
- A variety of more advanced random walk models are used to incorporate different assumptions, such as time-varying volatility and “spikes,” or jumps, in the asset price. They are not as tractable analytically as the classical random walk models, but can be simulated.
- The Wiener process, a stochastic process in continuous time, is a basic building block for many of the stochastic processes used to model asset prices. The increments of a Wiener process are independent, normally distributed random variables with variance proportional to the length of the time period.
- An Ito process is a generalized Wiener process with drift and volatility terms that can be functions of the asset price and time.

- An important result in stochastic calculus is Ito's lemma, which states that a variable that is a function of a variable that follows an Ito process follows an Ito process itself with specific drift and volatility terms.

REFERENCES

- Fabozzi, F. J., Focardi, S., and Kolm, P. N. (2006). *Financial Modeling of the Equity Markets: CAPM to Cointegration*. Hoboken, NJ: J. Wiley & Sons.
- Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. New York: Springer-Verlag.
- Hull, J. (2008). *Options, Futures and Other Derivatives*, 7th Edition. Upper Saddle River, NJ: Prentice Hall.
- Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Finance*, 29: 449–470.
- Pachamanova, D. A., and Fabozzi, F. J. (2010). *Simulation and Optimization in Finance: Modeling with MATLAB, @RISK, or VBA*. Hoboken, NJ: John Wiley & Sons.
- Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics*, 5: 177–188.

Arbitrage Pricing: Finite-State Models

SERGIO M. FOCARDI, PhD

Partner, The Intertek Group

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Arbitrage in its most basic form involves the simultaneous buying and selling of an asset at two different prices in two different markets. In real-world financial markets, arbitrage opportunities rarely, if ever, exist. Less obvious arbitrage opportunities exist in situations where a package of assets can be assembled that have a payoff (return) that is identical to an asset that is priced differently. A market is said to be a complete market if an arbitrary payoff can be replicated by a portfolio. The most fundamental principle in asset pricing theory is the absence of arbitrage opportunities.

The principle of *absence of arbitrage* or the no-arbitrage principle is perhaps the most fundamental principle of finance theory. In the presence of arbitrage opportunities, there is no trade-off between risk and returns because it is possible to make unbounded risk-free gains. The principle of absence of arbitrage is fundamental for understanding asset valuation in a competitive market. This entry discusses arbitrage pricing in a *finite-state*, discrete-time setting. However, it is important to note that there are well-known limits to arbitrage, first identified by Shleifer and Vishny (1997), resulting from restrictions imposed on rational traders and, as a result, pricing inefficiencies may exist for a period of time.

THE ARBITRAGE PRINCIPLE

Let's begin by defining what is meant by arbitrage. In its simple form, *arbitrage* is the

simultaneous buying and selling of an asset at two different prices in two different markets. The arbitrageur profits without risk by buying cheap in one market and simultaneously selling at the higher price in the other market. Such opportunities for arbitrage are rare. In fact, a single arbitrageur with unlimited ability to sell short could correct a mispricing condition by financing purchases in the underpriced market with proceeds from short sales in the overpriced market. This means that riskless arbitrage opportunities are short-lived.

Less obvious arbitrage opportunities exist in situations where a package of assets can produce a payoff (return) identical to an asset that is priced differently. This arbitrage relies on a fundamental principle of finance called the *law of one price*, which states that a given asset must have the same price regardless of the location where the asset is traded and the means by which one goes about creating that asset. The

law of one price implies that if the payoff of an asset can be synthetically created by a package of assets, the price of the package and the price of the asset whose payoff it replicates must be equal.

When a situation is discovered whereby the price of the package of assets differs from that of an asset with the same payoff, rational investors will trade these assets in such a way so as to restore price equilibrium. This market mechanism is founded on the fact that an arbitrage transaction does not expose the investor to any adverse movement in the market price of the assets in the transaction.

For example, consider how we can produce an arbitrage opportunity involving three assets A, B, and C. These assets can be purchased today at the prices shown below, and can each produce only one of two payoffs (referred to as State 1 and State 2) a year from now:

Asset	Price	Payoff in State 1	Payoff in State 2
A	\$70	\$50	\$100
B	60	30	120
C	80	38	112

While it is not obvious from the data presented above, an investor can construct a portfolio of assets A and B that will have the identical payoff as asset C in both State 1 and State 2. Let w_A and w_B be the proportion of assets A and B, respectively, in the portfolio. Then the payoff (i.e., the terminal value of the portfolio) under the two states can be expressed mathematically as follows:

- If State 1 occurs: $\$50 w_A + \$30 w_B$
- If State 2 occurs: $\$100 w_A + \$120 w_B$

We create a portfolio consisting of A and B that will reproduce the payoff of C regardless of the state that occurs one year from now. Here is how: For either condition (State 1 and State 2), we set the payoff of the portfolio equal to the payoff for C as follows:

- State 1: $\$50 w_A + \$30 w_B = \$38$
- State 2: $\$100 w_A + \$120 w_B = \$112$

We also know that $w_A + w_B = 1$. If we solved for the weights for w_A and w_B that would simultaneously satisfy the above equations, we would find that the portfolio should have 40% in asset A (i.e., $w_A = 0.4$) and 60% in asset B (i.e., $w_B = 0.6$). The cost of that portfolio will be equal to

$$(0.4)(\$70) + (0.6)(\$60) = \$64$$

Our portfolio (i.e., package of assets) comprised of assets A and B has the same payoff in State 1 and State 2 as the payoff of asset C. The cost of asset C is \$80 while the cost of the portfolio is only \$64. This is an arbitrage opportunity that can be exploited by buying assets A and B in the proportions given above and shorting (selling) asset C.

For example, suppose that \$1 million is invested to create the portfolio with assets A and B. The \$1 million is obtained by selling short asset C. The proceeds from the short sale of asset C provide the funds to purchase assets A and B. Thus, there would be no cash outlay by the investor. The payoffs for States 1 and 2 are shown below:

Asset	Investment	State 1	State 2
A	\$ 400,000	\$ 285,715	\$ 571,429
B	600,000	300,000	1,200,000
C	-1,000,000	-475,000	-1,400,000
Total	0	\$110,715	\$371,429

ARBITRAGE PRICING IN A ONE-PERIOD SETTING

We can describe the concepts of arbitrage pricing in a more formal mathematical context. It is useful to start in a simple one-period, finite-state setting as in the example of the previous section. This means that we consider only one period and that there is only a finite number M of states of the world. In this setting, asset prices can assume only a finite number of values.

The assumption of finite states is not as restrictive as it might appear. In practice,

security prices can only assume a finite number of values. Stock prices, for example, are not real numbers but integer fractions of a dollar. In addition, stock prices are nonnegative numbers and it is conceivable that there is some very high upper level that they cannot exceed. In addition, whatever simulation we might perform is a finite-state simulation given that the precision of computers is finite.

The finite number of states represents uncertainty. There is uncertainty because the world can be in any of the M states. At time 0 it is not known in what state the world will be at time 1. Uncertainty is quantified by probabilities but a lot of arbitrage pricing theory can be developed without any reference to probabilities. Suppose there are N securities. Each security i pays d_{ij} number of dollars (or of any other unit of account) in each state of the world j . The payoff of each security need not be a positive number. For instance, a derivative instrument might have negative payoffs in some states of the world. Therefore, in a one-period setting, the securities are formally represented by an $N \times M$ matrix $\mathbf{D} = \{d_{ij}\}$ where the d_{ij} entry is the payoff of security i in state j . The matrix \mathbf{D} can also be written as a set of N row vectors:

$$\mathbf{D} = \begin{bmatrix} \mathbf{d}_1 \\ \cdot \\ \mathbf{d}_N \end{bmatrix}, \quad \mathbf{d}_i = [d_{i1} \cdot d_{iM}]$$

where the M -vector \mathbf{d}_i represents the payoffs of security i in each of the M states.

Each security is characterized by a price S . Therefore, the set of N securities is characterized by an N -vector \mathbf{S} and an $N \times M$ matrix \mathbf{D} . Suppose, for instance, there are two states and three securities. Then the three securities are represented by

$$\mathbf{S} = \begin{bmatrix} S_1 \\ S_2 \\ S_3 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \\ d_{31} & d_{32} \end{bmatrix}$$

Every row of the \mathbf{D} matrix represents one security, every column one state. Note that in a one-period setting, prices are defined at time 0

while payoffs are defined at time 1. There is no payoff at time 0 and there is no price at time 1. A portfolio is represented by an N -vector of weights $\boldsymbol{\theta}$. In our example of a market with two states and three securities, a portfolio is a 3-vector:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

The market value $S_{\boldsymbol{\theta}}$ of a portfolio $\boldsymbol{\theta}$ at time 0 is a scalar given by the scalar product:

$$S_{\boldsymbol{\theta}} = \mathbf{S}\boldsymbol{\theta} = \sum_{i=1}^N S_i \theta_i$$

Its payoff $\mathbf{d}_{\boldsymbol{\theta}}$ at time 1 is the M -vector:

$$\mathbf{d}_{\boldsymbol{\theta}} = \mathbf{D}'\boldsymbol{\theta}$$

The price of a security and the market value of a portfolio can be a negative number. In the previous example of a two-state, three-security market we obtain

$$\begin{aligned} S_{\boldsymbol{\theta}} &= \mathbf{S}\boldsymbol{\theta} = S_1\theta_1 + S_2\theta_2 + S_3\theta_3 \\ \mathbf{d}_{\boldsymbol{\theta}} &= \mathbf{D}'\boldsymbol{\theta} = \begin{bmatrix} d_{11} & d_{21} & d_{31} \\ d_{12} & d_{22} & d_{32} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} \\ &= \begin{bmatrix} d_{11}\theta_1 + d_{21}\theta_2 + d_{31}\theta_3 \\ d_{12}\theta_1 + d_{22}\theta_2 + d_{32}\theta_3 \end{bmatrix} \end{aligned}$$

Let's introduce the concept of arbitrage in this simple setting. As we have seen, arbitrage is essentially the possibility of making money by trading without any risk. Therefore, we define an arbitrage as any portfolio $\boldsymbol{\theta}$ that has a negative market value $S_{\boldsymbol{\theta}} = \mathbf{S}\boldsymbol{\theta} < 0$ and a non-negative payoff $D_{\boldsymbol{\theta}} = \mathbf{D}'\boldsymbol{\theta} \geq 0$ or, alternatively, a nonpositive market value $S_{\boldsymbol{\theta}} = \mathbf{S}\boldsymbol{\theta} \leq 0$ and a positive payoff $D_{\boldsymbol{\theta}} = \mathbf{D}'\boldsymbol{\theta} > 0$.

State Prices

Next we define *state prices*. A state-price vector is a strictly positive M -vector $\boldsymbol{\psi}$ such that security prices can be written as $\mathbf{S} = \mathbf{D}\boldsymbol{\psi}$. In other words, given a state-price vector, if it exists,

security prices can be recovered as a weighted average of the securities' payoffs, where the state-price vector gives the weights. In the previous two-state, three-security example we can write:

$$\boldsymbol{\psi} = \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix}$$

$$\mathbf{S} = \mathbf{D}\boldsymbol{\psi}$$

$$\begin{bmatrix} S_1 \\ S_2 \\ S_3 \end{bmatrix} = \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \\ d_{31} & d_{32} \end{bmatrix} \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix} = \begin{bmatrix} d_{11}\psi_1 + d_{12}\psi_2 \\ d_{21}\psi_1 + d_{22}\psi_2 \\ d_{31}\psi_1 + d_{32}\psi_2 \end{bmatrix}$$

Given security prices and payoffs, state prices can be determined solving the system:

$$\begin{aligned} d_{11}\psi_1 + d_{12}\psi_2 &= S_1 \\ d_{21}\psi_1 + d_{22}\psi_2 &= S_2 \\ d_{31}\psi_1 + d_{32}\psi_2 &= S_3 \end{aligned}$$

This system admits solutions if and only if there are two linearly independent equations and the third equation is a linear combination of the other two. Note that this condition is necessary but not sufficient to ensure that there are state prices as state prices must be strictly positive numbers.

A portfolio $\boldsymbol{\theta}$ is characterized by payoffs $\mathbf{d}_\theta = \mathbf{D}'\boldsymbol{\theta}$. Its price is given, in terms of state prices, by: $S_\theta = \mathbf{S}\boldsymbol{\theta} = \mathbf{D}\boldsymbol{\psi}\boldsymbol{\theta} = \mathbf{d}_\theta\boldsymbol{\psi}$.

It can be demonstrated that there is no arbitrage if and only if there is a state-price vector. The formal demonstration is quite complicated given the inequalities that define an arbitrage portfolio. It hinges on the separating hyperplane theorem, which says that, given any two convex disjoint sets in R^M , it is possible to find a hyperplane separating them. A hyperplane is the locus of points x_i that satisfy a linear equation of the type:

$$a_0 + \sum_{i=1}^M a_i x_i = 0$$

Intuitively, however, it is clear that the existence of state prices ensures that the law of

one price introduced in the previous section is automatically satisfied. In fact, if there are state prices, two identical payoffs have the same price, regardless of how they are constructed. This is because the price of a security or of any portfolio is univocally determined as a weighted average of the payoffs, with the state prices as weights.

Risk-Neutral Probabilities

Let's now introduce the concept of risk-neutral probabilities. Given a state-price vector, consider the sum of its components $\psi_0 = \psi_1 + \psi_2 + \dots + \psi_M$. Normalize the state-price vector by dividing each component by the sum ψ_0 . The normalized state-price vector

$$\boldsymbol{\psi} = \{\psi_j\} = \left\{ \frac{\psi_j}{\psi_0} \right\}$$

is a set of positive numbers whose sum is one. These numbers can be interpreted as probabilities. They are not, in general, the real probabilities associated with states. They are called *risk-neutral probabilities*. We can then write

$$\mathbf{S} \frac{1}{\psi_0} = \mathbf{D}\boldsymbol{\psi}$$

We can interpret the above relationship as follows: The normalized security prices are their expected payoffs under these special probabilities. In fact, we can rewrite the above equation as

$$\bar{S}_i = \frac{S_i}{\psi_0} = E[d_i]$$

where expectation is taken with respect to risk-neutral probabilities. In this case, security prices are the discounted expected payoffs under these special risk-neutral probabilities.

Suppose that there is a portfolio $\bar{\boldsymbol{\theta}}$ such that $\mathbf{d}_{\bar{\theta}} = \mathbf{D}'\bar{\boldsymbol{\theta}} = \{1, 1, \dots, 1\}$. This portfolio can be one individual risk-free security. As we have seen above, $\mathbf{S}\boldsymbol{\theta} = \mathbf{d}_\theta\boldsymbol{\psi}$, which implies that $\psi_0 = \bar{\boldsymbol{\theta}}\mathbf{S}$ is the discount on riskless borrowing.

Complete Markets

Let's now define the concept of *complete markets*, a concept that plays a fundamental role in finance theory. In the simple setting of the one-period finite-state market, a complete market is one in which the set of possible portfolios is able to replicate an arbitrary payoff. Call $\text{span}(D)$ the set of possible portfolio payoffs, which is given by the following expression:

$$\text{span}(D) \equiv \{\mathbf{D}'\boldsymbol{\theta} : \boldsymbol{\theta} \in R^M\}$$

A market is complete if $\text{span}(D) = R^M$.

A one-period finite-state complete market is one where the equation

$$\mathbf{D}'\boldsymbol{\theta} = \boldsymbol{\xi} : \boldsymbol{\xi} \in R^M$$

always admits a solution. Recall from matrix algebra that this is the case if and only if the rank of \mathbf{D} is M . This means that there are at least M linearly independent payoffs—that is, there are as many linearly independent payoffs as there are states. Let's write down explicitly the system in the two-state, three-security market.

$$\begin{aligned} \mathbf{D}'\boldsymbol{\theta} &= \boldsymbol{\xi} \\ \begin{bmatrix} d_{11} & d_{21} & d_{31} \\ d_{12} & d_{22} & d_{32} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} &= \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \\ d_{11}\theta_1 + d_{21}\theta_2 + d_{31}\theta_3 &= \xi_1 \\ d_{12}\theta_1 + d_{22}\theta_2 + d_{32}\theta_3 &= \xi_2 \end{aligned}$$

This system of linear equations admits solutions if and only if the rank of the coefficient matrix is 2. This condition is not verified, for example, if the securities have the same payoff in each state. In this case, the relationship $\xi_1 = \xi_2$ must always be verified. In other words, the three securities can only replicate portfolios that have the same payoff in each state.

In this simple setting it is easy to associate risk-neutral probabilities with real probabilities. In fact, suppose that the vector of real probabilities \mathbf{p} is associated to states so that p_i is the probability of the i -th state. For any given

M -dimensional vector \mathbf{x} , we write its expected value under the real probabilities as

$$E[\mathbf{x}] = \mathbf{p}\mathbf{x} = \sum_{i=1}^M p_i x_i$$

It can be demonstrated that there is no arbitrage if and only if there is a strictly positive M -vector $\boldsymbol{\pi}$ such that: $\mathbf{S} = E[\mathbf{D}\boldsymbol{\pi}]$. Any such vector $\boldsymbol{\pi}$ is called a *state-price deflator*. To see this point, define

$$\pi_i = \frac{\psi_i}{p_i}$$

Prices can then be expressed as

$$S_i = \sum_{j=1}^M d_{ij}\psi_j = \sum_{j=1}^M p_j d_{ij} \frac{\psi_j}{p_i} = \sum_{j=1}^M p_j d_{ij} \pi_j$$

which demonstrates that $\mathbf{S} = E[\mathbf{D}\boldsymbol{\pi}]$.

We can now specialize the above calculations in the numerical case of the previous section. Recall that in the previous section we gave the example of three securities with the following prices and payoffs expressed in dollars:

$$\mathbf{S} = \begin{bmatrix} 70 \\ 60 \\ 80 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 50 & 100 \\ 30 & 120 \\ 38 & 112 \end{bmatrix}$$

We first compute the relative state prices:

$$\begin{aligned} 50\psi_1 + 100\psi_2 &= 70 \\ 30\psi_1 + 120\psi_2 &= 60 \\ 38\psi_1 + 112\psi_2 &= 80 \end{aligned}$$

Solving the first two equations, we obtain

$$\begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix} = \begin{bmatrix} 4/5 \\ 3/10 \end{bmatrix}$$

However, the third equation is not satisfied by these values for the state prices. As a consequence, there does not exist a state-price vector, which confirms that there are arbitrage opportunities as observed in the first section.

Now suppose that the price of security C is \$64 and not \$80. In this case, the third equation is satisfied and the state-price vector is the one shown above. Risk-neutral probabilities can now be easily computed. Here is how. First sum the two state prices: $\frac{4}{5} + \frac{3}{10} = \frac{11}{10}$ to obtain

$$\psi_0 = \psi_1 + \psi_2 = \frac{11}{10}$$

and consequently the risk-neutral probabilities:

$$\boldsymbol{\psi} = \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix} = \begin{bmatrix} \psi_1/\psi_0 \\ \psi_2/\psi_0 \end{bmatrix} = \begin{bmatrix} \frac{8}{11} \\ \frac{3}{11} \end{bmatrix}$$

Risk-neutral probabilities sum to one while state prices do not. We can now check if our market is complete. Write the following equations:

$$\begin{aligned} 50\theta_1 + 30\theta_2 + 38\theta_3 &= \xi_1 \\ 100\theta_1 + 120\theta_2 + 112\theta_3 &= \xi_2 \end{aligned}$$

The rank of the coefficient matrix is clearly 2 as the determinant of the first minor is different from zero:

$$\begin{bmatrix} 50 & 30 \\ 100 & 120 \end{bmatrix} = 50 \times 120 - 100 \times 30 = 300 \neq 0$$

Our sample market is therefore complete and arbitrage-free. A portfolio composed of the first two securities can replicate any payoff and the third security can be replicated as a portfolio of the first two.

ARBITRAGE PRICING IN A MULTIPERIOD FINITE-STATE SETTING

The above basic results can be extended to a multiperiod finite-state setting using probabilistic concepts. The economy is represented by a probability space $(\Omega, \mathfrak{S}, P)$ where Ω is the set of possible states, \mathfrak{S} is the algebra of events (recall that we are in a finite-state setting and therefore there are only a finite number of events), and P is a probability function. As the number of states is finite, finite probabilities $P(\{\omega\}) \equiv$

$P(\omega) \equiv p_\omega$ are defined for each state. There is only a finite number of dates from 0 to T .

Propagation of Information

The *propagation of information* is represented by a filtration \mathfrak{S}_t that, in the finite case, is equivalent to an information structure I_t . The latter is a discrete, hierarchical organization of partitions I_t with the following properties:

$$\begin{aligned} I_k &\equiv (\{A_{ik}\}); \quad k = 0, \dots, T; \quad i = 1, \dots, M_k; \\ 1 &= M_1 \leq \dots \leq M_k \leq \dots \leq M_T = M \\ A_{ik} \cap A_{jk} &= \emptyset \text{ if } i \neq j \text{ and } \bigcup_{i=1}^{M_k} A_{ik} = \Omega \end{aligned}$$

and, in addition, given any two sets A_{ik} , A_{jh} , with $h > k$, either their intersection is empty $A_{ik} \cap A_{jh} = \emptyset$ or $A_{ik} \supseteq A_{jh}$. In other words, the partitions become more refined with time.

Each security i is characterized by a payoff process d_t^i and by a price process S_t^i . In this finite-state setting, d_t^i and S_t^i are discrete variables that, given that there are M states, can be represented by M -vectors $\mathbf{d}_t^i = [d_t^i(\omega)]$ and $\mathbf{S}_t^i = [S_t^i(\omega)]$ where $d_t^i(\omega)$ and $S_t^i(\omega)$ are, respectively, the payoff and the price of the i -th asset at time t , $0 \leq t \leq T$ and in state $\omega \in \Omega$. All payoffs and prices are stochastic processes adapted to the filtration \mathfrak{S}_t . Given that d_t^i and S_t^i are adapted processes in a finite probability space, they have to assume a constant value on each partition of the information structure I_t . It is convenient to introduce the following notation:

$$\begin{aligned} d_{A_{jt}}^i &= d_t^i(\omega), \quad \omega \in A_{jt} \\ S_{A_{jt}}^i &= S_t^i(\omega), \quad \omega \in A_{jt} \end{aligned}$$

where $d_{A_{jt}}^i$ and $S_{A_{jt}}^i$ represent the constant values that the processes d_t^i and S_t^i assume on the states that belong to the sets A_{jt} of each partition I_t . There is $M_0 = 1$ value for $d_{A_{j0}}^i$ and $S_{A_{j0}}^i$, M_t values for $d_{A_{jt}}^i$ and $S_{A_{jt}}^i$ and $M_T = M$ values for $d_{A_{jT}}^i$ and $S_{A_{jT}}^i$. The same notation and the same consideration can be applied to any process adapted to the filtration \mathfrak{S}_t .

Trading Strategies

We have to define the meaning of trading strategies in this multiperiod setting. A trading strategy is a sequence of portfolios θ such that θ_t is the portfolio held at time t after trading. To ensure that there is no anticipation of information, each trading strategy θ must be an adapted process. The payoff d^θ generated by a trading strategy is an adapted process d_t^θ with the following time dynamics:

$$d_t^\theta = \theta_{t-1}(S_t + d_t) - \theta_t S_t$$

An arbitrage is a trading strategy whose payoff process is nonnegative and not always zero. In other words, an arbitrage is a trading strategy that is never negative and which is strictly positive for some instants and some states. Note that imposing the condition that payoffs are always nonnegative forbids any initial positive investment that is a negative payoff.

A consumption process is any nonnegative adapted process. Markets are said to be complete if any consumption process can be obtained as the payoff process of a trading strategy with some initial investment. Market completeness means that any nonnegative payoff process can be replicated with a trading strategy.

State-Price Deflator

We will now extend the concept of state-price deflator to a multiperiod setting. A state-price deflator is a strictly positive adapted process π_t such that the following set of M equations hold:

$$S_t^i = \frac{1}{\pi_t} E_t \left[\sum_{j=t+1}^T \pi_j d_j^i \right]$$

In other words, a state-price deflator is a strictly positive process such that prices S_t^i are random variables equal to the conditional expectation of discounted payoffs with respect to the filtration \mathfrak{F} . As noted above, in this finite-state setting a filtration is equivalent to an information structure I_t . Note that in the above stochastic equation—which is a set of M equations, one

for each state, the term on the left, the prices S_t^i , is an adapted process that, as mentioned, assumes constant values on each set of the partition I_t . The term on the right is a conditional expectation multiplied by a factor $1/\pi_t$. The process π_t is adapted by definition and, therefore, assumes constant values $\pi_{A_{it}}$ on each set of the partition I_t .

In this finite setting, conditional expectations are expectations computed with conditional probabilities. Conditional expectations are adapted processes. Therefore they assume one value at $t = 0$, M_j values for $t = j$, and M values at the last date.

To illustrate the above, let's write down explicitly the above equation in terms of the notation $d_{A_{jt}}^i$ and $S_{A_{jt}}^i$. Note first that

$$P(\{\omega\}|A_{kt}) = \frac{P(\{\omega\} \cap A_{kt})}{P(A_{kt})} = \frac{P(\{\omega\})}{P(A_{kt})},$$

if $\omega \in A_{kt}$, 0 if $\omega \notin A_{kt}$

Given that the probability space is finite,

$$P(A_{jt}) = \sum_{\omega \in A_{jt}} p_\omega$$

As we defined $P(\{\omega\}) \equiv p_\omega$, the previous equation becomes

$$\begin{aligned} P(\{\omega\}|A_{kt}) &= \frac{P(\{\omega\} \cap A_{kt})}{P(A_{kt})} = \frac{P(\{\omega\})}{P(A_{kt})} \\ &= \frac{p_\omega}{\left(\sum_{\omega \in A_{kt}} p_{\omega} \right)} \end{aligned}$$

if $\omega \in A_{kt}$, 0 if $\omega \notin A_{kt}$.

Pricing Relationships

We can now write the pricing relationship as follows:

$$\begin{aligned} S_{A_{kt}}^i &= \frac{1}{\pi_{A_{kt}}} \left[\sum_{\omega \in A_{kt}} \left(P(\{\omega\}|A_{kt}) \left(\sum_{j=t+1}^T \pi_j(\omega) d_j^i(\omega) \right) \right) \right] \\ &= \frac{1}{\pi_{A_{kt}}} \left[\sum_{\omega \in A_{kt}} \left(\frac{p_\omega}{\left(\sum_{\omega \in A_{kt}} p_\omega \right)} \left(\sum_{j=t+1}^T \pi_j(\omega) d_j^i(\omega) \right) \right) \right] \end{aligned}$$

$A_{kt} \in I_t, 1 \leq k \leq M_t$

The above formulas generalize to any trading strategy. In particular, if there is a state-price deflator, the market value of any trading strategy is given by

$$\begin{aligned} \theta_t \times \mathbf{S}_t &= \frac{1}{\pi_t} E \left[\sum_{j=t+1}^T \pi_j d_j^\theta \right] \\ (\theta_t \mathbf{S}_t)_{A_{kt}} &= \frac{1}{\pi_{A_{kt}}} \left[\sum_{\omega \in A_{kt}} \left(P(\{\omega\} | A_{kt}) \left(\sum_{j=t+1}^T \pi_j(\omega) d_j^\theta(\omega) \right) \right) \right] \\ &= \frac{1}{\pi_{A_{kt}}} \left[\sum_{\omega \in A_{kt}} \left(\frac{p_\omega}{\left(\sum_{\omega \in A_{kt}} p_\omega \right)} \left(\sum_{j=t+1}^T \pi_j(\omega) d_j^\theta(\omega) \right) \right) \right] \end{aligned}$$

It is possible to demonstrate that the payoff-price pair (d_t^i, S_t^i) admits no arbitrage if and only if there is a state-price deflator. These concepts and formulas generalize those of a one-period setting to a multiperiod setting.

Given a payoff-price pair (d_t^i, S_t^i) it is possible to compute the stateprice deflator, if it exists, from the previous equations. In fact, it is possible to write a set of linear equations in the π_t, π_{t-1} for each period. One can proceed backward from the period T to period 1 writing a homogeneous system of linear equations. As the system is homogeneous, one of the variables can be arbitrarily fixed; for example, the initial value π_0 can be assumed equal to 1. If the system admits nontrivial solutions and if all solutions are strictly positive, then there are state-price deflators.

To illustrate the above, let's write down explicitly the previous formulas for prices, extending the example of the previous section to a two-period setting. We assume there are three securities and two periods, that is, three dates $(0,1,2)$ and four states, indicated with the integers $1,2,3,4$, so that $\Omega = \{1,2,3,4\}$. Assume that the information structure is given by the following partitions of events:

$$\begin{aligned} I_0 &\equiv (I_0 \equiv \{A_{1,0}\}), \quad I_1 \equiv \{A_{1,1}, A_{2,1}\}, \\ I_2 &\equiv \{A_{1,2}, A_{2,2}, A_{3,2}, A_{4,2}\} \end{aligned}$$

$$\begin{aligned} A_{1,0} &= \{1 + 2 + 3 + 4\}, \quad A_{1,1} = \{1 + 2\}, \\ A_{2,1} &= \{3 + 4\} \\ A_{1,2} &= \{1\}, \quad A_{2,2} = \{2\}, \quad A_{3,2} = \{3\}, \quad A_{4,2} = \{4\} \end{aligned}$$

where we use $+$ to indicate logical union, so that, for example, $\{1 + 2\}$ is the event formed by states 1 and 2. The interpretation of the above notation is the following. At time zero the world can be in any possible state, that is, the securities can take any possible path. Therefore the partition at time zero is formed by the event $\{1 + 2 + 3 + 4\}$. At time 1, the set of states is partitioned into two mutually exclusive events, $\{1 + 2\}$ or $\{3 + 4\}$. At time 2 the partition is formed by all individual states. Note that this is a particular example; different partitions would be logically admissible.

Figure 1 represents the above structure. Each security is characterized by a price process and a payoff process adapted to the information structure. Each process is a collection of three discrete random variables indexed with the time indexes $0,1,2$. Each discrete random variable is a 4-vector as it assumes as many values as states. However, as processes are adapted, they must assume the same value on each partition of the information structure. Note also that payoffs are zero at date zero and prices are zero at date 2. Therefore, in this example, we can put together these vectors in two 3×4 matrices for

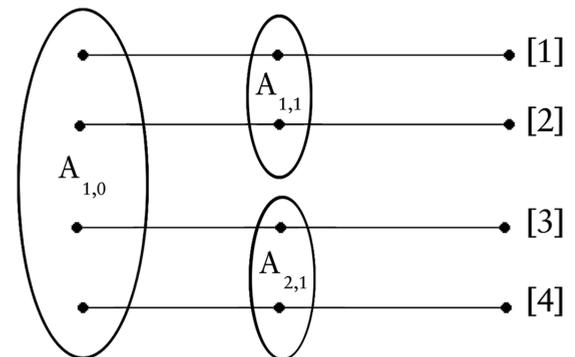


Figure 1 An Information Structure with Four States and Three Dates

each security as follows

$$\{S_t^i(\omega)\} \equiv \begin{bmatrix} S_0^i(1) & S_1^i(1) & 0 \\ S_0^i(2) & S_1^i(2) & 0 \\ S_0^i(3) & S_1^i(3) & 0 \\ S_0^i(4) & S_1^i(4) & 0 \end{bmatrix};$$

$$\{d_t^i(\omega)\} \equiv \begin{bmatrix} 0 & d_1^i(1) & d_2^i(1) \\ 0 & d_1^i(2) & d_2^i(2) \\ 0 & d_1^i(3) & d_2^i(3) \\ 0 & d_1^i(4) & d_2^i(4) \end{bmatrix}$$

The following relationships hold:

$$\begin{aligned} S_0^i(1) &= S_0^i(2) = S_0^i(3) = S_0^i(4) = S_{A_{1,0}}^i; \\ S_1^i(1) &= S_1^i(2) = S_{A_{1,1}}^i; \\ S_1^i(3) &= S_1^i(4) = S_{A_{2,1}}^i; \\ d_1^i(1) &= d_1^i(2) = d_{A_{1,1}}^i; \quad d_1^i(3) = d_1^i(4) = d_{A_{2,1}}^i \end{aligned}$$

where, as above, $S_t^i(\omega)$ is the price of security i in state ω at moment t and $d_t^i(\omega)$ is the payoff of security i in state ω at time t with the restriction that processes must assume the same value on partitions. This is because processes are adapted to the information structure so that there is no anticipation of information. One must not be able to discriminate at time 0 events that will be revealed at time 1 and so on.

Observe that there is no payoff at time 0 and no price at time 2 and that the payoffs at time 2 have to be intended as the final liquidation of the security as in the one-period case. Payoffs at time 1, on the other hand, are intermediate payments. Note that the number of states is chosen arbitrarily for illustration purposes. Each state of the world represents a path of prices and payoffs for the set of three securities. To keep the example simple, we assume that of all the possible paths of prices and payoffs only four are possible.

The state-price deflator can be represented as follows:

$$\{\pi_t(\omega)\} \equiv \begin{bmatrix} \pi_0(1) & \pi_1(1) & \pi_2(1) \\ \pi_0(2) & \pi_1(2) & \pi_2(2) \\ \pi_0(3) & \pi_1(3) & \pi_2(3) \\ \pi_0(4) & \pi_1(4) & \pi_2(4) \end{bmatrix}$$

$$\pi_0(1) = \pi_0(2) = \pi_0(3) = \pi_0(4)$$

$$\pi_1(1) = \pi_1(2) = \pi_1(3) = \pi_1(4)$$

A probability p_ω is assigned to each of the four states of the world. The probability of each event is simply the sum of the probabilities of its states. We can write down the formula for security prices in this way:

$$\begin{aligned} S_{A_{1,2}}^i &= S_2^i(1) = S_{A_{2,2}}^i = S_2^i(2) = S_{A_{3,2}}^i = S_2^i(3) \\ &= S_{A_{4,2}}^i = S_2^i(4) = 0 \end{aligned}$$

$$\begin{aligned} S_{A_{1,1}}^i &= S_1^i(1) = S_1^i(2) \\ &= \frac{1}{\pi_{A_{1,1}}} [P(A_{1,2}|A_{1,1})\pi_2(1)d_2^i(1) \\ &\quad + P(A_{2,2}|A_{1,1})\pi_2(2)d_2^i(2)] \\ &= \frac{1}{\pi_{A_{1,1}}} \left[\frac{p_1}{p_1 + p_2} \pi_2(1)d_2^i(1) \right. \\ &\quad \left. + \frac{p_2}{p_1 + p_2} \pi_2(2)d_2^i(2) \right] \end{aligned}$$

$$\begin{aligned} S_{A_{2,1}}^i &= S_1^i(3) = S_1^i(4) \\ &= \frac{1}{\pi_{A_{2,1}}} [P(A_{3,2}|A_{2,1})\pi_2(3)d_2^i(3) \\ &\quad + P(A_{4,2}|A_{2,1})\pi_2(4)d_2^i(4)] \\ &= \frac{1}{\pi_{A_{2,1}}} \left[\frac{p_3}{p_3 + p_4} \pi_2(3)d_2^i(3) \right. \\ &\quad \left. + \frac{p_4}{p_3 + p_4} \pi_2(4)d_2^i(4) \right] \end{aligned}$$

$$\begin{aligned} S_{A_{1,0}}^i &= \{p_1[\pi_{A_{1,1}}d_{A_{1,1}}^i + \pi_2(1)d_2^i(1)] + p_2[\pi_{A_{1,1}}d_{A_{1,1}}^i \\ &\quad + \pi_2(2)d_2^i(2)] + p_3[\pi_{A_{1,2}}d_{A_{1,2}}^i + \pi_2(3)d_2^i(3)] \\ &\quad + p_4[\pi_{A_{1,2}}d_{A_{1,2}}^i + \pi_2(4)d_2^i(4)]\} \end{aligned}$$

These equations illustrate how to compute the state-price deflator knowing prices,

payoffs, and probabilities. They form a homogeneous system of linear equations in $\pi_2(1), \pi_2(2), \pi_2(3), \pi_2(4), \pi_{A_{1,1}}, \pi_{A_{2,1}}, \pi_{A_{1,0}}$.

$$\begin{aligned} p_1 d_2^i(1) \pi_2(1) + p_2 d_2^i(2) \pi_2(2) - S_{A_{1,1}}^i (p_1 + p_2) \\ \pi_{A_{1,1}} &= 0 \\ p_3 d_4^i(3) \pi_2(3) + p_4 d_4^i(4) \pi_2(4) - S_{A_{2,1}}^i (p_3 + p_4) \\ \pi_{A_{2,1}} &= 0 \\ p_1 d_2^i(1) \pi_2(1) + p_2 d_2^i(2) \pi_2(2) + p_3 d_2^i(3) \pi_2(3) \\ + p_4 d_4^i(4) \pi_2(4) + (p_1 + p_2) d_{A_{1,1}}^i \pi_{A_{1,1}} \\ + (p_3 + p_4) d_{A_{2,3}}^i \pi_{A_{2,3}} - S_{A_{1,0}}^i \pi_{A_{1,0}} &= 0 \end{aligned}$$

Substituting, we obtain

$$\begin{aligned} p_1 d_2^i(1) \pi_2(1) + p_2 d_2^i(2) \pi_2(2) - S_{A_{1,1}}^i (p_1 + p_2) \\ \pi_{A_{1,1}} &= 0 \\ p_3 d_4^i(3) \pi_2(3) + p_4 d_4^i(4) \pi_2(4) - S_{A_{2,1}}^i (p_3 + p_4) \\ \pi_{A_{2,1}} &= 0 \\ [(p_1 + p_2) S_{A_{1,1}}^i + (p_1 + p_2) d_{A_{1,1}}^i] \pi_{A_{1,1}} \\ + [(p_3 + p_4) S_{A_{2,1}}^i + (p_3 + p_4) d_{A_{2,1}}^i] \pi_{A_{2,1}} \\ - S_{A_{1,0}}^i \pi_{A_{1,0}} &= 0 \end{aligned}$$

This homogeneous system must admit a strictly positive solution to yield a state-price deflator. There are seven unknowns. However, as the system is homogeneous, if nontrivial so-

lutions exist, one of the unknowns can be arbitrarily fixed, for example $\pi_{A_{1,0}}$. Therefore, six independent equations are needed. Each asset provides two conditions, so a minimum of three assets are needed.

To illustrate the point, we assume that all states (which are also events in this discrete example) have the same probability 0.25. Thus the events of the information structure have the following probabilities: the single event at time zero has probability 1, the two events at time 1 have probability 0.5, and the four events at time 2 coincide with individual states and have probability 0.25. Conditional probabilities are shown in Table 1.

For illustration purposes, let's write the following matrices for payoffs for each security at each date in each state:

$$\begin{aligned} \{d_1^i(\omega)\} &\equiv \begin{bmatrix} 0 & 15 & 50 \\ 0 & 15 & 100 \\ 0 & 20 & 70 \\ 0 & 20 & 110 \end{bmatrix}; \{d_2^i(\omega)\} \equiv \begin{bmatrix} 0 & 8 & 30 \\ 0 & 8 & 120 \\ 0 & 15 & 40 \\ 0 & 15 & 140 \end{bmatrix}; \\ \{d_3^i(\omega)\} &\equiv \begin{bmatrix} 0 & 5 & 38 \\ 0 & 5 & 112 \\ 0 & 8 & 42 \\ 0 & 8 & 130 \end{bmatrix} \end{aligned}$$

Table 1 Conditional Probabilities

$P(A_{1,1} A_{1,0}) = \frac{P(A_{1,1} \cap A_{1,0})}{P(A_{1,0})} = \frac{P\{1+2\}}{P\{1+2+3+4\}} = 0.5$	$P(A_{2,1} A_{1,0}) = \frac{P(A_{2,1} \cap A_{1,0})}{P(A_{1,0})} = \frac{P\{3+4\}}{P\{1+2+3+4\}} = 0.5$
$P(A_{1,2} A_{1,0}) = \frac{P(A_{1,2} \cap A_{1,0})}{P(A_{1,0})} = \frac{P\{1\}}{P\{1+2+3+4\}} = 0.25$	$P(A_{2,2} A_{1,0}) = \frac{P(A_{2,2} \cap A_{1,0})}{P(A_{1,0})} = \frac{P\{2\}}{P\{1+2+3+4\}} = 0.25$
$P(A_{3,2} A_{1,0}) = \frac{P(A_{3,2} \cap A_{1,0})}{P(A_{1,0})} = \frac{P\{3\}}{P\{1+2+3+4\}} = 0.25$	$P(A_{4,2} A_{1,0}) = \frac{P(A_{4,2} \cap A_{1,0})}{P(A_{1,0})} = \frac{P\{4\}}{P\{1+2+3+4\}} = 0.25$
$P(A_{1,2} A_{1,1}) = \frac{P(A_{1,2} \cap A_{1,1})}{P(A_{1,1})} = \frac{P\{1\}}{P\{1+2\}} = \frac{0.25}{0.5} = 0.5$	$P(A_{1,2} A_{2,1}) = \frac{P(A_{1,2} \cap A_{2,1})}{P(A_{2,1})} = \frac{P\{\emptyset\}}{P\{1+2\}} = 0$
$P(A_{2,2} A_{1,1}) = \frac{P(A_{2,2} \cap A_{1,1})}{P(A_{1,1})} = \frac{P\{2\}}{P\{1+2\}} = \frac{0.25}{0.5} = 0.5$	$P(A_{2,2} A_{2,1}) = \frac{P(A_{2,2} \cap A_{2,1})}{P(A_{2,1})} = \frac{P\{\emptyset\}}{P\{1+2\}} = 0$
$P(A_{3,2} A_{1,1}) = \frac{P(A_{3,2} \cap A_{1,1})}{P(A_{1,1})} = \frac{P\{\emptyset\}}{P\{1+2\}} = 0$	$P(A_{3,2} A_{2,1}) = \frac{P(A_{3,2} \cap A_{2,1})}{P(A_{2,1})} = \frac{P\{3\}}{P\{3+4\}} = 0.5$
$P(A_{4,2} A_{1,1}) = \frac{P(A_{4,2} \cap A_{1,1})}{P(A_{1,1})} = \frac{P\{\emptyset\}}{P\{1+2\}} = 0$	$P(A_{4,2} A_{2,1}) = \frac{P(A_{4,2} \cap A_{2,1})}{P(A_{2,1})} = \frac{P\{4\}}{P\{3+4\}} = 0.5$

We will assume that the state-price deflator is the following given process:

$$\{\pi_{t(\omega)}\} = \begin{bmatrix} 1 & 0.8 & 0.7 \\ 1 & 0.8 & 0.75 \\ 1 & 0.9 & 0.75 \\ 1 & 0.9 & 0.8 \end{bmatrix}$$

Each price is computed according to the previous equations. For example, calculations related to asset 1 are as follows:

$$S_2^1(1) = S_2^1(2) = S_2^1(3) = S_2^1(4) = 0$$

$$S_{A_{1,1}}^1 = \frac{1}{0.8}(0.5 \times 0.7 \times 50 + 0.5 \times 0.75 \times 100) = 68.75$$

$$S_{A_{2,1}}^1 = \frac{1}{0.9}(0.5 \times 0.75 \times 70 + 0.5 \times 0.8 \times 110) = 78.05$$

$$S_{A_{1,0}}^1 = \frac{1}{1}[0.25(0.8 \times 15 + 0.7 \times 50) + 0.25(0.8 \times 15 + 0.75 \times 100) + 0.25(0.9 \times 20 + 0.75 \times 70) + 0.25(0.9 \times 20 + 0.8 \times 110)] = 68.75$$

$$S_2^2(1) = S_2^2(2) = S_2^2(3) = S_2^2(4) = 0$$

$$S_{A_{1,1}}^2 = \frac{1}{0.8}(0.5 \times 0.7 \times 30 + 0.5 \times 0.75 \times 120) = 69.37$$

$$S_{A_{2,1}}^2 = \frac{1}{0.9}(0.5 \times 0.75 \times 40 + 0.5 \times 0.8 \times 140) = 78.88$$

$$S_{A_{1,0}}^2 = \frac{1}{1}[0.25(0.8 \times 8 + 0.7 \times 30) + 0.25(0.8 \times 8 + 0.75 \times 120) + 0.25(0.9 \times 15 + 0.75 \times 40) + 0.25(0.9 \times 15 + 0.8 \times 140)] = 73.2$$

$$S_2^3(1) = S_2^3(2) = S_2^3(3) = S_2^3(4) = 0$$

$$S_{A_{1,1}}^3 = \frac{1}{0.8}(0.5 \times 0.7 \times 38 + 0.5 \times 0.75 \times 112) = 69.12$$

$$S_{A_{2,1}}^3 = \frac{1}{0.9}(0.5 \times 0.75 \times 42 + 0.5 \times 0.8 \times 130) = 75.27$$

$$S_{A_{1,0}}^3 = \frac{1}{1}[0.25(0.8 \times 5 + 0.7 \times 38) + 0.25(0.8 \times 5 + 0.75 \times 112) + 0.25(0.9 \times 8 + 0.75 \times 42) + 0.25(0.9 \times 8 + 0.8 \times 130)] = 67.125$$

With the above equations we computed prices from payoffs and state-price deflators. If prices and payoffs were given, we could compute state-price deflators from the homogeneous system for state prices established above. Suppose that the following price processes were given:

$$\{S_t^1(\omega)\} = \begin{bmatrix} 68.75 & 68.75 & 0 \\ 68.75 & 68.75 & 0 \\ 68.75 & 78.05 & 0 \\ 68.75 & 78.05 & 0 \end{bmatrix}$$

$$\{S_t^2(\omega)\} = \begin{bmatrix} 73.2 & 69.37 & 0 \\ 73.2 & 69.37 & 0 \\ 73.2 & 78.88 & 0 \\ 73.2 & 78.88 & 0 \end{bmatrix}$$

$$\{S_t^3(\omega)\} = \begin{bmatrix} 67.125 & 69.12 & 0 \\ 67.125 & 69.12 & 0 \\ 67.125 & 75.27 & 0 \\ 67.125 & 75.27 & 0 \end{bmatrix}$$

We could then write the following system of equations to compute state-price deflators:

$$\begin{aligned} 0.25 \times 50 \times \pi_2(1) + 0.25 \times 100 \times \pi_2(2) - 68.75 \times 0.5 \times \pi_{A_{1,1}} &= 0 \\ 0.25 \times 70 \times \pi_2(1) + 0.25 \times 110 \times \pi_2(2) - 78.05 \times 0.5 \times \pi_{A_{1,1}} &= 0 \\ (55 \times 0.5 + 0.5 \times 15) \times \pi_{A_{1,1}} + (70.25 \times 0.5 + 0.5 \times 20) \times \pi_{A_{2,1}} - 68.75 \times \pi_{A_{1,0}} &= 0 \\ 0.25 \times 30 \times \pi_2(1) + 0.25 \times 120 \times \pi_2(2) - 69.37 \times 0.5 \times \pi_{A_{1,1}} &= 0 \\ 0.25 \times 40 \times \pi_2(1) + 0.25 \times 140 \times \pi_2(2) - 78.88 \times 0.5 \times \pi_{A_{1,1}} &= 0 \\ (55.5 \times 0.5 + 0.5 \times 8) \times \pi_{A_{1,1}} + (71 \times 0.5 + 0.5 \times 15) \times \pi_{A_{2,1}} - 73.2 \times \pi_{A_{1,0}} &= 0 \end{aligned}$$

$$\begin{aligned}
& 0.25 \times 38 \times \pi_2(1) + 0.25 \times 115 \times \pi_2(2) \\
& \quad - 69.12 \times 0.5 \times \pi_{A_{1,1}} = 0 \\
& 0.25 \times 42 \times \pi_2(1) + 0.25 \times 130 \times \pi_2(2) \\
& \quad - 75.27 \times 0.5 \times \pi_{A_{1,1}} = 0 \\
& (55 \times 0.5 + 0.5 \times 15) \times \pi_{A_{1,1}} + (70.25 \times 0.5 \\
& \quad + 0.5 \times 20) \times \pi_{A_{2,1}} - 67.125 \times \pi_{A_{1,0}} = 0
\end{aligned}$$

It can be verified that this system, obviously, is solvable and returns the same state-price deflators as in the previous example.

Equivalent Martingale Measures

We now introduce the concept and properties of *equivalent martingale measures*. This concept has become fundamental for the technology of derivative pricing. The idea of equivalent martingale measures is the following. A martingale is a process X_t such that at any time t its conditional expectation at time s , $s > t$ coincides with its present value: $X_t = E_t[X_s]$. In discrete time, a martingale is a process such that its value at any time is equal to its conditional expectation one step ahead. In our case, this principle can be expressed in a different but equivalent way by stating that prices are the discounted expected values of future payoffs. The law of iterated expectation then implies that price plus payoff processes are martingales.

In fact, assume that we can write

$$S_t = E_t \left[\sum_{j=t+1}^T d_j \right]$$

then the following relationship holds:

$$\begin{aligned}
S_t &= E_t \left[\sum_{j=t+1}^T d_j \right] = E_t \left[d_{t+1} + E_{t+1} \left[\sum_{j=t+1+1}^T d_j \right] \right] \\
&= E_t[d_{t+1} + S_{t+1}]
\end{aligned}$$

Given a probability space, price processes are not, in general, martingales. However it can be demonstrated that, in the absence of arbitrage, there is an artificial probability measure in which all price processes, appropriately dis-

counted, become martingales. More precisely, we will see that in the absence of arbitrage there is an artificial probability measure Q in which the following discounted present value relationship holds:

$$S_t^i = E_t^Q \left[\sum_{j=t+1}^T \frac{d_j^i}{R_{t,j}} \right]$$

We can rewrite this equation explicitly as follows:

$$\begin{aligned}
S_t^i &= E_t^Q \left[\sum_{j=t+1}^T \frac{d_j^i}{R_{t,j}} \right] \\
&= E_t^Q \left[\frac{d_{t+1}^i}{R_{t,t+1}} + \frac{1}{R_{t,t+1}} \sum_{j=t+2}^T \frac{d_j^i}{R_{t+1,j}} \right] \\
&= E_t^Q \left[\frac{d_{t+1}^i}{R_{t,t+1}} + \frac{E_{t+1}^Q}{R_{t,t+1}} \left[\sum_{j=t+2}^T \frac{d_j^i}{R_{t,j}} \right] \right] \\
&= E_t^Q \left[\frac{d_{t+1}^i + S_{t+1}^i}{R_{t,t+1}} \right]
\end{aligned}$$

which shows that the discounted price plus payoff process is a martingale. The terms on the left are the price processes, the terms on the right are the conditional expectations under the probability measure Q of the payoffs discounted with the risk-free payoff.

The measure Q is a mathematical construct. The important point is that this new probability measure can be computed either from the real probabilities if the state-price deflators are known or directly from the price and payoff processes. This last observation illustrates that the concept of arbitrage depends only on the structure of the price and payoff processes and not on the actual probabilities. As we will see later in this entry, equivalent martingale measures greatly simplify the computation of the pricing of derivatives.

Let's assume that there is short-term risk-free borrowing in the sense that there is a trading strategy able to pay for any given interval (t, s) one sure dollar at time s given that

$(d_t d_{t+1} \dots d_{s-1})^{-1}$ has been invested at time t . Equivalently, we can define for any time interval (t, s) the payoff of a dollar invested risk-free at time t as $R_{t,s} = (d_t d_{t+1} \dots d_{s-1})$.

We now define the concept of *equivalent probability measures*. Given a probability measure P the probability measure Q is said to be equivalent to P if both assign probability zero to the same events. An equivalent probability measure Q is an equivalent martingale measure if all price processes discounted with $R_{t,j}$ become martingales. More precisely, Q is an equivalent martingale measure if and only if the market value of any trading strategy is a martingale:

$$\theta_t \times S_t = E_t^Q \left[\sum_{j=t+1}^T \frac{d_j^\theta}{R_{t,j}} \right]$$

Risk-Neutral Probabilities

Probabilities computed according to the equivalent martingale measure Q are the risk-neutral probabilities. Risk-neutral probabilities can be explicitly computed. Here is how. Call q_ω the risk-neutral probability of state ω . Let's write explicitly the relationship

$$S_t^i = E_t^Q \left[\frac{d_j^i}{R_{t,j}} \right]$$

as follows:

$$\begin{aligned} S_{A_{kt}}^i &= \sum_{\omega \in A_{kt}} \frac{q_\omega}{Q(A_{kt})} \left[\sum_{j=t+1}^T \frac{d_j^i(\omega)}{R_{t,j}} \right] \\ &= \sum_{\omega \in A_{kt}} \frac{q_\omega}{\left(\sum_{\omega \in A_{kt}} q_\omega \right)} \left[\sum_{j=t+1}^T \frac{d_j^i(\omega)}{R_{t,j}} \right] \end{aligned}$$

The above system of equations determines the risk-neutral probabilities. In fact, we can write, for each risky asset, M_t linear equations, where M_t is the number of sets in the partition I_t plus the normalization equation for probabilities. From the above equation, one can see that

the system can be written as

$$\sum_{\omega \in A_{kt}} q_\omega \left[\sum_{j=t+1}^T \frac{d_j^i(\omega)}{R_{t,j}} - S_{A_{kt}}^i \right] = 0$$

$$\sum_{\omega=1}^S q_\omega = 1$$

This system might be determined, indetermined, or impossible. The system will be impossible if there are arbitrage opportunities. This system will be indetermined if there is an insufficient number of securities. In this case, there will be an infinite number of equivalent martingale measures and the market will not be complete.

Now consider the relationship between risk-neutral probabilities and state-price deflators. Consider a probability measure P and a nonnegative random variable Y with expected value on the entire space equal to 1. Define a new probability measure as $Q(B) = E[1_B Y]$ for any event B and where 1_B is the indicator function of the event B . The random variable Y is called the Radon-Nikodym derivative of Q and it is written

$$Y = \frac{dQ}{dP}$$

It is clear from the definition that P and Q are equivalent probability measures as they assign probability zero to the same events. Note that in the case of a finite-state probability space the new probability measure is defined on each state and is equal to

$$q_\omega = Y(\omega) p_\omega$$

Suppose π_t is a state-price deflator. Let Q be the probability measure defined by the Radon-Nikodym derivative:

$$\xi_T = \frac{\pi_T R_{0,T}}{\pi_0}$$

The new state probabilities under Q are the following:

$$q_\omega = \frac{\pi_T(\omega) R_{0,T}}{\pi_0(\omega)} p_\omega$$

Define the density process ξ_t for Q as $\xi_t = E_t[\xi_T]$. As $\xi_t = E_t[\xi_T]$ is an adapted process, we can write:

$$\begin{aligned} (E_t[\xi_T])_{A_{kt}} &= \xi_{A_{kt}} = \sum_{\omega \in A_{kt}} \frac{p_\omega}{P(A_{kt})} \xi_T(\omega) \\ &= \sum_{\omega \in A_{kt}} \frac{p_\omega}{P(A_{kt})} \frac{\pi_T(\omega) R_{0,T}}{\pi_0(\omega)} = \frac{\pi_{A_{kt}} R_{0,t}}{\pi_0(\omega)} \\ &\quad \times \frac{1}{\pi_{A_{kt}}} \sum_{\omega \in A_{kt}} \frac{p_\omega}{P(A_{kt})} \pi_T[\pi_0(\omega)] R_{t,T} \\ &= \frac{\pi_{A_{kt}} R_{0,t}}{\pi_0} \end{aligned}$$

As $R_{t,s} = (d_t d_{t+1} \dots d_{s-1})$ is the payoff at time s of one dollar invested in a risk-free asset at time t , $s > t$, we can then write the following equation:

$$1 = \frac{1}{\pi_t} E_t[\pi_s R_{t,s}]$$

Therefore,

$$\begin{aligned} 1 &= \frac{1}{\pi_{A_{kt}}} \left[\sum_{\omega \in A_{kt}} P(\{\omega\} | A_{kt}) \pi_s(\omega) R_{t,s} \right] \\ &= \frac{1}{\pi_{A_{kt}}} \left[\sum_{\omega \in A_{kt}} \frac{p_\omega}{P(A_{kt})} \pi_s(\omega) R_{t,s} \right] \\ 1 &\leq k \leq M_t \end{aligned}$$

Substituting in the previous equation, we obtain, for each interval (t, T) ,

$$\xi_{A_{kt}} = (E_t[\xi_T])_{A_{kt}} = \frac{\pi_{A_{kt}} R_{0,t}}{\pi_{A_{10}}}$$

which we can rewrite in the usual notation as

$$\xi_t = E_t[\xi_T] = \frac{\pi_t R_{0,t}}{\pi_{10}}$$

We can now state the following result. Consider any \mathcal{F}_j -measurable variable x_j . This condition can be expressed equivalently stating that x_j assumes constant values on each set of the partition I_j . Then the following relationship holds:

$$E_t^Q[x_j] = E_t^P \frac{1}{\xi_t} [\xi_j x_j]$$

To see this, consider the following demonstration, which hinges on the fact that x_j assumes a constant value on each A_{lj} and, therefore, can be taken out of sums. In addition, as demonstrated above, from

$$1 = \frac{1}{\pi_t} E_t[\pi_s R_{t,s}]$$

the following relationship holds:

$$\begin{aligned} P(A_{kt}) \pi_{A_{kt}} &= \sum_{\omega \in A_{kt}} p_\omega \pi_s(\omega) R_{t,s} \\ 1 &\leq k \leq M_t \end{aligned}$$

$$\begin{aligned} (E_t^Q[x_j])_{A_{kt}} &= \sum_{\omega \in A_{kt}} \frac{q_\omega}{Q(A_{kt})} x_j(\omega) = \sum_{\omega \in A_{kt}} \frac{p_\omega}{Q(A_{kt})} \frac{\pi_T(\omega) R_{0,T}}{\pi_0(\omega)} x_j(\omega) \\ &= \frac{1}{Q(A_{kt})} \sum_{A_{hj} \subset A_{kt}} \left[\sum_{\omega \in A_{hj}} \frac{R_{0,j} R_{j,T} p_\omega \pi_T(\omega) x_j(\omega)}{\pi_0(\omega)} \right] \\ &= \frac{1}{Q(A_{kt})} \sum_{A_{hj} \subset A_{kt}} \left[\frac{x_{A_{hj}} R_{0,j}}{\pi_0(\omega)} \sum_{\omega \in A_{hj}} R_{j,T} p_\omega \pi_T(\omega) \right] \\ &= \frac{1}{Q(A_{kt})} \sum_{A_{hj} \subset A_{kt}} \left[\frac{x_{A_{hj}} R_{0,j} \pi_{A_{hj}} P(A_{hj})}{\pi_0(\omega)} \right] \\ &= \frac{1}{Q(A_{kt})} \sum_{A_{hj} \subset A_{kt}} [x_{A_{hj}} \xi_{A_{hj}} P(A_{hj})] \\ &= \frac{1}{\xi_{A_{kt}}} \sum_{A_{hj} \subset A_{kt}} \frac{x_{A_{hj}} \xi_{A_{hj}} P(A_{hj})}{P(A_{kt})} \\ &= \frac{1}{\xi_{A_{kt}}} [E_t^P(\xi_j x_j)_{A_{kt}}] \end{aligned}$$

Let's now apply the above result to the relationship:

$$\begin{aligned} S_t^i &= \frac{1}{\pi_t} E_t \left[\sum_{j=t+1}^T \pi_j d_j^i \right] = \frac{\pi_0}{\pi_t} E_t \left[\sum_{j=t+1}^T \frac{\pi_j R_{t,j}}{\pi_0} \frac{d_j^i}{R_{t,j}} \right] \\ &= \frac{\pi_0}{\pi_t R_{0,j}} E_t \left[\sum_{j=t+1}^T \frac{\pi_j R_{0,j}}{\pi_0} \frac{d_j^i}{R_{t,j}} \right] = E_t^Q \left(\frac{d_j^i}{R_{t,j}} \right) \end{aligned}$$

We have thus demonstrated the following results: There is no arbitrage if and only if there is an equivalent martingale measure. In addition,

π_t is a state-price deflator if and only if an equivalent martingale measure Q has the density process defined by

$$\xi_t = \frac{\pi_t R_{0,t}}{\pi_0}$$

In addition, it can be demonstrated that, if there is no arbitrage, markets are complete if and only if there is a unique equivalent martingale measure.

To illustrate the above we now proceed to detail the calculations for the previous example of three assets, three dates, and four states. Let's first write the equations for the risk-free asset:

$$\begin{aligned} 1 &= \frac{1}{\pi_{A_{kt}}} \left[\sum_{\omega \in A_{kt}} \frac{p_\omega}{P(A_{kt})} \pi_s(\omega) R_{t,s} \right] \\ 1 &= \frac{1}{\pi_{A_{11}}} \left(\frac{p_1}{p_1 + p_2} \pi_2(1) R_{1,2} + \frac{p_2}{p_1 + p_2} \pi_2(2) R_{1,2} \right) \\ 1 &= \frac{1}{\pi_{A_{21}}} \left(\frac{p_3}{p_3 + p_4} \pi_2(3) R_{1,2} + \frac{p_4}{p_3 + p_4} \pi_2(4) R_{1,2} \right) \\ 1 &= \frac{1}{\pi_{A_{10}}} [p_1 \pi_2(1) R_{0,2} + p_2 \pi_2(2) R_{0,2} \\ &\quad + p_3 \pi_2(3) R_{0,2} + p_4 \pi_2(4) R_{0,2}] \end{aligned}$$

$$\pi_{A_{11}} = \pi_1(1) = \pi_1(2)$$

$$\pi_{A_{21}} = \pi_1(3) = \pi_1(4)$$

$$\pi_{A_{10}} = \pi_0(1) = \pi_0(2) = \pi_0(3) = \pi_0(4)$$

We can now rewrite the pricing relationships for the other risky assets as follows:

$$\text{At date 2, prices are zero: } S_2^i = 0.$$

At date 1, the relationship

$$S_1^i = E_1 \left[\frac{d_2^i}{R_{1,2}} \right]$$

holds. In fact, we can write the following:

$$\begin{aligned} S_{A_{1,1}}^i &= S_1^i(1) = S_1^i(2) \\ &= \frac{1}{\pi_1(2)} [P(A_{1,2}|A_{1,1}) \pi_2(1) d_2^i(1) \\ &\quad + P(A_{2,2}|A_{1,1}) \pi_2(2) d_2^i(2)] \\ &= \frac{1}{\pi_{11}} \left(\frac{p_1}{p_1 + p_2} \pi_2(1) R_{1,2} \frac{d_2^i(1)}{R_{1,2}} \right. \\ &\quad \left. + \frac{p_2}{p_1 + p_2} \pi_2(2) R_{1,2} \frac{d_2^i(2)}{R_{1,2}} \right) \end{aligned}$$

$$\begin{aligned} &= \left[Q(A_{1,2}|A_{1,1}) \frac{d_2^i(1)}{R_{1,2}} + Q(A_{2,2}|A_{1,1}) \frac{d_2^i(2)}{R_{1,2}} \right] \\ &= \left[\frac{q_1}{q_1 + q_2} \frac{d_2^i(1)}{R_{1,2}} + \frac{q_2}{q_1 + q_2} \frac{d_2^i(2)}{R_{1,2}} \right] \end{aligned}$$

$$S_{A_{2,1}}^i = S_1^i(3) = S_1^i(4)$$

$$\begin{aligned} &= \left[Q(A_{3,2}|A_{1,1}) \frac{d_2^i(3)}{R_{1,2}} + Q(A_{4,2}|A_{1,1}) \frac{d_2^i(4)}{R_{1,2}} \right] \\ &= \left[\frac{q_3}{q_3 + q_4} \frac{d_2^i(3)}{R_{1,2}} + \frac{q_4}{q_3 + q_4} \frac{d_2^i(4)}{R_{1,2}} \right] \end{aligned}$$

At date 0, the relationship

$$S_0^i = E_0 \left[\frac{d_1^i}{R_{0,1}} + \frac{d_2^i}{R_{0,2}} \right]$$

holds. In fact we can write the following:

$$\begin{aligned} S_{A_{1,0}}^i &= S_0^i(1) = S_0^i(2) = S_0^i(3) = S_0^i(4) \\ &= \frac{1}{\pi_{A_{10}}} \left\{ \begin{aligned} &p_1 [\pi_1(1) d_1^i(1) + \pi_2(1) d_2^i(1)] \\ &+ p_2 [\pi_1(2) d_1^i(2) + \pi_2(2) d_2^i(2)] \\ &+ p_3 [\pi_1(3) d_1^i(3) + \pi_2(3) d_2^i(3)] \\ &+ p_4 [\pi_1(4) d_1^i(4) + \pi_2(4) d_2^i(4)] \end{aligned} \right\} \\ &= p_1 \left[\frac{\pi_1(1) R_{0,1}}{\pi_{A_{1,0}}} \frac{d_1^i(1)}{R_{0,1}} + \frac{\pi_2(1) R_{0,2}}{\pi_{A_{1,0}}} \frac{d_2^i(1)}{R_{0,2}} \right] \\ &\quad + p_2 \left[\frac{\pi_1(2) R_{0,1}}{\pi_{A_{1,0}}} \frac{d_1^i(2)}{R_{0,1}} + \frac{\pi_2(2) R_{0,2}}{\pi_{A_{1,0}}} \frac{d_2^i(2)}{R_{0,2}} \right] \\ &\quad + p_3 \left[\frac{\pi_1(3) R_{0,1}}{\pi_{A_{1,0}}} \frac{d_1^i(3)}{R_{0,1}} + \frac{\pi_2(3) R_{0,2}}{\pi_{A_{1,0}}} \frac{d_2^i(3)}{R_{0,2}} \right] \\ &\quad + p_4 \left[\frac{\pi_1(4) R_{0,1}}{\pi_{A_{1,0}}} \frac{d_1^i(4)}{R_{0,1}} + \frac{\pi_2(4) R_{0,2}}{\pi_{A_{1,0}}} \frac{d_2^i(4)}{R_{0,2}} \right] \\ &= p_1 \left\{ \frac{\pi_1(1) R_{0,1}}{\pi_{A_{1,0}}} \frac{d_1^i(1)}{R_{0,1}} \frac{1}{\pi_{11}} \left[\frac{p_1}{p_1 + p_2} \pi_2(1) R_{1,2} \right. \right. \\ &\quad \left. \left. + \frac{p_2}{p_1 + p_2} \pi_2(2) R_{1,2} \right] \right\} \\ &\quad + p_2 \left\{ \frac{\pi_1(2) R_{0,1}}{\pi_{A_{1,0}}} \frac{d_1^i(2)}{R_{0,1}} \frac{1}{\pi_{21}} \left[\frac{p_1}{p_1 + p_2} \pi_2(1) R_{1,2} \right. \right. \\ &\quad \left. \left. + \frac{p_2}{p_1 + p_2} \pi_2(2) R_{1,2} \right] \right\} \\ &\quad + p_3 \left\{ \frac{\pi_1(3) R_{0,1}}{\pi_{A_{1,0}}} \frac{d_1^i(3)}{R_{0,1}} \frac{1}{\pi_{31}} \left[\frac{p_3}{p_3 + p_4} \pi_2(3) R_{1,2} \right. \right. \\ &\quad \left. \left. + \frac{p_4}{p_3 + p_4} \pi_2(4) R_{1,2} \right] \right\} \end{aligned}$$

$$\begin{aligned}
& + p_4 \left\{ \frac{\pi_1(4)R_{0,1}}{\pi_{A_{1,0}}} \frac{d_1^i(4)}{R_{0,1}} \frac{1}{\pi_{41}} \left[\frac{p_3}{p_3 + p_4} \pi_2(3)R_{1,2} \right. \right. \\
& \left. \left. + \frac{p_3}{p_3 + p_4} \pi_2(4)R_{1,2} \right] \right\} \\
& + q_1 \frac{d_2^i(1)}{R_{0,2}} + q_2 \frac{d_2^i(2)}{R_{0,2}} + q_3 \frac{d_2^i(3)}{R_{0,2}} + q_4 \frac{d_2^i(4)}{R_{0,2}} \\
& = \frac{d_1^i(1)}{R_{0,1}} \left[\frac{p_1 \pi_2(1)}{\pi_{A_{1,0}}} R_{0,2} + \frac{p_2 \pi_2(2)}{\pi_{A_{1,0}}} R_{0,2} \right] \\
& + \frac{d_1^i(3)}{R_{0,1}} \left[\frac{p_3 \pi_2(3)}{\pi_{A_{1,0}}} R_{0,2} + \frac{p_4 \pi_2(4)}{\pi_{A_{1,0}}} R_{0,2} \right] \\
& + q_1 \frac{d_2^i(1)}{R_{0,2}} + q_2 \frac{d_2^i(2)}{R_{0,2}} + q_3 \frac{d_2^i(3)}{R_{0,2}} + q_4 \frac{d_2^i(4)}{R_{0,2}} \\
& = q_1 \frac{d_1^i(1)}{R_{0,1}} + q_2 \frac{d_1^i(2)}{R_{0,1}} + q_3 \frac{d_1^i(3)}{R_{0,1}} + q_4 \frac{d_1^i(4)}{R_{0,1}} \\
& + q_1 \frac{d_2^i(1)}{R_{0,2}} + q_2 \frac{d_2^i(2)}{R_{0,2}} + q_3 \frac{d_2^i(3)}{R_{0,2}} + q_4 \frac{d_2^i(4)}{R_{0,2}}
\end{aligned}$$

The value of a derivative instrument might depend on the path of its past values. Consider a lookback option on a stock—that is, a derivative instrument on a stock whose payoff at time t is the maximum difference between the price of the stock and a given value K at any moment prior to t . Call V_t the payoff of the lookback option at time t . We can then write:

$$\begin{aligned}
V_t &= \max_{0 \leq k < t} (S_k - K)^+ \\
(S_k - K)^+ S_k - K & (S_k - K)^+ = \max(S_k - K, 0)
\end{aligned}$$

THE BINOMIAL MODEL

Let's now introduce the simple but important multiperiod finite-state model known as the binomial model. The binomial model is important because it gives a simple and mathematically tractable model of stock price behavior that tends, in the limit of a zero time step, to a Brownian motion.¹ We introduce a market populated by one risk-free asset and by one or more risky assets whose price(s) follow(s) a binomial or trinomial model. In the next section we will see how to compute the price of derivative instruments in this market.

In the binomial model of stock prices, we assume that at each time step the stock price will assume one of two possible values. This is a restriction of the general multiperiod finite-state model described in the previous sections on probability theory. The latter is, as we have seen in the previous section, a hierarchical structure of partitions of the set of states. The number of sets in any partition is arbitrary, provided that partitions grow more refined with time.

The binomial model assumes that there are two positive numbers, d and u , such that $0 < d < u$ and such that at each time step the price S_t of the risky asset changes to dS_t or to uS_t . In general one assumes that $0 < d < 1 < u$ so that d represents a price decrease (a movement down) while u represents a price increase (a movement up). It is often required that

$$d = \frac{1}{u}$$

In this case an equal number of movements up and down leave prices unchanged. The binomial model is a Markov model as the distribution of S_t clearly depends only on the value of S_{t-1} .

A binomial model can be graphically represented by a tree. For example, Figure 2 shows a binomial model for three periods. A binomial model over T time steps, from 0 to T , produces a total of 2^T paths. Therefore, the corresponding space of states has 2^T states. However, the

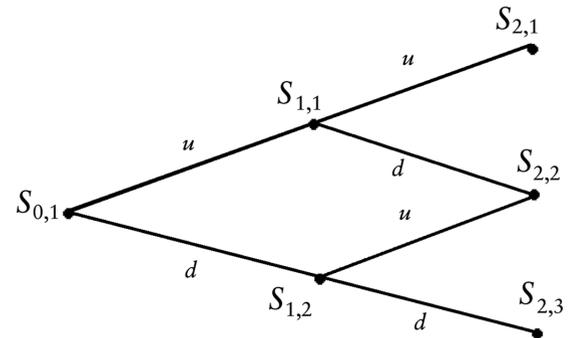


Figure 2 Binomial Model: Illustration of a Binomial Tree with Three Dates, Three Final Prices, and Four States: uu , ud , du , dd

number of different final prices $S_T = u^k d^{T-k} S_0$, $k = 0, 1, \dots, T$ is determined solely by the number of u and d in each path and increases by 1 at each time step; there are as many final prices as dates. For example, the model in Figure 2 shows three final prices and four states.

Note that there is a simple relationship between the numbers d and u and returns. In fact, we can write,

$$R_t(\text{up}) = \frac{S_{t+1} - S_t}{S_t} = \frac{uS_t - S_t}{S_t} = u - 1$$

$$R_t(\text{down}) = d - 1$$

Real probabilities of states are typically constructed from the probabilities of a movement up or down. Call p the probability of a movement up; $1 - p$ is thus the probability of a movement down. Suppose that the state s , which is identified by a price path, has k movements up and $T - k$ movements down. The probability of the state s is

$$p_s = p^k (1 - p)^{T-k}$$

Consider the final date T . Each of the possible final prices $S_T = u^k d^{T-k} S_0$, $k = 0, 1, \dots, T$ can be obtained through

$$\binom{T}{k} = \frac{T!}{k!(T-k)!}$$

paths with k movements up and $T - k$ movements down. The probability distribution of final prices is therefore a binomial distribution:

$$P(S_T = u^k d^{T-k} S_0) = \binom{T}{k} p^k (1 - p)^{T-k}$$

Following the same reasoning, one can demonstrate that at any intermediate date the probability distribution of prices is a binomial distribution as follows:

$$P(S_t = u^k d^{t-k} S_0) = \binom{t}{k} p^k (1 - p)^{t-k}$$

Next introduce a risk-free security. In the setting of a binomial model, a risk-free security is simply a security such that $d = u = 1 + r$ where

$r > 0$ is the positive risk-free rate. To avoid arbitrage it is clearly necessary that $d < 1 + r < u$. In fact, if the interest rate is inferior to both the up and down returns, one can make a sure profit by buying the risky asset and shorting the risk-free asset. If the interest rate is superior to both the up and down returns, one can make a sure profit by shorting the risky asset and buying the risk-free asset. Denote by b_t the price of the risk-free asset at time t . From the definition of price movement in the binomial model we can write: $b_t = (1 + r)^t b_0$.

Risk-Neutral Probabilities for the Binomial Model

Let's now compute the risk-neutral probabilities. In the setting of binomial models, the computation of risk-neutral probabilities is simple. In fact we have to impose the condition:

$$q_t = E_t^Q[q_{t+1}]$$

which we can explicitly write as follows:

$$S_t = \frac{quS_t + (1 - q)dS_t}{1 + r}$$

$$1 + r = qu + d - qd$$

$$q = \frac{1 + r - d}{u - d}$$

$$1 - q = \frac{u - 1 - r}{u - d}$$

As we have assumed $0 < d < 1 + r < u$, the condition $0 < q < 1$ holds. Therefore we can state that the unique risk-neutral probabilities are

$$q = \frac{1 + r - d}{u - d}$$

$$1 - q = \frac{u - 1 - r}{u - d}$$

The binomial model is complete and arbitrage free.

Suppose that there is more than one risky asset, for example two risky assets, in addition to the risk-free asset. At each time step each

risky asset can go either up or down. Therefore there are four possible joint movements at each time step: uu, ud, du, dd that we identify with the states 1,2,3,4. Four probabilities must be determined at each time step; four equations are therefore needed. Two equations are provided by the martingale conditions:

$$S_t^1 = \frac{q_1 u S_t^1 + q_2 u S_t^1 + q_3 u S_t^1 + q_4 u S_t^1}{1+r}$$

$$S_t^2 = \frac{q_1 u S_t^2 + q_3 u S_t^2 + q_2 u S_t^2 + q_4 u S_t^2}{1+r}$$

A third equation is provided by the fact that probabilities must sum to 1. The fourth condition, however, is missing. The model is incomplete.

The problem of approximating price processes when there are two stocks and one bond and where the stock prices follow two correlated lognormal processes has long been of interest to financial economists. As seen above, with two stocks and one bond available for trading, markets cannot be completed by dynamic trading. This is not the case in the continuous-time model, in which markets can be completed by continuous trading in the two stocks and the bond. Different solutions to this problem have been proposed in the literature.²

ARBITRAGE PRICING IN A DISCRETE-TIME, CONTINUOUS-STATE SETTING

Let's now discuss the discrete-time, continuous-state setting. This is an important setting as it is, for example, the setting of the arbitrage pricing theory (APT) model.³

As in the previous discrete-time, discrete-state setting, we apply probabilistic concepts. The economy is represented by a probability space (Ω, σ, P) where Ω is the set of possible states, σ is the σ -algebra of events (formed, in

this continuous-state setting, by a nondenumerable number of events), and P is a probability function. As the number of states is infinite, the probability of each state is zero and only events, in general, formed by nondenumerable states have a finite probability. There are only a finite number of dates from 0 to T . The propagation of information is represented by a finite filtration \mathfrak{S}_t , $t = 0, 1, \dots, T$. In this case, the filtration \mathfrak{S}_t is not equivalent to an information structure I_t .

Each security i is characterized by a payoff process d_t^i and by a price process S_t^i . In this continuous-state setting, d_t^i and S_t^i are formed by a finite number of continuous variables. As before, $d_t^i(\omega)$ and $S_t^i(\omega)$ are, respectively, the payoff and the price of the i -th asset at time t , $0 \leq t \leq T$ and in state $\omega \in \Omega$. All payoffs and prices are stochastic processes adapted to the filtration \mathfrak{S} .

To develop an intuition for continuous-state arbitrage pricing, consider the previous multiperiod, finite-state case with a very large number M of states, $M \gg N$ where N is the number of securities. Recall from our earlier discussion that risk-neutral probabilities can be computed solving the following system of linear equations:

$$\sum_{\omega \in A_{k,t}} q_\omega \left[\sum_{j=t+1}^T \frac{d_j^i(\omega)}{R_{t,j}} - S_{A_{k,t}}^i \right] = 0$$

$$\sum_{\omega=1}^M q_\omega = 1$$

Recall also that at each date t the information structure I_t partitions the set of states into M_t subsets. Each partition therefore yields $N \times M_t$ equations and the system is formed by a total of

$$N \times \sum_{t=0}^{T-1} M_t$$

equation plus the probability normalizing equation. Consider that the previous system can be broken down, at each date t , into separate blocks formed by N equations (one for

each asset) of the following type:

$$\sum_{\omega \in A_{kt}} q_{\omega}^* \sum_{j=t+1}^T \frac{d_j^i}{R_{t,j}} = S_{A_{kt}}$$

$$q_{\omega}^* = \frac{q_{\omega}}{\sum_{\omega \in A_{kt}} q_{\omega}}$$

Each of these systems can be solved individually for the conditional probabilities q_{ω}^* . Recall that a system of this type admits a solution if and only if the coefficient matrix and the augmented coefficient matrix have the same rank. If the system is solvable, its solution will be unique if and only if the number of unknowns is equal to the rank of the coefficient matrix.

If the above system is not solvable, then there are arbitrage opportunities. This occurs if the payoffs of an asset are a linear combination of those of other assets, but its price is not the same linear combination of the prices of the other assets. This happens, in particular, if two assets have the same payoff in each state but different prices. In these cases, in fact, the rank of the coefficient matrix is inferior to the rank of the augmented matrix.

Under the assumption

$$M \gg N \times \sum_{t=0}^{T-1} M_t$$

this system, if it is solvable, will be undetermined. Therefore, there will be infinite equivalent risk-neutral probabilities and the market will not be complete. Going to the limit of an infinite number of states, the above reasoning proves, heuristically, that a discrete-time continuous-state market with a finite number of securities is inherently incomplete. In addition, there will be arbitrage opportunities only if the random variable that represents the payoff of an asset is a linear combination of the random variables that represent the payoffs of other assets, but the random variables that represent prices are not in the same relationship.

The above discussion can be illustrated in the case of multiple assets, each following a binomial model. If there are N linearly indepen-

dent assets, the price paths in the interval $(0, T)$ will form a total of 2^{NT} states. In a binomial model, we can limit our considerations to one time step as the other steps are identical. In one step, each price S_t^i at time t can go up to $S_t^i u^i$ or down to $S_t^i d^i$ at time $t + 1$. Given the prices $\{S_t^i\} \equiv \{S_t^1, S_t^2, \dots, S_t^N\}$ at time t , there will be at the next time step, $2N$ possible combinations $\{S_t^1 w^1, S_t^2 w^2, \dots, S_t^N w^N\}$, $w^i = u^i$ or d^i .

Suppose that there are 2^N states and that each combination of prices identifies a state. This means that at each date t the information structure I_t partitions the set of states into 2^{Nt} subsets. Each set of the partition is partitioned into 2^N subsets at the next time step. This yields $2^N(t + 1)$ subsets at time $t + 1$.

Note that this partitioning is compatible with any correlation structure between the random variables that represent prices. In fact, correlations depend on the value of the probability assigned to each state while the partitioning we assume depends on how different prices are assigned to different states.

Risk-neutral probabilities q_i , $i = 1, 2, \dots, 2^N$ can be determined solving the following system of martingale conditions:

$$\sum_{j=1}^{2^N} q_j S_t^i w^i(j) = S_t^i$$

$$\sum_{j=1}^{2^N} q_j = 1$$

$$j = 1, 2, \dots, 2^N, i = 1, 2, \dots, N$$

which becomes, after dividing each equation by S_t^i , the following:

$$\sum_{i=1}^{2^N} q^1 w_j(j) = 1$$

$$\sum_{j=1}^{2^N} q_j = 1$$

where $w^i(j) = u^i$ or d^i for asset i in state j .

It can be verified that, under the previous assumptions and provided prices are positive, the above system admits infinite solutions. In fact,

as $N + 1 < 2^N$, the number of equations is larger than the number of unknowns. Therefore, if the system is solvable it admits infinite solutions. To verify that the system is indeed solvable, let's choose the first asset and partition the set of states into two events corresponding to the movement up or down of the same asset. Assign to these events probabilities as in the binomial model

$$q_t^1 = \frac{1 - r + d_t^1}{u_t^1 - d_t^1} \text{ and } 1 - q_t^1$$

Choose a second asset and partition each of the previous events into two events corresponding to the movements up or down of the second asset. We can now assign the following probabilities to each of the following four events:

$$q_t^1 q_t^2, q_t^1(1 - q_t^2), (1 - q_t^1)q_t^2, (1 - q_t^1)(1 - q_t^2)$$

It can be verified that these numbers sum to one. The same process can be repeated for each additional asset. We obtain a set of positive numbers that sum to one and that satisfy the system by construction. There are infinite other possible constructions. In fact, at each step, we could multiply probabilities by "correlation factors" (i.e., numbers that form a 2×2 correlation matrix) and still obtain solutions to the system.

We can therefore conclude that a system of positive binomial prices such as the one above plus a risk-free asset is arbitrage-free and forms an incomplete market. If we let the number of states tend to infinity, the binomial distribution converges to a normal distribution. We have therefore demonstrated heuristically that a multivariate normal distribution plus a risk-free asset forms an incomplete and arbitrage-free market. Note that the presence of correlations does not change this conclusion.

Let's now see under what conditions this conclusion can be changed. Go back to the multiple binomial model, assuming, as before, that there are N assets and T time steps. There is no logical reason to impose that the number of states be 2^{NT} . As we can consider each time step separately, suppose that there is only one time step

and that there are a number of states less than or equal to the number of assets plus 1: $M \leq N + 1$. In this case, the martingale condition that determines risk-neutral probabilities becomes:

$$\sum_{j=1}^M q_j w^i(j) \\ \sum_{j=1}^N q_j = 1$$

There are M equations and $N + 1$ unknowns with $M \leq N + 1$. This system will either determine unique risk-neutral probabilities or will be unsolvable. Therefore, the market will be either complete and arbitrage-free or will exhibit arbitrage opportunities. Note that in this case we cannot use the constructive procedure used in the previous case.

What is the economic meaning of the condition that the number of states be less than or equal to the number of assets? To illustrate this point, assume that the number of states is $M = 2^K \leq N + 1$. This means that we can choose K assets whose independent price processes identify all the states as in the previous case. Now add one more asset. This asset will go up or down not in specific states but in events formed by a number of states. Suppose it goes up in the event A and goes down in the event B . These events are determined by the value of the first K assets. In other words, the new asset will be a function of the first K assets. An interesting case is when the new asset can be expressed as a linear function of the first K assets. We can then say that the first K assets are factors and that any other asset is expressed as a linear combination of the factors.

Consider that, given the first K assets, it is possible to determine state-price deflators. These state-price deflators will not be uniquely determined. Any other price process must be expressed as a linear combination of state-price deflators to avoid arbitrage. If all price processes are arbitrage-free, the market will be

complete if it is possible to determine uniquely the risk-neutral probabilities.

If we let the number of states become very large, the number of assets must become large as well. Therefore it is not easy to develop simple heuristic arguments in the limit of a large economy. What we can say is that in a large discrete economy where the number of states is less than or equal to the number of assets, if there are no arbitrage opportunities the market might be complete. If the market is complete and arbitrage-free, there will be a number of factors while all other processes will be linear combinations of these factors.

KEY POINTS

- The law of one price states that a given asset must have the same price regardless of the means by which one goes about creating that asset.
- Arbitrage is the simultaneous buying and selling of an asset at two different prices in two different markets.
- A finite-state one-period market is represented by a vector of prices and a matrix of payoffs.
- A state-price vector is a strictly positive vector such that prices are the product of the state-price vector and the payoff matrix.
- There is no arbitrage if and only if there is a state-price vector.
- A market is complete if an arbitrary payoff can be replicated by a portfolio.
- A finite-state one-period market is complete if there are as many linearly independent assets as states.
- A multiperiod finite-state economy is represented by a probability space plus an information structure.
- In a multiperiod finite-state market each security is represented by a payoff process and a price process.
- An arbitrage is a trading strategy whose payoff process is nonnegative and not always zero.
- A market is complete if any nonnegative payoff process can be replicated with a trading strategy.
- A state-price deflator is a strictly positive process such that prices are random variables equal to the conditional expectation of discounted payoffs.
- A martingale is a process such that at any time t its conditional expectation at time s , $s > t$ coincides with its present value.
- In the absence of arbitrage there is an artificial probability measure in which all price processes, appropriately discounted, become martingales.
- Given a probability measure P , the probability measure Q is said to be equivalent to P if both assign probability zero to the same events.
- The binomial model assumes that there are two positive numbers, d and u , such that $0 < d < u$ and such that at each time step the price S of the risky asset changes to dS or to uS .
- The distribution of prices of a binomial model is a binomial distribution.
- The binomial model is complete.

NOTES

1. The binomial model was first suggested for the pricing of options by Cox, Ross, and Rubinstein (1979), Rendleman and Bartter (1979), and Sharpe (1978).
2. See He (1990).
3. For an application of the principles discussed here to the APT, see Focardi and Fabozzi (2004).

REFERENCES

- Cox, J. C., Ross, S. A., and Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of Financial Economics* 7:229–263.

- Focardi, S. M., and Fabozzi, F. J. (2004). *The Mathematics of Financial Modeling and Investment Management*. Hoboken, NJ: John Wiley & Sons.
- He, H. (1990). Convergence from discrete- to continuous-time contingent claims prices. *Review of Financial Studies* 3:523–546.
- Rendleman, R. J., Jr., and Bartter, B. J. (1979). Two-state option pricing. *Journal of Finance* 24:1093–1110.
- Sharpe, W. F. (1978). *Investments*. Englewood Cliffs, NJ: Prentice Hall.
- Shleifer, A., and Vishny, R. W. (1997). The limits of arbitrage. *Journal of Finance* 52:35–55.

Arbitrage Pricing: Continuous-State, Continuous-Time Models

SERGIO M. FOCARDI, PhD
Partner, The Intertek Group

FRANK J. FABOZZI, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

Abstract: The principle of absence of arbitrage is perhaps the most fundamental principle of finance theory. In the presence of arbitrage opportunities, there is no trade-off between risk and returns because it is possible to make unbounded risk-free gains. The principle of absence of arbitrage is fundamental for understanding asset valuation in a competitive market. Arbitrage pricing can be developed in a finite-state, discrete-time setting and a continuous-time, continuous-state setting.

In this entry, we describe *arbitrage pricing* in the *continuous-state, continuous-time setting*. There are a number of important conceptual changes in going from a discrete-state, discrete-time setting (as described in the entry “Arbitrage Pricing: Finite-State Models”) to a continuous-state, continuous-time setting. First, each state of the world has probability zero. This precludes the use of standard conditional probabilities for the definition of conditional expectation and requires the use of filtrations (rather than of information structures) to describe the propagation of information. Second, the tools of matrix algebra are inadequate; the more complex tools of calculus and stochastic calculus are required. Third, simple generalizations are rarely possible as many pathological cases appear in connection with infinite sets.

THE ARBITRAGE PRINCIPLE IN CONTINUOUS TIME

Let’s start with the definition of basic concepts. The economy is represented by a probability space $(\Omega, \mathfrak{F}, P)$ where Ω is the set of possible states, \mathfrak{F} is the σ -algebra of events, and P is a probability measure. Time is a continuous variable in the interval $[0, T]$. The propagation of information is represented by a filtration \mathfrak{F}_t . The latter is a family of σ -algebras such that $\mathfrak{F}_t \subseteq \mathfrak{F}_s, t < s$.

Each security i is characterized by a payoff-rate process δ_t^i and by a price process S_t^i . In this continuous-state setting, δ_t^i and S_t^i are real variables with a continuous range such that $\delta_t^i(\omega)$ and $S_t^i(\omega)$ are, respectively, the payoff-rate and the price of the i -th asset at time $t, 0 \leq t \leq T$

and in state $\omega \in \Omega$. Note that δ_t^i represents a rate of payoff and not a payoff as was the case in the discrete-time setting. The payoff-rate process must be interpreted in the sense that the cumulative payoff of each individual asset is

$$D_t^i = \int_0^t \delta_s^i ds$$

We assume that the number of assets is finite. We can therefore use the vector notation to indicate a set of processes. For example, we write δ_t and S_t to indicate the vector process of payoff rates and prices respectively. All payoff-rates and prices are stochastic processes adapted to the filtration \mathfrak{F} . One can make assumptions about the price and the payoff-rate processes. For example, it can be assumed that price and payoff-rate processes satisfy a set of stochastic differential equations or that they exhibit finite jumps. Later in this entry we will explore a number of these processes.

Conditional expectations are defined as partial averaging. In fact, given a variable X_s , $s > t$, its conditional expectation $E_t[X_s]$ is defined as a variable that is \mathfrak{F}_t -measurable and whose average on each set $A \in \mathfrak{F}_t$ is the same as that of X :

$$Y_t = E_t[X_s] \Leftrightarrow E[Y_t(\omega)] = E[X_s(\omega)]$$

for $\omega \in A$, $\forall A \in \mathfrak{F}_t$ and Y is \mathfrak{F}_t -measurable.

The law of iterated expectations applies as in the finite-state case:

$$E_t[E_u(X_s)] = E_t[X_s]$$

In a continuous-state setting, conditional expectations are variables that assume constant values on the sets of infinite partitions. Imagine the evolution of a variable X . At the initial date, X_0 identifies the entire space Ω . At each subsequent date t , the space Ω is partitioned into an infinite number of sets, each determined by one of the infinite values of X_t .¹ However, these sets have measure zero. In fact, they are sets of the type: $\{A: \omega \in A \Leftrightarrow X_t(\omega) = x\}$ determined by specific values of the variable X_t . These sets

have probability zero as there is an infinite number of values X_t . As a consequence, we cannot define conditional expectation as expectation under the usual definition of conditional probabilities the same way we did in the case of finite-state setting.

Trading Strategies and Trading Gains

We have to define the meaning of trading strategies in the continuous-state, continuous-time setting; this requires the notion of continuous trading. Mathematically, continuous trading means that the composition of portfolios changes continuously at every instant and that these changes are associated with trading gains or losses. A trading strategy is a (vector-valued) process $\theta = \{\theta^i\}$ such that $\theta_t = \{\theta_t^i\}$ is the portfolio held at time t . To ensure that there is no anticipation of information, each trading strategy θ must be an adapted process.

Given a trading strategy, we have to define the gains or losses associated with it. In discrete time, the trading gains equal the sum of payoffs plus the change of a portfolio's value

$$\sum_{t=0}^T \left(\sum_t d_t^i \theta_t^i \right) + \sum_i S_T^i \theta_T^i - \sum_i S_0^i \theta_0^i$$

over a finite interval $[0, T]$.

We must define trading gains when time is a continuous variable. It is not possible to replace finite sums of stochastic increments with pathwise Riemann-Stieltjes integrals after letting the time interval go to zero. The reason is that, though we can assume that paths are continuous, we cannot assume that they have bounded variation. As a consequence, pathwise Riemann-Stieltjes integrals generally do not exist. However, we can assume that paths are of bounded quadratic variation. Under this latter assumption, using Itô isometry, we can define pathwise Itô integrals and stochastic integrals.

Let's first assume that the payoff-rate process is zero, so that there are only price processes. Under this assumption, the trading gain

T_t of a trading strategy can be represented by a stochastic integral:

$$T_t = \int_0^t \theta_s dS_s = \sum_i \int_0^t \theta_s^i dS_s^i$$

In the rest of this section, we will not strictly adhere to the vector notation when there is no risk of confusion. For example, we will write $\theta \cdot S$ to represent the scalar product $\theta \cdot S$. If a payoff-rate process is associated with each asset, we have to add the gains consequent to the payoff-rate process. We therefore define the gain process

$$G_t^i = S_t^i + D_t^i$$

as the sum of the price processes plus the cumulative payoff-rate processes, and we define the trading gains as the stochastic integral

$$T_t = \int_0^t \theta_s dG_s = \sum_i \int_0^t \theta_s^i dG_s^i$$

How can we match the abstract notion of a stochastic integral with the buying and selling of assets? In discrete time, trading gains have a meaning that is in agreement with the practical notion of buying a portfolio of assets, holding it for a period, and then selling it at market prices, thus realizing either a gain or a loss. One might object that in continuous time this meaning is lost. How can a process where prices change so that their total variation is unbounded be a reasonable representation of financial reality? This is a question of methodology that is relevant to every field of science. In classical physics, the use of continuous models was assumed to reflect reality; time and space, for example, were considered continuous. Quantum physics upset the conceptual cart of classical physics, and the reality of continuous processes has since been questioned at every level. In quantum physics, a theory is considered to be nothing but a model useful as a mathematical device to predict measurements. This is, in essence, the theory set forth in the 1930s by Niels Bohr and the school

of Copenhagen; it has now become mainstream methodology in physics. It is also, ultimately, the point of view of positive economics. In a famous and widely quoted essay, Milton Friedman (1953) wrote:

The relevant question to ask about the "assumptions" of a theory is not whether they are descriptively "realistic," for they never are, but whether they are sufficiently good approximations for the purpose in hand. And this question can be answered only by seeing whether the theory works, which means if it yields sufficiently accurate predictions.

In the spirit of positive economics, continuous-time financial models are mathematical devices used to predict, albeit in a probabilistic sense, financial observations made at discrete intervals of time. Stochastic gains predict trading gains only at discrete intervals of time—the only intervals that can be observed. Continuous-time finance should be seen as a logical construction that meets observations only at a finite number of dates, not as a realistic description of financial trading.

Let's consider processes without any intermediate payoff. A self-financing trading strategy is a trading strategy such that the following relationships hold:

$$\theta_t S_t = \sum_i \theta_t^i S_t^i = \sum_i \left(\theta_0^i S_0^i + \int_0^t \theta_t^i dS_t^i \right), t \in [0, T]$$

We first define arbitrage in the absence of a payoff-rate process. An arbitrage is a self-financing trading strategy such that: $\theta_0 S_0 < 0$ and $\theta_T S_T \geq 0$, or $\theta_0 S_0 \leq 0$ and $\theta_T S_T > 0$. If there is a payoff-rate process, a self-financing trading strategy is a trading strategy such that the following relationships hold:

$$\theta_t S_t = \sum_i \theta_t^i S_t^i = \sum_i \left(\theta_0^i S_0^i + \int_0^t \theta_t^i dG_t^i \right), t \in [0, T]$$

where $G_t^i = S_t^i + D_t^i$ is the gain process as previously defined. An arbitrage is a self-financing trading strategy such that: $\theta_0 S_0 < 0$ and $\theta_T S_T \geq 0$, or $\theta_0 S_0 \leq 0$ and $\theta_T S_T > 0$.

ARBITRAGE PRICING IN CONTINUOUS-STATE, CONTINUOUS-TIME

The abstract principles of arbitrage pricing are the same in a discrete-state, discrete-time setting as in a continuous-state, continuous-time setting. Arbitrage pricing is relative pricing. In the absence of arbitrage, the price and payoff-rate processes of a set of basic assets fix the prices of other assets given the payoff-rate process of the latter. If markets are complete, every price process can be computed in this way. In a discrete-state, discrete-time setting, the computation of arbitrage pricing is done with matrix algebra. In fact, in the absence of arbitrage, every price process can be expressed in two alternative ways:

1. Prices S_t^i are equal to the normalized conditional expectation of payoffs deflated with state prices under the real probabilities:

$$S_t^i = \frac{1}{\pi_t} E_t \left[\sum_{j=t+1}^T \pi_j d_j^i \right]$$

2. Prices S_t^i are equal to the conditional expectation of discounted payoffs under the risk-neutral probabilities

$$S_t^i = E_t^Q \left[\sum_{j=t+1}^T \frac{d_j^i}{R_{t,j}} \right]$$

State-price deflators and risk-neutral probabilities can be computed solving systems of linear equations for a kernel of basic assets. The above relationships are algebraic linear equations that fix all price processes.

In a continuous-state, continuous-time setting, the principle of arbitrage pricing is the same. In the absence of arbitrage, given a number of basic price and payoff stochastic processes, other processes are fixed. The latter are called *redundant securities* as they are not necessary to fix prices. If markets are complete, every price process can be fixed in this way. In

order to make computations feasible, some additional assumptions are made, in particular, all payoff-rate and price processes are assumed to be Itô processes.

The theory of arbitrage pricing in a continuous-state, continuous-time setting uses the same tools as in a discrete-state, discrete-time setting. Under an equivalent martingale measure, all price processes become martingales. Therefore prices can be determined as discounted present value relationships. *Equivalent martingale measures* are the same concept as state-price deflators: After appropriate deflation, all processes become martingales. The key point of arbitrage pricing theory is that both equivalent martingale measures and state-price deflators can be determined from a subset of the market. All other processes are redundant.

In the following sections we will develop the theory of arbitrage pricing in steps. First, we will illustrate the principles of arbitrage pricing in the case of options, arriving at the *Black-Scholes option pricing formula*. We will then extend this theory to more general derivative securities. Subsequently, we will state arbitrage pricing theory in the context of equivalent martingale measures and of state-price deflators.

OPTION PRICING

We will now apply the concepts of arbitrage pricing to option pricing in a continuous-state, continuous-time setting. Suppose that a market consists of three assets: a risk-free asset (which allows risk-free borrowing and lending at the risk-free rate of interest), a stock, and a European option. We will show that the price processes of a stock and of a risk-free asset fix the price process of an option on that stock.

Suppose the risk-free rate is a constant r . The value V_t of a risk-free asset with constant rate r evolves according to the deterministic differential equation of continually compounding interest rates:

$$dV_t = rV_t dt$$

The above is a differential equation with separable variables. After separating the variables, the equation can be written as

$$\frac{dV_t}{V_t} = r dt$$

which admits the solution $V_t = V_0 e^{rt}$ where V_0 is the initial value of the bank account. This formula can also be interpreted as the price process of a risk-free bond with deterministic rate r .

Stock Price Processes

Let's now examine the price process of the stock. Consider the process $y = \alpha t + \sigma B_t$ where B_t is a standard Brownian motion. From the definition of Itô integrals, it can be seen that this process, which is called an arithmetic Brownian motion, is the solution of the following diffusion equation:

$$dy_t = \alpha dt + \sigma dB_t$$

where α is a constant called the drift of the diffusion and σ is a constant called the *volatility of the diffusion*.

Consider now the process $S_t = S_0 e^{(\alpha t + \sigma B_t)}$, $t \geq 0$. Applying Itô's lemma it is easy to see that this process, which is called a geometric Brownian motion, is an Itô process that satisfies the following stochastic differential equation:

$$dS_t = \mu S_t dt + \sigma S_t dB_t; S_0 = x$$

where x is an initial value, $\mu = \alpha + 1/2\sigma^2$ and B_t is a standard Brownian motion. We assume that the stock price process follows a geometric Brownian motion and that there is no payoff-rate process.

Now consider a European call option, which gives the owner the right but not the obligation to buy the underlying stock at the exercise price K at the expiry date T . Call Y_t the price of the option at time t . The price of the option as a function of the stock price is known at the final

expiry date. If the option is rationally exercised, the final value of the option is

$$Y_T = \max(S_T - K, 0)$$

In fact, the option can be rationally exercised only if the price of the stock exceeds K . In that case, the owner of the option can buy the underlying stock at the price K , sell it immediately at the current price S_t and make a profit equal to $(S_t - K)$. If the stock price is below K , the option is clearly worthless. After T , the option ceases to exist.

How can we compute the option price at every other date? We can arrive at the solution in two different but equivalent ways: (1) through hedging arguments and (2) the equivalent martingale measures. In the following sections we will introduce hedging arguments and equivalent martingale measures.

Hedging

To hedge means to protect against an adverse movement. The seller of an option is subject to a liability as, from his point of view, the option has a negative payoff in some states. In our context, hedging this option means to form a self-financing trading strategy formed with the stock plus the risk-free asset in appropriate proportions such that the option plus this hedging portfolio is risk free. Hedging the option implies that the hedging portfolio perfectly replicates the option payoff in every possible state.

A European call option has only one payoff at the expiry date. It therefore suffices that the hedging portfolio replicates the option payoff at that date. Suppose that there is a self-financing trading strategy (θ_t^1, θ_t^2) in the bond and the stock such that

$$\theta_t^1 V_T + \theta_t^2 S_T = Y_T$$

To avoid arbitrage, the price of the option at any moment must be equal to the value of the hedging self-financing trading strategy. In fact, suppose that at any time $t < T$ the self-financing

strategy (θ_t^1, θ_t^2) has a value lower than the option:

$$\theta_t^1 V_t + \theta_t^2 S_t < Y_t$$

An investor could then sell the option for Y_t , make an investment $\theta_t^1 V_t + \theta_t^2 S_t$ in the trading strategy, and at time T liquidate both the option and the trading strategy. As $\theta_T^1 V_T + \theta_T^2 S_T = Y_T$ the final liquidation has value zero in every state of the world, so that the initial profit $Y_t - \theta_t^1 V_t + \theta_t^2 S_t$ is a risk-free profit. A similar reasoning could be applied if, at any time $t < T$, the strategy (θ_t^1, θ_t^2) had a value higher than the option. Therefore, we can conclude that if there is a self-financing trading strategy that replicates the option's payoff, the value of the strategy must coincide with the option's price at every instant prior to the expiry date.

Observe that the above reasoning is an instance of the law of one price. If two portfolios have the same payoffs at every moment and in every state of the world, their price must be the same. In particular, if a trading strategy has the same payoffs of an asset, its value must coincide with the price of that asset.

The Black-Scholes Option Pricing Formula

Let's now see how the price of the option can be computed. Assume that the price of the option is a function of time and of the price of the underlying stock: $Y_t = C(S_t, t)$. This assumption is reasonable but needs to be justified; for the moment it is only a hint as to how to proceed with the calculations. It will be justified later by verifying that the pricing formula produces the correct final payoff.

As S_t is assumed to be an Itô process, in particular a geometric Brownian motion, $Y_t = C(S_t, t)$ —which is a function of S_t —is an Itô process as well. Therefore, using Itô's formula, we can write down the stochastic equation that Y_t

must satisfy. Itô's formula prescribes that:

$$dY_t = \left[\frac{\partial C(S_t, t)}{\partial t} + \frac{\partial C(S_t, t)}{\partial S_t} S_t \mu + \frac{1}{2} \frac{\partial^2 C(S_t, t)}{\partial S_t^2} S_t^2 \sigma^2 \right] dt + \frac{\partial C(S_t, t)}{\partial S_t} \sigma S_t dB$$

Suppose now that there is a self-financing trading strategy $Y_t = \theta_t^1 V_t + \theta_t^2 S_t$. We can write this equation as

$$\int_0^t dY_t = \theta_t^1 \int_0^t dV_t + \theta_t^2 \int_0^t dS_t$$

or, in differential form, as

$$dY_t = \theta_t^1 dV_t + \theta_t^2 dS_t = (\theta_t^1 r V_t + \theta_t^2 \mu S_t) dt + \theta_t^2 \sigma S_t dB_t$$

If the trading strategy replicates the option price process, the two expressions for dY_t —the one obtained through Itô's lemma and the other obtained through the assumption that there is a replicating self-financing trading strategy—must be equal:

$$\begin{aligned} & (\theta_t^1 r V_t + \theta_t^2 \mu S_t) dt + \theta_t^2 \sigma S_t dB_t \\ &= \left[\frac{\partial C(S_t, t)}{\partial t} + \frac{\partial C(S_t, t)}{\partial S_t} S_t \mu + \frac{1}{2} \frac{\partial^2 C(S_t, t)}{\partial S_t^2} S_t^2 \sigma^2 \right] dt \\ &+ \frac{\partial C(S_t, t)}{\partial S_t} \sigma S_t dB_t \end{aligned}$$

The equality of these two expressions implies the equality of the coefficients in dt and dB respectively. Equating the coefficients in dB yields

$$\theta_t^2 = \frac{\partial C(S_t, t)}{\partial S_t}$$

As $Y_t = C(S_t, t) = \theta_t^1 V_t + \theta_t^2 S_t$, substituting, we obtain

$$\theta_t^1 = \frac{1}{V_t} \left[C(S_t, t) - \frac{\partial C(S_t, t)}{\partial S_t} S_t \right]$$

We have now obtained the self-financing trading strategy in function of the stock and option prices. Substituting and equating the

coefficients of dt yields

$$\begin{aligned} & \frac{1}{V_t} \left[C(S_t, t) - \frac{\partial C(S_t, t)}{\partial S_t} S_t \right] r V_t + \frac{\partial C(S_t, t)}{\partial S_t} \mu S_t \\ &= \frac{\partial C(S_t, t)}{\partial t} + \frac{\partial C(S_t, t)}{\partial S_t} S_t \mu + \frac{1}{2} \frac{\partial^2 C(S_t, t)}{\partial S_t^2} S_t^2 \sigma^2 \end{aligned}$$

Simplifying and eliminating common terms, we obtain

$$\begin{aligned} -rC(S_t, t) + r \frac{\partial C(S_t, t)}{\partial S_t} S_t + \frac{\partial C(S_t, t)}{\partial t} \\ + \frac{1}{2} \frac{\partial^2 C(S_t, t)}{\partial S_t^2} S_t^2 \sigma^2 = 0 \end{aligned}$$

If the function $C(S_t, t)$ satisfies this relationship, then the coefficients in dt match. The above relationship is a partial differential equation (PDE). This equation can be solved with suitable boundary conditions. Boundary conditions are provided by the payoff of the option at the expiry date:

$$Y_T = C(S_T, T) = \max(S_T - K, 0)$$

The closed-form solution of the above PDE with the above boundary conditions was derived by Black and Scholes (1973) and referred to as the *Black-Scholes option pricing formula*:

$$C(S_t, t) = x\Phi(z) - e^{-r(T-t)}K\Phi(z - \sigma\sqrt{T-t})$$

with

$$z = \frac{\log(S_t/K) + (r + \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}}$$

and where Φ is the cumulative normal distribution.

Let's stop for a moment and review the logical steps we have followed thus far. First, we defined a market made by a stock whose price process follows a geometric Brownian motion and a bond whose price process is a deterministic exponential. We introduced into this market a European call option. We then made two assumptions: (1) The option's price process is a deterministic function of the stock price process; and (2) the option's price process can be replicated by a self-financing trading strategy.

If the above assumptions are true, we can write a stochastic differential equation for the option's price process in two different ways: (1) Using Itô's lemma, we can write the option price stochastic process as a function of the stock stochastic process; and (2) using the assumption that there is a replicating trading strategy, we can write the option price stochastic process as the stochastic process of the trading process. As the two equations describe the same process, they must coincide. Equating the coefficients in the deterministic and stochastic terms, we can determine the trading strategy and write a deterministic PDE that the pricing function of the option must satisfy. The latter PDE together with the boundary conditions provided by the known value of the option at the expiry date uniquely determine the option pricing function.

Note that the above is neither a demonstration that there is an option pricing function, nor a demonstration that there is a replicating trading strategy. However, if both a pricing function and a replicating trading strategy exist, the above process allows one to determine both by solving a partial differential equation. After determining a solution to the PDE, one can verify if it provides a pricing function and if it allows the creation of a self-financing trading strategy. Ultimately, the justification of the existence of an option's pricing function and of a replicating self-financing trading strategy resides in the possibility of actually determining both. Absence of arbitrage ensures that this solution is unique.

Generalizing the Pricing of European Options

We can now generalize the above pricing methodology to a generic European option and to more general price processes for the bond and for the underlying stock. In the most general case, the process underlying a derivative need not be a stock price process. However, we

suppose that the underlying is a stock price process so that replicating portfolios can be formed. We generalize in three ways:

- The option's payoff is an arbitrary finite-variance random variable.
- The stock price process is an Itô process.
- The short-rate process is stochastic.

Following the definition given in the finite-state setting, we define a European option on some underlying process S_t as an asset whose payoff at time T is given by the random variable $Y_T = g(S_T)$ where $g(x)$, $x \in R$ is a continuous real-valued function. In other words, a European option is defined as a security whose payoff is determined at a given expiry date T as a function of some underlying random variable. The option has a zero payoff at every other date $t \in [0, T]$. This definition clearly distinguishes European options from American options, which yield payoffs at random stopping times.

Let's now generalize the price process of the underlying stock. We represent the underlying stock price process as a generic Itô process. A generic univariate Itô process can be represented through the differential stochastic equation:

$$dS_t = \mu(S_t, t)dt + \sigma(S_t, t)dB_t; S_0 = x$$

where x is the initial condition, B is a standard Brownian motion, and $\mu(S_t, t)$ and $\sigma(S_t, t)$ are given functions $R \times (0, \infty) \rightarrow R$. The geometric Brownian motion is a particular example of an Itô process.

Let's now define the bond price process. We retain the risk-free nature of the bond but let the interest rate be stochastic. Recall that in a discrete-state, discrete-time setting, a bond was defined as a process that, at each time step, exhibits the same return for each state though the return can be different in different time steps. Consequently, in continuous-time we define a bond price process as the following integral:

$$V_t = V_0 e^{\int_0^t r(S_u, u) du}$$

where r is a given function that represents the stochastic rate. In fact, the rate r depends on the time t and on the stock price process S_t . Application of Itô's lemma shows that the bond price process satisfies the following equation:

$$dV_t = V_t r(S_t, t) dt$$

We can now use the same reasoning that led to the Black-Scholes formula. Suppose that there are both an option pricing function $Y_t = C(S_t, t)$ and a replicating self-financing trading strategy

$$Y_t = \theta_t^1 V_t + \theta_t^2 S_t$$

We can now write a stochastic differential equation for the process Y_t in two ways:

- Applying Itô's lemma to $Y_t = C(S_t, t)$
- Directly to $Y_t = \theta_t^1 V_t + \theta_t^2 S_t$

The first approach yields

$$dY_t = \left[\frac{\partial C(S_t, t)}{\partial t} + \frac{\partial C(S_t, t)}{\partial S_t} \mu(S_t, t) + \frac{1}{2} \frac{\partial^2 C(S_t, t)}{\partial S_t^2} \sigma^2(S_t, t) \right] dt + \frac{\partial C(S_t, t)}{\partial S_t} \sigma(S_t, t) dB_t$$

The second approach yields

$$dY_t = [\theta_t^1 r(S_t, t) V_t + \theta_t^2 \mu(S_t, t)] dt + \theta_t^2 \sigma(S_t, t) dB_t$$

Equating coefficients in dt, dB we obtain the trading strategy

$$\theta_t^1 = \frac{1}{V_t} \left[C(S_t, t) - \frac{\partial C(S_t, t)}{\partial S_t} S_t \right]$$

$$\theta_t^2 = \frac{\partial C(S_t, t)}{\partial S_t}$$

and the PDE

$$-r(x, t)C(x, t) + r(x, t) \frac{\partial C(x, t)}{\partial x} x + \frac{\partial C(x, t)}{\partial t} + \frac{1}{2} \frac{\partial^2 C(x, t)}{\partial x^2} \sigma^2(x, t) = 0$$

with the boundary conditions $C(S_T, T) = g(S_T)$. Solving this equation we obtain a candidate option pricing function. In each specific case, one

can then verify that the option pricing function effectively solves the option pricing problem.

STATE-PRICE DEFLATORS

We now extend the concepts of state prices and equivalent martingale measures to a continuous-state, continuous-time setting. As in the previous sections, the economy is represented by a probability space $(\Omega, \mathfrak{F}, P)$ where Ω is the set of possible states, \mathfrak{F} is the σ -algebra of events, and P is a probability measure. Time is a continuous variable in the interval $[0, T]$. The propagation of information is represented by a filtration \mathfrak{F}_t . A multivariate standard Brownian motion $B = (B_1, \dots, B_D)$ in R^D adapted to the filtration \mathfrak{F}_t is defined over this probability space. We know that there are mathematical subtleties that we will not take into consideration, as regards whether (1) the filtration is given and the Brownian motion is adapted to the filtration or (2) the filtration is generated by the Brownian motion.

Suppose that there are N price processes $\mathbf{X} = (X^1, \dots, X^N)$ that form a multivariate Itô process in R^N . Trading strategies are adapted processes $\theta = (\theta^1, \dots, \theta^N)$ that represent the quantity of each asset held at each instant. In order to ensure the existence of stochastic integrals, we require the processes (X^1, \dots, X^N) and any trading strategy to be of bounded variation. Let's first suppose that there is no payoff-rate process. This assumption will be relaxed in a later section. Suppose also that one of these processes, say X_t^1 , is defined by a short-rate process r , so that

$$X_t^1 = e^{\int_0^t r_u du}$$

or

$$dX_t^1 = r_t X_t^1 dt$$

where r_t is a deterministic function of t called the short-rate process. Note that X_t^1 could be replaced by a trading strategy. We can think of r_t as the risk-free short-term continuously compounding interest rate and of X_t^1

as a risk-free continuously compounding bank account.

The concept of arbitrage and of trading strategy was defined in the previous section. We now introduce the concept of deflators in a continuous-time continuous-state setting. Any strictly positive Itô process is called a *deflator*. Given a deflator Y we can deflate any process X , obtaining a new deflated process

$$X_t^Y = X_t Y_t$$

For example, any stock price process of a non-defaulting firm or the risk-free bank account is a deflator. For technical reasons it is necessary to introduce the concept of regular deflators. A *regular deflator* is a deflator that, after deflation, leaves unchanged the set of admissible bounded-variation trading strategies.

We can make the first step towards defining a theory of pricing based on equivalent martingale measures. It can be demonstrated that if Y is a regular deflator, a trading strategy θ is self-financing with respect to the price process $\mathbf{X} = (X^1, \dots, X^N)$ if and only if it is self-financing with respect to the deflated price process

$$\mathbf{X}^Y = (Y_t X_t^1, \dots, Y_t X_t^N)$$

In addition, it can be demonstrated that the price process $\mathbf{X} = (X^1, \dots, X^N)$ admits no arbitrage if and only if the deflated price process

$$\mathbf{X}^Y = (Y_t X_t^1, \dots, Y_t X_t^N)$$

admits no arbitrage.

A *state-price deflator* is a deflator π with the property that the deflated price process \mathbf{X}^π is a martingale. A martingale is a stochastic process M_t such that its current value equals the conditional expectation of the process at any future time: $M_t = E_t[M_s]$, $s > t$. For each price process X_t^i , the following relationship therefore holds:

$$\pi_t X_t^i = E_t[\pi_s X_s^i], s > t$$

This definition is the equivalent in continuous time of the definition of a state-price deflator in discrete time. In fact, a state-price deflator is

defined as a process π such that

$$S_t^i = \frac{1}{\pi_t} E_t \left[\sum_{j=t+1}^T \pi_j d_j^i \right]$$

If there is no intermediate payoff, as in our present case, the previous relationship can be written as

$$\begin{aligned} \pi_t S_t^i &= E_t[\pi_T S_T^i] = E_t[E_{t+1}[\pi_T S_T^i]] \\ &= E_t[\pi_{t+1} S_{t+1}^i] \end{aligned}$$

The next proposition states that if there is a regular state-price deflator, then there is no arbitrage. The demonstration of this proposition hinges on the fact that, as the deflated price process is a martingale, the following relationship holds:

$$E \left[\int_0^T \theta_u dS_u^\pi \right] = 0$$

and therefore any self-financing trading strategy is a martingale. We can thus write

$$\theta_0 S_0^\pi = E[\theta_T S_T^\pi]$$

If

$$\begin{aligned} \theta_T S_T^\pi \geq 0 \quad \text{then} \quad \theta_0 S_0^\pi \geq 0 \\ \text{and if} \quad \theta_T S_T^\pi > 0 \quad \text{then} \quad \theta_0 S_0^\pi > 0 \end{aligned}$$

which shows that there cannot be any arbitrage.

We have now stated that the existence of state-price deflators ensures the absence of arbitrage. The converse of this statement in a continuous-state, continuous-time setting is more delicate and will be dealt with later. We will now move on to equivalent martingale measures.

EQUIVALENT MARTINGALE MEASURES

In the previous section we saw that if there is a regular state-price deflator then there is no arbitrage. A state-price deflator transforms every price process and every self-financing trading strategy into a martingale. We will now see that,

after discounting by an appropriate process, price processes become martingales through a transformation of the real probability measure into an equivalent martingale measure.² This theory parallels the theory of equivalent martingale measures developed in the discrete-state, discrete-time setting in the entry "Arbitrage Pricing: Finite-State Models." First some definitions must be discussed.

Given a probability measure P , the probability measure Q is said to be equivalent to P if both assign probability zero to the same events, that is, if $P(A) = 0$ if and only if $Q(A) = 0$ for every event A . The equivalent probability measure Q is said to be an *equivalent martingale measure* for the process X if X is a martingale with respect to Q and if the Radon-Nikodym derivative

$$\xi = \frac{dQ}{dP}$$

has finite variance. The definition of the Radon-Nikodym derivative is the same here as it is in the finite-state context. The Radon-Nikodym derivative is a random variable ξ such that $Q(A) = E^P[\xi I_A]$ for every event A where I_A is the indicator function of the event A .

To develop an intuition for this definition, consider that any stochastic process X is a time-dependent random variable X_t . The latter is a family of functions $\Omega \rightarrow R$ from the set of states to the real numbers indexed with time such that the sets $\{X_t(\omega) \leq x\}$ are events for any real x . Given the probability measure P , the finite-dimension distributions of the process X are determined. The equivalent measure Q determines another set of finite-dimension distributions. However, the correspondence between the process paths and the states remains unchanged.

The requirement that P and Q are equivalent is necessary to ensure that the process is effectively the same under the two measures. There is no assurance that given an arbitrary process an equivalent martingale measure exists. Let's assume that an equivalent martingale measure does exist for the N -dimensional price process

$\mathbf{X} = (X^1, \dots, X^N)$. It can be demonstrated that if the price process $\mathbf{X} = (X^1, \dots, X^N)$ admits an equivalent martingale measure, then there is no arbitrage.

The proof is similar to that for state-price deflators as discussed above. Under the equivalent martingale measure Q , which we assume exists, every price process and every self-financing trading strategy becomes a martingale. Using the same reasoning as above it is easy to see that there is no arbitrage.

This result can be generalized; here is how. If there is a regular deflator Y such that the deflated price process $\mathbf{X}^Y = (Y_t X_t^1, \dots, Y_t X_t^N)$ admits an equivalent martingale measure, then there is no arbitrage. The proof hinges on the result established in the previous section that, if there is a regular deflator Y , the price process \mathbf{X} admits no arbitrage if and only if the deflated price process \mathbf{X}^Y admits no arbitrage.

Note that none of these results is constructive. They only state that the existence of an equivalent martingale measure with respect to a price process ensures the absence of arbitrage. Conditions to ensure the existence of an equivalent martingale measure with respect to a price process are given in the next section.

EQUIVALENT MARTINGALE MEASURES AND GIRSANOV'S THEOREM

We first need to establish an important mathematical result known as *Girsanov's theorem*. This theorem applies to Itô processes. Let's first state Girsanov's theorem in simple cases. Let X be a single-valued Itô process where B is a single-valued standard Brownian motion:

$$X_t = x + \int_0^t \mu_s ds + \int_0^t \sigma_s dB_s$$

Suppose that a process v and a process θ such that $\sigma_t \theta_t = \mu_t - v_t$ are given. Suppose, in addition, that the process θ satisfies the Novikov

condition which requires

$$E \left[e^{\left(\frac{1}{2} \int_0^t \theta_s^2 ds \right)} \right] < \infty$$

Then, there is a probability measure Q equivalent to P such that the following integral

$$\hat{B}_t = B_t + \int_0^t \theta_s ds$$

defines a standard Brownian motion \hat{B}_t in R on $(\Omega, \mathfrak{F}, Q)$ with the same standard filtration of the original Brownian motion B_t . In addition, under Q the process X becomes

$$X_t = x + \int_0^t v_s ds + \int_0^t \sigma_s d\hat{B}_s$$

Girsanov's theorem states that we can add drift to a standard Brownian motion and still obtain a standard Brownian motion under another probability measure. In addition, by changing the probability measure we can arbitrarily change the drift of an Itô process.

The same theorem can be stated in multiple dimensions. Let X be an N -valued Itô process:

$$X_t = x + \int_0^t \mu_s ds + \int_0^t \sigma_s dB_s$$

In this process, μ_s is an N -vector process and σ_s is an $N \times D$ matrix. Suppose that there are both a vector process $v = (v^1, \dots, v^N)$ and a vector process $\theta = (\theta^1, \dots, \theta^N)$ such that $\sigma_t \theta_t = \mu_t - v_t$ where the product $\sigma_t \theta_t$ is not a scalar product but is performed component by component. Suppose, in addition, that the process θ satisfies the Novikov condition:

$$E \left[e^{\left(\frac{1}{2} \int_0^t \theta \cdot \theta ds \right)} \right] < \infty$$

Then there is a probability measure Q equivalent to P such that the following integral

$$\hat{B}_t = B_t + \int_0^t \theta_s ds$$

defines a standard Brownian motion \hat{B}_t in R^D on $(\Omega, \mathfrak{F}, Q)$ with the same standard filtration of the original Brownian motion B_t . In addition, under Q the process X becomes

$$X_t = x + \int_0^t v_s ds + \int_0^t \sigma_s d\hat{B}_s$$

Girsanov's theorem essentially states that under technical conditions (the Novikov condition) by changing the probability measure, it is possible to transform an Itô process into another Itô process with arbitrary drift. Prima facie, this result might seem unreasonable. In the end the drift of a process seems to be a fundamental feature of the process as it defines, for example, the average of the process. Consider, however, that a stochastic process can be thought as the set of all its possible paths. In the case of an Itô process, we can identify the process with the set of all continuous and square integrable functions. As observed above, the drift is an average, and it is determined by the probability measure on which the process is defined. Therefore, it should not be surprising that by changing the probability measure it is possible to change the drift.

The Diffusion Invariance Principle

Note that Girsanov's theorem requires neither that the process X be a martingale nor that Q be an equivalent martingale measure. If X is indeed a martingale under Q , an implication of Girsanov's theorem is the diffusion invariance principle, which can be stated as follows. Let X be an Itô process:

$$dX_t = \mu_t dt + \sigma_t dB_t$$

If X is a martingale with respect to an equivalent probability measure Q , then there is a standard Brownian motion \hat{B}_T in R^D under Q such that

$$dX_t = \sigma_t d\hat{B}_t$$

Let's now apply the previous results to a price process $X = (V, S^1, \dots, S^{N-1})$ where

$$dS_t = \mu_t dt + \sigma_t dB_t$$

and

$$dV_t = r_t V_t dt$$

If the short-term rate r is bounded, V_t^{-1} is a regular deflator. Consider the deflated processes:

$$Z_t = S_t V_t^{-1}$$

By Itô's lemma, this process satisfies the following stochastic equation:

$$dZ_t = \left(-r_t Z_t + \frac{\mu_t}{V_t} \right) dt + \frac{\sigma_t}{V_t} dB_t$$

Suppose there is an equivalent martingale measure Q . Under the equivalent martingale measure Q , the discounted price process

$$Z_t = S_t V_t^{-1}$$

is a martingale. In addition, by the diffusion invariance principle there is a standard Brownian motion \hat{B}_t in R^D under Q such that:

$$dZ_t = \frac{\sigma_t}{V_t} d\hat{B}_t$$

Applying Itô's lemma, given that $Z_t V_t = S_t$, we obtain the fundamental result:

$$dS_t = r_t dt + \sigma_t d\hat{B}_t$$

This result states that, under the equivalent martingale measure, all price processes become Itô processes with the same drift.

Application of Girsanov's Theorem to Black-Scholes Option Pricing Formula

To illustrate Girsanov's theorem, let's see how the Black-Scholes option pricing formula can be obtained from an equivalent martingale measure. In the previous setting, let's assume that $N = 3$, $d = 1$, r_t is a constant and

$$\sigma_t = \sigma S_t$$

with σ constant. Let S be the stock price process and C be the option price process. The option's price at time T is

$$C = \max(S_T^1 - K)$$

In this setting, therefore, the following three equations hold:

$$\begin{aligned} dS_t &= \mu_t^S dt + \sigma S_t^S dB_t \\ dC_t^2 &= \mu_t^C dt + \sigma_t^C dB_t \\ dV_t &= rV_t dt \end{aligned}$$

Given that $C_t V_t^{-1}$ is a martingale, we can write

$$C_t = V_t E_t^Q \left[\frac{C_T^2}{V_t} \right] = E_t^Q [e^{-r(T-t)} \max(S_T - K)]$$

It can be demonstrated by direct computation that the above formula is equal to the Black-Scholes option pricing formula presented earlier in this entry.

EQUIVALENT MARTINGALE MEASURES AND COMPLETE MARKETS

In the continuous-state, continuous-time setting, a market is said to be complete if any finite-variance random variable Y can be obtained as the terminal value at time T of a self-financing trading strategy θ : $Y = \theta_T X_T$. A fundamental theorem of arbitrage pricing states that, in the absence of arbitrage, a market is complete if and only if there is a unique equivalent martingale measure. This condition can be made more specific given that the market is populated with assets that follow Itô processes. Suppose that the price process is $\mathbf{X} = (V, S^1, \dots, S^{N-1})$ where, as in the previous section:

$$\begin{aligned} dS_t &= \mu_t dt + \sigma_t dB_t \\ dV_t &= rV_t dt \end{aligned}$$

and \mathbf{B} is a standard Brownian motion $B = (B^1, \dots, B^D)$ in R^D .

It can be demonstrated that markets are complete if and only if $\text{rank}(\sigma) = d$ almost everywhere. This condition should be compared with

the conditions for completeness we established in the discrete-state setting. In that setting, we demonstrated that markets are complete if and only if the number of linearly independent price processes is equal to the maximum number of branches leaving a node. In fact, market completeness is equivalent to the possibility of solving a linear system with as many equations as branches leaving each node.

In the present continuous-state setting, there are infinite states and so we need different types of considerations. Roughly speaking, each price process (which is an Itô process) depends on D independent sources of uncertainty as we assume that the standard Brownian motion is D -dimensional. In a finite-state setting this means that, if processes are Markovian, at each time step any process can jump to D different values. The market is complete if there are D independent price processes. Note that the number D is arbitrary.

EQUIVALENT MARTINGALE MEASURES AND STATE PRICES

We will now show that equivalent martingale measures and state prices are the same concept. We use the same setting as in the previous sections. Suppose that Q is an equivalent martingale measure after deflation by the process

$$\frac{1}{V_t^1} = e^{\int_0^t -r_u du}$$

where r is a bounded short-rate process. The density process ξ_t for Q is defined as

$$\xi_t = E_r \left[\frac{dQ}{dP} \right], t \in [0, T]$$

where

$$\left[\frac{dQ}{dP} \right]$$

is the Radon-Nikodym derivative of Q with respect to P . As in the discrete-state setting, the Radon-Nikodym derivative of Q with respect

to P is a random variable

$$\xi = \left[\frac{dQ}{dP} \right]$$

with average value on the entire space equal to 1 and such that, for every event A , the probability of A under Q is the average of ξ :

$$P^Q(A) = E_A[\xi]$$

It can be demonstrated that, given any \mathfrak{F}_t -measurable random variable W , the density process ξ_t for Q has the following property:

$$E_t^Q[W] = \frac{E_t[W\xi_t]}{\xi_t}$$

To gain an intuition for the Radon-Nikodym derivative in a continuous-state setting, let's assume that the probability space is the real line equipped with the Borel σ -algebra and with a probability measure P . In this case, $\xi = \xi(x)$, $R \rightarrow R$ and we can write

$$Q(A) = \int_A \xi dP$$

or, $dQ = \xi dP$. Given any random variable X with density f under P and density q under Q , we can then write

$$E^Q[X] = \int_R xq(x)dx = \int_R x\xi(x)f(x)dx$$

In other words, the random variable ξ is a function that multiplies the density f to yield the density q .

We can now show the following key result. Given an equivalent martingale measure with density process ξ_t a state-price deflator is given by the process

$$\pi_t = \xi_t e^{\int_0^t -r_u du}$$

Conversely, given a state-price deflator π_t , the density process

$$\xi_t = e^{\int_0^t r_u du} \frac{\pi_t}{\pi_0}$$

defines an equivalent martingale measure. In fact, suppose that Q is an equivalent martingale

measure for X^Y with $\pi_t = \xi_t Y_t$ where

$$Y_t = e^{\int_0^t -r_u du}$$

Then, using the above relationship we can write:

$$\begin{aligned} E_t[\pi_t X_t] &= E_t[\xi_t X_t^Y] = \xi_t E_t^Q[\xi_t X_t^Y] = \xi_t X_t^Y \\ &= \pi_t X_t \end{aligned}$$

which shows that π_t is a state-price deflator. The same reasoning in reverse order demonstrates that if π_t is a state-price deflator then:

$$\xi_t = e^{\int_0^t r_u du} \frac{\pi_t}{\pi_0}$$

is a density process for Q .

ARBITRAGE PRICING WITH A PAYOFF RATE

In the analysis thus far, we assumed that there is no intermediate payoff. The owner of an asset makes a profit or a loss due only to the changes in value of the asset. Let's now introduce a payoff-rate process δ_t^i for each asset i . The payoff-rate process must be interpreted in the sense that the cumulative payoff of each individual asset is

$$D_t^i = \int_0^t \delta_s^i ds$$

We define a gain process

$$G_t^i = S_t^i + D_t^i$$

By the linearity of the Itô integrals, we can write any trading strategy as

$$\int_0^t \theta_t dG_t = \int_0^t \theta_t dX_t + \int_0^t \theta_t dD_t$$

If there is a payoff-rate process, a self-financing trading strategy is a trading strategy

such that the following relationship holds:

$$\theta_t \mathbf{S}_t = \sum_i \theta_t^i S_t^i = \sum_i \left(\theta_t^i S_t^i + \int_0^t \theta_t^i dG_t^i \right), t \in [0, T]$$

An arbitrage is, as before, a self-financing trading strategy such that

$$\theta_0 \mathbf{S}_0 < 0 \text{ and } \theta_T \mathbf{S}_T \geq 0, \text{ or } \theta_0 \mathbf{S}_0 \leq 0 \text{ and } \theta_T \mathbf{S}_T > 0$$

The previous arguments extend to this case. An equivalent martingale measure for the pair (D, S) is defined as an equivalent probability measure Q such that the Radon-Nikodym derivative

$$\xi = \left[\frac{dQ}{dP} \right]$$

has finite variance and the process $G = S + D$ is a martingale. Under these conditions, the following relationship holds:

$$S_t = E_t^Q \left[e^{\int_t^T -r_u du} + \int_t^T e^{\int_t^s -r_u du} dD_s \right]$$

IMPLICATIONS OF THE ABSENCE OF ARBITRAGE

We saw that the existence of an equivalent martingale measure or of state-price deflators implies absence of arbitrage. We have also seen that, in the absence of arbitrage, markets are complete if and only if there is a unique equivalent martingale measure.

In a discrete-state, discrete-time context we could establish the complete equivalence between the existence of state-price deflators, equivalent martingale measures and absence of arbitrage, in the sense that any of these conditions implies the other two. In addition, the existence of a unique equivalent martingale measure implies absence of arbitrage and market completeness.

In the present continuous-state context, however, absence of arbitrage implies the existence

of an equivalent martingale measure and of state price deflators only under rather restrictive and complex technical conditions. If we want to relax these conditions, the condition of absence of arbitrage has to be slightly modified. These discussions are quite technical and will not be presented in this entry.³

WORKING WITH EQUIVALENT MARTINGALE MEASURES

The concepts established in the preceding sections of this entry might seem very complex, abstract, and scarcely useful. On the contrary, they entail important simplifications in the computation of derivative prices. Applications of these computations can be found in the pricing of bonds and credit derivatives. Here we want to make a few general comments on how these tools are used.

The key result of the arbitrage pricing theory is that, under the equivalent martingale measure, all discounted price processes become martingales and all price processes have the same drift. Therefore, all calculations can be performed under the assumption that the change to an equivalent martingale measure has been made. This environment allows important simplifications. For example, as we have seen, the option pricing problem becomes a problem of computing the present value of simpler processes.

Obviously one has to go back to a real environment at the end of the pricing exercise. This is essentially a calibration problem, as risk-neutral probabilities have to be estimated from real probabilities. Despite this complication, the equivalent martingale methodology has proved to be an important tool in derivative pricing.

KEY POINTS

- A trading strategy is a vector-valued process that represents portfolio weights at each moment.

- Trading gains are defined as stochastic integrals.
- A self-financing trading strategy is one whose value at every moment is the initial value plus the trading gains at that moment.
- An arbitrage is a self-financing trading strategy whose initial value is either negative and the final value nonnegative or the initial value non-negative and the final value positive.
- The Black-Scholes option pricing formula can be established by replicating self-financing trading strategies.
- The Black-Scholes pricing argument is based on constructing a self-financing trading strategy that replicates the option price in each state and for each time.
- Absence of arbitrage implies that a replicating self-financing trading strategy must have the same price as the option.
- The Black-Scholes option pricing formula is obtained by solving the partial differential equation implied by the equality of the replicating self-financing trading strategy and the option price process.
- A deflator is any strictly positive Itô process; a state-price deflator is a deflator with the property that the deflated price process is a martingale.
- If there is a (regular) state-price deflator, then there is no arbitrage; the converse is true only under a number of technical conditions.
- Two probability measures are said to be equivalent if they assign probability zero to the same event.
- Given a process X on a probability space with probability measure P , the probability measure Q is said to be an equivalent martingale measure if it is equivalent to P and X is a martingale with respect to Q (plus other conditions).
- If there is a regular deflator such that the deflated price process admits an equivalent martingale measure, then there is no arbitrage.
- Under the equivalent martingale measure, all Itô price processes have the same drift.
- In the absence of arbitrage, a market is complete if and only if there is a unique equivalent martingale measure.

NOTES

1. One can visualize this process as a tree structure with an infinite number of branches and an infinite number of branching points. However, as the number of branches and of branching points is a continuum, intuition might be misleading.
2. The theory of equivalent martingale measures was developed in Harrison and Pliska (1981, 1985) and Harrison and Kreps (1979).
3. See Delbaen and Schachermayer (1994, 1999).

REFERENCES

- Black, B., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.
- Friedman, M. (1953). *Essays in the Theory of Positive Economics*. Chicago: University of Chicago Press.
- Delbaen, F., and Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing. *Mathematische Annalen* 300 (3): 463–520.
- Delbaen, F., and Schachermayer, W. (1999). The fundamental theorem of asset pricing for unbounded stochastic processes. *Mathematische Annalen* 312 (2): 215–250.
- Harrison, J. M., and Kreps, D. M. (1979). Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* 20: 381–408.
- Harrison, J. M., and Pliska, S. R. (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Process Application* 11: 215–260.
- Harrison, J. M., and Pliska, S. R. (1985). A stochastic calculus model of continuous trading: Complete markets. *Stochastic Process Application* 15: 313–316.

Bayesian Analysis and Financial Modeling Applications

Basic Principles of Bayesian Analysis

BILIANA S. GÜNER, PhD

Assistant Professor of Statistics and Econometrics, Ozyegin University, Turkey

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

JOHN S. J. HSU, PhD

Professor of Statistics and Applied Probability, University of California, Santa Barbara

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: One of the basic mechanisms of learning is assimilating the information arriving from the external environment and then updating the existing knowledge base with that information. This mechanism lies at the heart of the Bayesian framework. A Bayesian decision maker learns by revising beliefs in light of the new data that become available. From the Bayesian point of view, probabilities are interpreted as degrees of belief. Therefore, the Bayesian learning process consists of revising probabilities. Contrast this with the way probability is interpreted in the classical (frequentist) statistical theory—as the relative frequency of occurrence of an event in the limit, as the number of observations goes to infinity. Bayes' theorem provides the formal means of putting that mechanism into action; it is a simple expression combining the knowledge about the distribution of the model parameters and the information about the parameters contained in the data.

Quantitative financial models describe in mathematical terms the relationships between financial random variables through time and/or across assets. The fundamental assumption is that the model relationship is valid independent of the time period or the asset class under consideration. Financial data contain both meaningful information and random noise. An adequate financial model not only extracts optimally the relevant information from the his-

torical data but also performs well when tested with new data. The uncertainty brought about by the presence of data noise makes imperative the use of statistical analysis as part of the process of financial model building, model evaluation, and model testing.

Statistical analysis is employed from the vantage point of either of the two main statistical philosophical traditions—frequentist and Bayesian. An important difference between the

two lies with the interpretation of the concept of probability. As the name suggests, advocates of frequentist statistics adopt a *frequentist* interpretation: The probability of an event is the limit of its long-run relative frequency (i.e., the frequency with which it occurs as the amount of data increases without bound). Strict adherence to this interpretation is not always possible in practice. When studying rare events, for instance, large samples of data may not be available and in such cases proponents of frequentist statistics resort to theoretical results. The Bayesian view of the world is based on the subjectivist interpretation of probability: Probability is subjective, a degree of belief that is updated as information or data are acquired.

The concept of subjective probability is derived from arguments for rationality of the preferences of agents. It originated in the 1930s with the (independent) works of Bruno de Finetti (1931) and Frank Ramsey (1931), and was further developed by Leonard Savage (1954) and Dennis Lindley (1971). The subjective probability interpretation can be traced back to the Scottish philosopher and economist David Hume, who also had philosophical influence over Harry Markowitz (by Markowitz's own words in his autobiography published in *Les Prix Nobel*, 1991).

Closely related to the concept of probability is that of uncertainty. Proponents of the frequentist approach consider the source of uncertainty to be the randomness inherent in realizations of a random variable. The probability distributions of variables are not subject to uncertainty. In contrast, Bayesian statistics treats probability distributions as uncertain and subject to modification as new information becomes available. Uncertainty is implicitly incorporated by probability updating. The probability beliefs based on the existing knowledge base take the form of the *prior probability*.

The *posterior probability* represents the updated beliefs. Since the beginning of the last century, when quantitative methods and models became a mainstream tool to aid in un-

derstanding financial markets and formulating investment strategies, the framework applied in finance has been the frequentist approach. The term *frequentist* usually refers to the Fisherian philosophical approach named after Sir Ronald Fisher.

Strictly speaking, "Fisherian" has a broader meaning as it includes not only frequentist statistical concepts such as unbiased estimators, hypothesis tests, and confidence intervals, but also the maximum likelihood estimation framework pioneered by Fisher. Only in the last two decades has Bayesian statistics started to gain greater acceptance in financial modeling, despite its introduction about 250 years ago by Thomas Bayes, a British minister and mathematician. It has been the advancements of computing power and the development of new computational methods that has fostered the growing use of Bayesian statistics in finance.

On the applicability of the Bayesian conceptual framework, consider an excerpt from the speech of the former chairman of the Board of Governors of the Federal Reserve System, Alan Greenspan, at the Meeting of the American Statistical Association in San Diego, California, January 3, 2004:

The Federal Reserve's experiences over the past two decades make it clear that uncertainty is not just a pervasive feature of the monetary policy landscape; it is the defining characteristic of that landscape. The term "uncertainty" is meant here to encompass both "Knightian uncertainty," in which the probability distribution of outcomes is unknown, and "risk," in which uncertainty of outcomes is delimited by a known probability distribution. . . . This conceptual framework emphasizes understanding as much as possible the many sources of risk and uncertainty that policymakers face, quantifying those risks when possible, and assessing the costs associated with each of the risks. In essence, the risk management approach to monetary policymaking is an application of Bayesian [decision making].

The three steps of Bayesian decision making that Alan Greenspan outlines are:

1. Formulating the prior probabilities to reflect existing information.

2. Constructing the quantitative model, taking care to incorporate the uncertainty intrinsic in model assumptions.
3. Selecting and evaluating a utility function describing how uncertainty affects alternative model decisions.

While these steps constitute the rigorous approach to Bayesian decision making, applications of Bayesian methods to financial modeling often only involve the first two steps or even only the second step. This tendency is a reflection of the pragmatic Bayesian approach that financial modelers often favor.

Applications of the *Bayesian framework* to financial modeling include:

- Bayesian approach to mean-variance portfolio selection.
- Reflecting degrees of belief in an asset pricing model when selecting an optimal portfolio.
- Bayesian methods of portfolio selection within the context of the Black-Litterman model.
- Computing measures of market efficiency.
- Estimating complex volatility models.

All of these applications are presented in Rachev et al. (2008).

In this entry, we discuss some of the basic principles of Bayesian analysis.

THE LIKELIHOOD FUNCTION

Suppose we are interested in analyzing the returns on a given stock and have available a historical record of returns. Any analysis of these returns, beyond a very basic one, would require that we make an educated guess about (propose) a process that might have generated these return data. Assume that we have decided on some statistical distribution and denote it by

$$p(y|\theta) \quad (1)$$

where y is a realization of the random variable Y (stock return) and θ is a parameter specific to

the distribution, p . Assuming that the distribution we proposed is the one that generated the observed data, we draw a conclusion about the value of θ . Obviously, central to that goal is our ability to summarize the information contained in the data. The likelihood function is a statistical construct with this precise role. Denote the n observed stock returns by y_1, y_2, \dots, y_n . The joint density function of Y , for a given value of θ , is

$$f(y_1, y_2, \dots, y_n | \theta)$$

By using the term “density function,” we implicitly assume that the distribution chosen for the stock return is continuous, which is invariably the case in financial modeling.

We can observe that the function above can also be treated as a function of the unknown parameter, θ , given the observed stock returns. That function of θ is called the *likelihood function*. We write it as

$$L(\theta | y_1, y_2, \dots, y_n) = f(y_1, y_2, \dots, y_n | \theta) \quad (2)$$

Suppose we have determined from the data two competing values of θ , θ_1 and θ_2 , and want to determine which one is more likely to be the true value (at least, which one is closer to the true value). The likelihood function helps us make that decision. Assuming that our data were indeed generated by the distribution in (1), θ_1 is more likely than θ_2 to be the true parameter value whenever $L(\theta_1 | y_1, y_2, \dots, y_n) > L(\theta_2 | y_1, y_2, \dots, y_n)$. This observation provides the intuition behind the method most often employed in “classical” statistical inference to estimate θ from the data alone—the *method of maximum likelihood*. The value of θ most likely to have yielded the observed sample of stock return data, y_1, y_2, \dots, y_n , is the maximum likelihood estimate, $\hat{\theta}$, obtained from maximizing the likelihood function in (2).

To illustrate the concept of a likelihood function, we briefly discuss two examples—one based on the Poisson distribution (a discrete

distribution) and another based on the normal distribution (one of the most commonly employed continuous distributions).

The Poisson Distribution Likelihood Function

The Poisson distribution is often used to describe the random number of events occurring within a certain period of time. It has a single parameter, θ , indicating the rate of occurrence of the random event, that is, how many events happen on average per unit of time. The probability distribution of a Poisson random variable, X , is described by the following expression:

$$p(X = k) = \frac{\theta^k}{k!} e^{-\theta}, \quad k = 0, 1, 2, \dots \quad (3)$$

The Poisson distribution is employed in the context of finance (most often, but not exclusively, in the areas of credit risk and operational risk) as the distribution of a stochastic process, called the Poisson process, which governs the occurrences of random events.

Suppose we are interested in examining the annual number of defaults of North American corporate bond issuers and we have gathered a sample of data for the period from 1986 through 2005. Assume that these corporate defaults occur according to a Poisson distribution. Denoting the 20 observations by x_1, x_2, \dots, x_{20} , we write the likelihood function for the Poisson parameter θ (the average rate of defaults) as¹

$$\begin{aligned} L(\theta | x_1, x_2, \dots, x_{20}) &= \prod_{i=1}^{20} p(X = x_i | \theta) = \prod_{i=1}^{20} \frac{\theta^{x_i}}{x_i!} e^{-\theta} \\ &= \frac{\theta^{\sum_{i=1}^{20} x_i}}{\prod_{i=1}^{20} x_i!} e^{-20\theta} \end{aligned} \quad (4)$$

It is often customary to retain in the expressions for the likelihood function and the probability distributions only the terms that contain the unknown parameter(s); that is, we get rid of the terms that are constant with respect to the pa-

rameter(s). Thus, (4) could be written as

$$L(\theta | x_1, x_2, \dots, x_{20}) \propto \theta^{\sum_{i=1}^{20} x_i} e^{-20\theta} \quad (5)$$

where \propto denotes “proportional to.” Clearly, for a given sample of data, the expressions in (4) and (5) are proportional to each other and therefore contain the same information about θ . Maximizing either of them with respect to θ , we obtain that the maximum likelihood estimator of the Poisson parameter, θ , is the sample mean, \bar{x} :

$$\hat{\theta} = \bar{x} = \frac{\sum_{i=1}^{20} x_i}{20}$$

For the 20 observations of annual corporate defaults, we get a sample mean of 51.6. The Poisson probability distribution function (evaluated at θ equal to its maximum-likelihood estimate, $\hat{\theta} = 51.6$) and the likelihood function for θ can be visualized, respectively, in the left-hand-side and right-hand-side plots in Figure 1.

The Normal Distribution Likelihood Function

The normal distribution (also called the Gaussian distribution) has been the predominant distribution of choice in finance because of the relative ease of dealing with it and the availability of attractive theoretical results resting on it.² It is certainly one of the most important distributions in statistics. Two parameters describe the normal distribution—the location parameter, μ , which is also its mean, and the scale (dispersion) parameter, σ , also called standard deviation. The probability density function of a normally distributed random variable Y is expressed as

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \quad (6)$$

where y and μ could take any real value and σ can only take positive values. We denote the distribution of Y by $Y \sim N(\mu, \sigma)$. The normal density is symmetric around the mean, μ , and its plot resembles a bell.

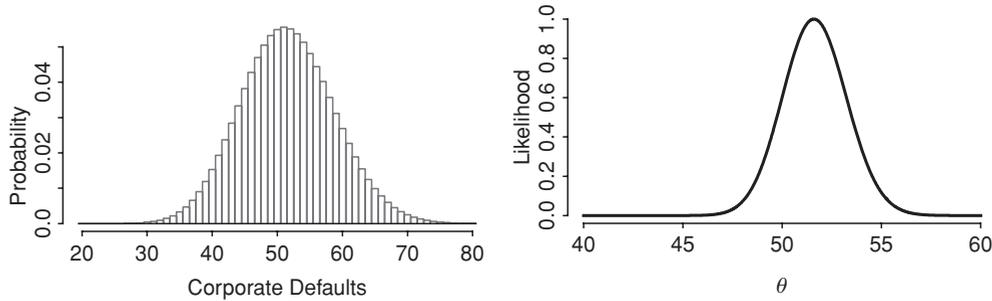


Figure 1 The Poisson Distribution Function and Likelihood Function

Note: The graph on the left represents the mass function of the Poisson random variable evaluated at the maximum-likelihood estimate, $\hat{\theta} = 51.6$. The graph on the right represents the likelihood function for the parameter of the Poisson distribution.

Suppose we have gathered daily dollar return data on the MSCI-Germany Index for the period January 2, 1998, through December 31, 2003 (a total of 1,548 returns), and we assume that the daily return is normally distributed. Then, given the realized index returns (denoted by $y_1, y_2, \dots, y_{1548}$), the likelihood function for the parameters μ and σ is written in the following way:

$$\begin{aligned}
 L(\mu, \sigma \mid y_1, y_2, \dots, y_{1548}) &= \prod_{i=1}^{1548} f(y_i) \\
 &= \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^{1548} e^{-\sum_{i=1}^{1548} \frac{(y_i - \mu)^2}{2\sigma^2}} \\
 &\propto \sigma^{-1548} e^{-\sum_{i=1}^{1548} \frac{(y_i - \mu)^2}{2\sigma^2}} \quad (7)
 \end{aligned}$$

We again implicitly assume that the MSCI-Germany index returns are independently and identically distributed (IID), that is, each daily return is a realization from a normal distribution with the same mean and standard deviation.

In the case of the normal distribution, since the likelihood is a function of two arguments, we can visualize it with a three-dimensional surface as in Figure 2. It is also useful to plot the so-called contours of the likelihood, which we obtain by “slicing” the shape in Figure 2 horizontally at various levels of the likelihood.

Each contour corresponds to a pair of parameter values (and the respective likelihood value). In Figure 3, for example, we could observe that the pair $(\mu, \sigma) = (-0.23e - 3, 0.31e - 3)$, with a likelihood value of 0.6, is more likely than the pair $(\mu, \sigma) = (0.096e - 3, 0.33e - 3)$, with a likelihood value of 0.1, since the corresponding likelihood is larger.

BAYES' THEOREM

Bayes' theorem is the cornerstone of the Bayesian framework. Formally, it is a result from introductory probability theory, linking the unconditional distribution of a random variable with its conditional distribution. For Bayesian

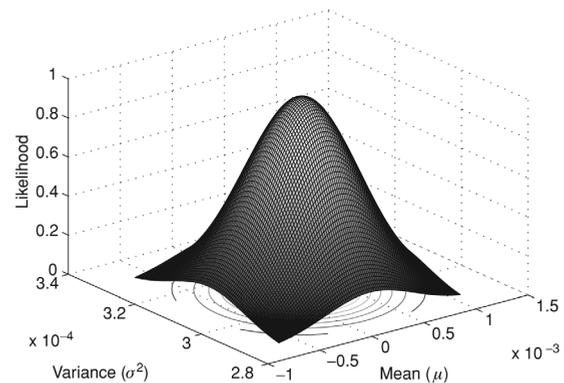


Figure 2 The Likelihood Function for the Parameters of the Normal Distribution

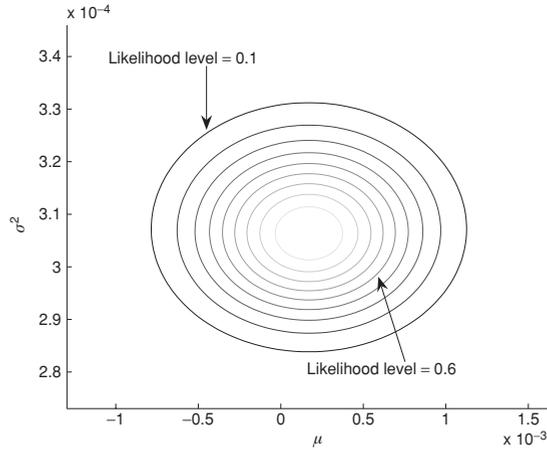


Figure 3 The Likelihood Function for the Parameters of the Normal Distribution: Contour Plot

proponents, it is the representation of the philosophical principle underlying the Bayesian framework that probability is a measure of the degree of belief one has about an uncertain event. Bayes' theorem is a rule that can be used to update the beliefs that one holds in light of new information (for example, observed data).

We first consider the discrete version of Bayes' theorem. Denote the evidence prior to observing the data by E and suppose that a researcher's belief in it can be expressed as the probability $P(E)$. The Bayes theorem tells us that, after observing the data, D , the belief in E is adjusted according to the following expression:

$$P(E | D) = \frac{P(D | E) \times P(E)}{P(D)} \quad (8)$$

where:

1. $P(D | E)$ is the conditional probability of the data given that the prior evidence, E , is true.
2. $P(D)$ is the unconditional (marginal) probability of the data, $P(D) > 0$; that is, the probability of D irrespective of E , also expressed as

$$P(D) = P(D | E) \times P(E) + P(D | E^c) \times P(E^c)$$

where the subscript c denotes a complementary event.³

The probability of E before seeing the data, $P(E)$, is called the *prior probability*, whereas the updated probability, $P(E | D)$, is called the *posterior probability*.⁴ Notice that the magnitude of the adjustment of the prior probability, $P(E)$, after observing the data is given by the ratio $P(D|E)/P(D)$. The conditional probability, $P(D|E)$, when considered as a function of E , is in fact the likelihood function, as will become clear further below.

As an illustration, consider a manager in an event-driven hedge fund. The manager is testing a strategy that involves identifying potential acquisition targets and examines the effectiveness of various company screens, in particular the ratio of stock price to free cash flow per share (PFCF). Let us define the following events:

D = Company X 's PFCF has been more than three times lower than the sector average for the past three years.

E = Company X becomes an acquisition target in the course of a given year.

Independently of the screen, the manager assesses the probability of company X being targeted at 40%. That is, denoting by E^c the event that X does not become a target in the course of the year, we have

$$P(E) = 0.4$$

and

$$P(E^c) = 0.6$$

Suppose further that the manager's analysis suggests that the probability a target company's PFCF has been more than three times lower than the sector average for the past three years is 75% while the probability that a nontarget company has been having that low of a PFCF for the past three years is 35%:

$$P(D | E) = 0.75$$

and

$$P(D | E^c) = 0.35$$

If a bidder does appear on the scene, what is the chance that the targeted company had been detected by the manager's screen? To answer this question, the manager needs to update the prior probability $P(E)$ and compute the posterior probability $P(E | D)$. Applying (8), we obtain

$$P(E | D) = \frac{0.75 \times 0.4}{0.75 \times 0.4 + 0.35 \times 0.6} \approx 0.59 \quad (9)$$

After taking into account the company's persistently low PFCF, the probability of a takeover increases from 40% to 59%.

In financial applications, the continuous version of the Bayes' theorem (as follows later) is predominantly used. Nevertheless, the discrete form has some important uses, two of which we briefly outline now.

Bayes' Theorem and Model Selection

The usual approach to modeling of a financial phenomenon is to specify the analytical and distributional properties of a process that one thinks generated the observed data and treat this process as if it were the true one. Clearly, in doing so, one introduces a certain amount of error into the estimation process. Accounting for model risk might be no less important than accounting for (within-model) parameter uncertainty, although it seems to preoccupy researchers less often.

One usually entertains a small number of models as plausible ones. The idea of applying the Bayes' theorem to model selection is to combine the information derived from the data with the prior beliefs one has about the degree of model validity. One can then select the single "best" model with the highest posterior probability and rely on the inference provided by it or one can weigh the inference of each model by its posterior probability and obtain an "averaged-out" conclusion.

Bayes' Theorem and Classification

Classification refers to assigning an object, based on its characteristics, into one out of several categories. It is most often applied in the area of credit and insurance risk, when a creditor (an insurer) attempts to determine the creditworthiness (riskiness) of a potential borrower (policyholder). Classification is a statistical problem because of the existence of information asymmetry—the creditor's (insurer's) aim is to determine with very high probability the unknown status of the borrower (policyholder). For example, suppose that a bank would like to rate a borrower into one of three categories: low risk (L), medium risk (M), and high risk (H). It collects data on the borrower's characteristics such as the current ratio, the debt-to-equity ratio, the interest coverage ratio, and the return on capital. Denote these observed data by the four-dimensional vector \mathbf{y} . The dynamics of \mathbf{y} depends on the borrower's category and is described by one of three (multivariate) distributions,

$$f(\mathbf{y} | C = L) \\ f(\mathbf{y} | C = M)$$

or

$$f(\mathbf{y} | C = H)$$

where C is a random variable describing the category. Let the bank's belief about the borrower's category be π_i , where

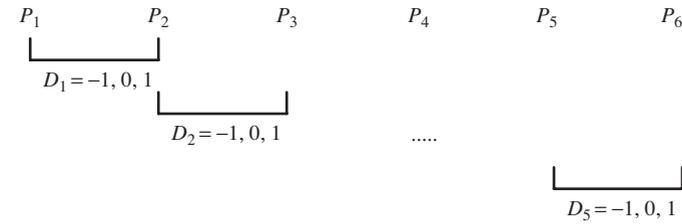
$$\pi_1 = \pi(C = L) \\ \pi_2 = \pi(C = M)$$

and

$$\pi_3 = \pi(C = H)$$

The discrete version of Bayes' theorem can be employed to evaluate the posterior (updated) probability, $\pi(C = i | \mathbf{y})$, $i = L, M, H$, that the borrower belongs to each of the three categories.

Let us now take our first steps in illustrating how Bayes' theorem helps in making inferences about an unknown distribution parameter.



where $D_i = -1$ if $P_{i+1} < P_i$
 $D_i = 0$ if $P_{i+1} = P_i$
 $D_i = 1$ if $P_{i+1} > P_i$

$$A_1 = D_1 + D_2$$

$$A_2 = D_2 + D_3$$

...

$$A_4 = D_4 + D_5$$

Note: X = number of occurrences of $A = 2$ within the sample period

Figure 4 The Number of Consecutive Trade-by-Trade Price Increases

Bayesian Inference for the Binomial Probability

Suppose we are interested in analyzing the dynamic properties of the intraday price changes for a stock. In particular, we want to evaluate the probability of consecutive trade-by-trade price increases. In an oversimplified scenario, this problem could be formalized as a binomial experiment.

The binomial experiment is a setting in which the source of randomness is a binary one (only takes on two alternative modes/states) and the probability of both states is constant throughout. The binomial random variable is the number of occurrences of the state of interest. In our illustration, the two states are “the consecutive trade-by-trade price change is an increase” and “the consecutive trade-by-trade price change is a decrease or null.” The random variable is the number of consecutive price increases. Denote it by X . Denote the probability of a consecutive increase by θ . Our goal is to draw a conclusion about the unknown probability, θ .

As an illustration, we consider the transaction data for the AT&T stock during the two-month period from January 4, 1993, through February

26, 1993 (a total of 55,668 price records). The diagram in Figure 4 shows how we define the binomial random variable given six price observations, P_1, \dots, P_6 . (Notice that the realizations of the random variable are one less than the number of price records.) A consecutive price increase is “encoded” as $A = 2$ and its probability is $\theta = P(A = 2)$; all other realizations of A ($A = -2, -1, 0$ or 1) have a probability of $1 - \theta$. We say that the number of consecutive price increases, X , is distributed as a binomial random variable with parameter θ . The probability mass function of X is represented by the expression

$$P(X = x | \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} \quad x = 0, 1, 2, \dots, n \quad (10)$$

where n is the sample size (the number of trade-by-trade price changes; a price change could be zero) and $\binom{n}{x} = \frac{n!}{x!(n-x)!}$. During the sample period, there are $X = 176$ trade-by-trade consecutive price increases. This information is embodied in the likelihood function for θ :

$$L(\theta | X = 176) = \theta^{176} (1 - \theta)^{55667-176} \quad (11)$$

We would like to combine that information with our prior belief about what the probability of a consecutive price increase is. We denote the prior distribution of an unknown parameter θ by $\pi(\theta)$, the posterior distribution of θ by $\pi(\theta | \text{data})$, and the likelihood function by $L(\theta | \text{data})$.

We consider two prior scenarios for the probability of consecutive price increases, θ :

1. We do not have any particular belief about the probability θ . Then, the prior distribution could be represented by a uniform distribution on the interval $[0, 1]$. Note that this prior assumption implies an expected value for θ of 0.5. The density function of θ is given by

$$\pi(\theta) = 1, \quad 0 \leq \theta \leq 1$$

2. Our intuition suggests that the probability of a consecutive price increase is around 2%. A possible choice of a prior distribution for θ is the beta distribution.⁵ The density function of θ is then written as

$$\pi(\theta | \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}, \quad 0 \leq \theta \leq 1 \tag{12}$$

where $\alpha > 0$ and $\beta > 0$ are the parameters of the beta distribution and $B(\alpha, \beta)$ is the so-called beta function. We set the parameters α and β to 1.6 and 78.4, respectively.

Figure 5 presents the plots of the two prior densities. Notice that under the uniform prior, all values of θ are equally likely, while under the beta prior, we assert higher prior probability for some values and lower prior probability for others.

Combining the sample information with the prior beliefs, we obtain θ 's posterior distribution. We rewrite Bayes' theorem with the notation in the current discussion:

$$p(\theta | x) = \frac{L(\theta | x)\pi(\theta)}{f(x)} \tag{13}$$

where $f(x)$ is the unconditional (marginal) distribution of the random variable X , given by

$$f(x) = \int L(\theta | x)\pi(x) d\theta \tag{14}$$

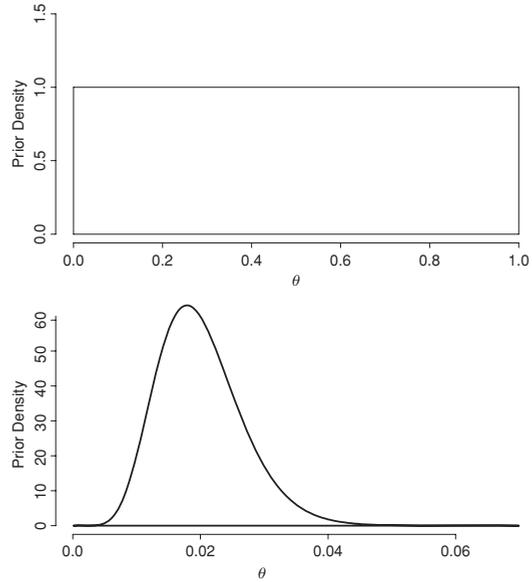


Figure 5 Density Curves of the Two Prior Distributions for the Binomial Parameter, θ
 Note: The density curve on top is the uniform density, while the one at the bottom is the beta density.

Since $f(x)$ is obtained by averaging over all possible values of θ , it does not depend on θ . Therefore, we can rewrite (8) as

$$\pi(\theta | x) \propto L(\theta | x)\pi(\theta) \tag{15}$$

The expression in (15) provides us with the posterior density of θ up to some unknown constant. However, in certain cases we would still be able to recognize the posterior distribution as a known distribution, as we see shortly.⁶ Since both assumed prior distributions of θ are continuous, the posterior density is also continuous and (13) and (15), in fact, represent the *continuous version of Bayes' theorem*.

Let us see what the posterior distribution for θ is under each of the two prior scenarios.

1. The posterior of θ under the uniform prior scenario is written as

$$\begin{aligned} \pi(\theta | x) &\propto L(\theta | x) \times 1 \\ &\propto \theta^{176} (1 - \theta)^{55667-176} \\ &= \theta^{177-1} (1 - \theta)^{55492-1} \end{aligned} \tag{16}$$

where the first α refers to omitting the marginal data distribution term in (14), while the second α refers to omitting the constant term from the likelihood function.

The expression $\theta^{177-1}(1-\theta)^{55492-1}$ above resembles the density function of the beta distribution in (12). The missing part is the term $B(177, 55492)$, which is a constant with respect to θ . We call $\theta^{\alpha-1}(1-\theta)^{\beta-1}$ the kernel of a beta distribution with parameters α and β . Obtaining it is sufficient to identify uniquely the posterior of θ as a beta distribution with parameters $\alpha = 177$ and $\beta = 55492$.

- The beta distribution is the conjugate prior distribution for the binomial parameter θ . This means that the posterior distribution of θ is also a beta distribution (of course, with updated parameters):

$$\begin{aligned}\pi(\theta | x) &\propto L(\theta | x)\pi(\theta) \\ &\propto \theta^{176}(1-\theta)^{55667-176}\theta^{1.6-1}(1-\theta)^{78.4-1} \\ &= \theta^{177.6-1}(1-\theta)^{55569.4-1}\end{aligned}\quad (17)$$

where again we omit any constants with respect to θ . As expected, we can recognize the expression in the last line above as the kernel of a beta distribution with parameters $\alpha = 177.6$ and $\beta = 55569.4$.

Finally, we might want to obtain a single number as an estimate of θ . In the classical (frequentist) setting, the usual estimator of θ is the maximum likelihood estimator (the value maximizing the likelihood function in (11)), which happens to be the sample proportion $\hat{\theta}$:

$$\hat{\theta} = \frac{176}{55667} = 0.00316 \quad (18)$$

or 0.316%.

In the Bayesian setting, one possible estimate of θ is the *posterior mean*, that is, the mean of θ 's posterior distribution. Since the mean of the beta distribution is given by $\alpha/(\alpha + \beta)$, the posterior mean of θ (the expected probability of consecutive trade-by-trade increase in the price of the AT&T stock) under the uniform prior

scenario is

$$\tilde{\theta}_U = \frac{177}{177 + 55492} = 0.00318$$

or 0.318%, while the posterior mean of θ under the beta prior scenario is

$$\tilde{\theta}_B = \frac{177.6}{177.6 + 55569.4} = 0.00319$$

or 0.319%.

The two posterior estimates and the maximum-likelihood estimate are the same for all practical purposes. The reason is that the sample size is so large that the information contained in the data sample “swamps out” the prior information.

KEY POINTS

- Statistical analysis is employed from the vantage point of either of the two main statistical philosophical traditions—frequentist and Bayesian.
- The frequentist interpretation of the probability of an event is that it is the limit of its long-run relative frequency (i.e., the frequency with which it occurs as the amount of data increases without bound).
- The Bayesian view of the world is based on the subjectivist interpretation of probability: Probability is subjective, a degree of belief that is updated as information or data are acquired.
- In the Bayesian framework, probability beliefs based on the existing knowledge base take the form of the prior probability; the posterior probability represents the updated beliefs.
- The likelihood function is a statistical construct summarizing the information contained in the sample of data.
- Bayes' theorem links the unconditional and unconditional probabilities. Under the Bayesian approach, prior beliefs are combined with sample information to create updated posterior beliefs.

- Two important applications of the discrete form of Bayes' theorem are model selection and classification.
- In financial applications, the continuous version of Bayes' theorem is predominantly used.

NOTES

1. In this example, we assume, perhaps unrealistically, that θ stays constant through time and that the annual number of defaults in a given year is independent from the number of defaults in any other year within the 20-year period. The independence assumption means that each observation of the number of annual defaults is regarded as a realization from a Poisson distribution with the same average rate of defaults, θ ; this allows us to represent the likelihood function as the product of the mass function at each observation.
2. One such result is the Central Limit Theorem which asserts that, under certain mild regularity conditions, sums of independent random variables are distributed with the normal distribution asymptotically (as the terms of the sum become indefinitely many).
3. The complement (complementary event) of E , E_c , includes all possible outcomes that could occur if E is not realized. The probabilities of an event and its complement always sum up to 1: $P(E) + P(E_c) = 1$.
4. The expression in (8) is easily generalized to the case when a researcher updates be-

liefs about one of many mutually exclusive events (such that two or more of them occur at the same time). Denote these events by E_1, E_2, \dots, E_K . The events are such that their probabilities sum up to 1: $P(E_1) + \dots + P(E_K) = 1$. Bayes' theorem then takes the form

$$P(E_k | D) = \frac{P(D | E_k) \times P(E_k)}{P(D | E_1) \times P(E_1) + P(D | E_2) \times P(E_2) + \dots + P(D | E_K) \times P(E_K)}$$

for $k = 1, \dots, K$ and $P(D) > 0$.

5. The beta distribution is the conjugate distribution for the parameter of the binomial distribution.
6. When the posterior distribution is not recognizable as a known distribution, inference about θ is accomplished with the help of numerical methods.

REFERENCES

- De Finetti, B. (1931). Probabilism: A critical essay on the theory of probability and on the value of science (translation of 1931 article). *Erkenntnis* 31, September 1989.
- Lindley, D. (1971) *Making Decisions*. New York: John Wiley & Sons.
- Markowitz, H. (1991). Autobiography. In Tore Frängsmyr (ed.), *Les Prix Nobel. The Nobel Prizes 1990*. Nobel Foundation.
- Rachev, S. T., Hsu, J. S. J., Bagasheva, B. S., and Fabozzi, F. J. (2008). *Bayesian Methods in Finance*. Hoboken, NJ: John Wiley & Sons.
- Ramsey, F. (1931). Truth and probability. Chapter VII in *The Foundations of Mathematics and Other Logical Essays*. New York: Harcourt, Brace.
- Savage, L. (1954). *The Foundations of Statistics*. New York: John Wiley & Sons.

Introduction to Bayesian Inference

BILIANA S. GÜNER, PhD

Assistant Professor of Statistics and Econometrics, Ozyegin University, Turkey

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

JOHN S. J. HSU, PhD

Professor of Statistics and Applied Probability, University of California, Santa Barbara

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Bayesian inference is the process of arriving at estimates of the model parameters reflecting the blending of information from different sources. Most commonly, two sources of information are considered: prior knowledge or beliefs and observed data. The discrepancy (or lack thereof) between them and their relative strength determines how far away the resulting Bayesian estimate is from the corresponding classical estimate. Along with the point estimate, which most often is the posterior mean, in the Bayesian setting one has available the whole posterior distribution, allowing for a richer analysis.

In this entry, we focus on the essentials of *Bayesian inference*. Formalizing the practitioner's knowledge and intuition into prior distributions is a key part of the inferential process. Especially when the data records are not abundant, the choice of prior distributions can influence greatly posterior conclusions. After presenting an overview of some approaches to prior specification, we focus on the elements of posterior analysis. Posterior and predictive results can be summarized in a few numbers, as in the classical statistical approach, but one

could also easily examine and draw conclusions about all other aspects of the posterior and predictive distributions of the (functions of the) parameters.

PRIOR INFORMATION

The prior distribution for the model parameters is an integral component of the Bayesian inference process. The updated (posterior) beliefs are the result of the trade-off between the prior

and data distributions. The continuous form of Bayes' theorem is:

$$p(\theta | \mathbf{y}) \propto L(\theta | \mathbf{y})\pi(\theta) \quad (1)$$

where

θ = unknown parameter whose inference we are interested in.

\mathbf{y} = a vector (or a matrix) of recorded observations.

$\pi(\theta)$ = prior distribution of θ depending on one or more parameters, called *hyperparameters*.

$L(\theta | \mathbf{y})$ = likelihood function for θ .

$p(\theta | \mathbf{y})$ = posterior (updated) distribution of θ .

Two factors determine the degree of posterior trade-off—the strength of the prior information and the amount of data available. Generally, unless the prior is very informative (in a sense that will become clear), the more observations, the greater the influence of the data on the posterior distribution. On the contrary, when very few data records are available, the prior distribution plays a predominant role in the updated beliefs.

How to translate the prior information about a parameter into the analytical (distributional) form, $\pi(\theta)$, and how sensitive the posterior inference is to the choice of prior have been questions of considerable interest in the Bayesian literature.¹ There is, unfortunately, no “best” way to specify the prior distribution and translating subjective views into prior values for the distribution parameters could be a difficult undertaking.

Before we review some commonly used approaches to prior elicitation, we make the following notational and conceptual note. It is often convenient to represent the posterior distribution, $p(\theta | \mathbf{y})$, in a logarithmic form. Then, it is easy to see that the expression in (1) is transformed according to

$$\log(p(\theta | \mathbf{y})) = \text{const} + \log(L(\theta | \mathbf{y})) + \log(\pi(\theta)),$$

where const is the logarithm of the constant of proportionality.

Informative Prior Elicitation

Prior beliefs are informative when they modify substantially the information contained in the data sample so that the conclusions we draw about the model parameters based on the posterior distribution and on the data distribution alone differ. The most commonly used approach to representing informative prior beliefs is to select a distribution for the unknown parameter and specify the hyperparameters so as to reflect these beliefs.

Informative Prior Elicitation for Location and Scale Parameters

Usually, when we think about the average value that a random variable takes, we have the typical value in mind. Therefore, we hold beliefs about the median of the distribution rather than its mean.² This distinction does not matter in the case of symmetric distributions, since then the mean and the median coincide. However, when the distribution we selected is not symmetric, care must be taken to ensure that the prior parameter values reflect our beliefs. Formulating beliefs about the spread of the distribution is less intuitive. The easiest way to do so is to ask ourselves questions such as, for instance: Which value of the random variable do a quarter of the observations fall below/above? Denoting the random variable by X , the answers to these questions give us the following probability statements:

$$P(X < x_{0.25}) = 0.25$$

and

$$P(X > x_{0.75}) = 0.25$$

where $x_{0.25}$ and $x_{0.75}$ are the values we have subjectively determined and are referred to as the first and third quartiles of the distribution, respectively. Other similar probability statements

can be formulated, depending on the prior beliefs.

As an example, suppose that we model the behavior of the monthly returns on some financial asset and the normal distribution, $N(\mu, \sigma^2)$ (along with the assumption that the returns are independently and identically distributed), describes their dynamics well. Assume for now that the variance is known, $\sigma^2 = \sigma^{2*}$, and thus we only need to specify a prior distribution for the unknown mean parameter, μ . We believe that a symmetric distribution is an appropriate choice and go for the simplicity of a normal prior:

$$\mu \sim N(\eta, \tau^2) \quad (2)$$

where η is the prior mean and τ^2 is the prior variance of μ ; to fully specify μ 's prior, we need to (subjectively) determine their values. We believe that the typical monthly return is around 1%, suggesting that the median of μ 's distribution is 1%. Therefore, we set η to 1%. Further, suppose we (subjectively) estimate that there is about a 25% chance that the average monthly return is less than 0.5% (i.e., $\mu_{0.25} = 0.5\%$). Then, using the tabulated cumulative probability values of the standard normal distribution, we find that the implied variance, τ^2 , is approximately equal to 0.74^2 .³ Our choice for the prior distribution of μ is thus $\pi(\mu) = N(1, 0.74^2)$.

Noninformative Prior Distributions

In many cases, our prior beliefs are vague and thus difficult to translate into an informative prior. We therefore want to reflect our uncertainty about the model parameter(s) without substantially influencing the posterior parameter inference. The so-called *noninformative priors*, also called vague or diffuse priors, are employed to that end.

Most often, the noninformative prior is chosen to be either a uniform (flat) density defined on the support of the parameter or the *Jeffreys' prior*.⁴ The noninformative distribution for a location parameter, μ , is given by a uniform

distribution on its support $((-\infty, \infty))$, that is,⁵

$$\pi(\mu) \propto 1 \quad (3)$$

The noninformative distribution for a scale parameter, σ (defined on the interval $(0, \infty)$) is⁶

$$\pi(\sigma) \propto \frac{1}{\sigma} \quad (4)$$

Notice that the prior densities in both (3) and (4) are not proper densities, in the sense that they do not integrate to one:

$$\int_{-\infty}^{\infty} 1 \, d\mu = \infty$$

and

$$\int_0^{\infty} \frac{1}{\sigma} \, d\sigma = \infty$$

Even though the resulting posterior densities are usually proper, care must be taken to ensure that this is indeed the case. To avoid impropriety of the posterior distributions, one could employ proper prior distributions but make them noninformative, as we discuss further on.

When one is interested in the joint posterior inferences for μ and σ , these two parameters are often assumed independent, giving the joint prior distribution

$$\pi(\mu, \sigma) \propto \frac{1}{\sigma} \quad (5)$$

The prior in (5) is often referred to as the *Jeffreys' prior*.⁷

Prior ignorance could also be represented by a (proper) standard distribution with a very large dispersion—the so-called flat or diffuse proper prior distribution. Let us turn again to the example for the monthly returns for some financial asset we considered earlier and suppose that we do not have particular prior information about the range of typical values the mean monthly return could take. To reflect this ignorance, we might center the normal distribution of μ around 0 (a neutral value, so to speak) and fix the standard deviation, τ , at a large value such as 10^6 , that is, $\pi(\mu) = N(0, (10^6)^2)$.

The prior of μ could take alternative distributional forms. For instance, a symmetric

Student's t -distribution could be asserted. A standard Student's t -distribution has a single parameter, the degrees of freedom, ν , which one can use to regulate the heaviness of the prior's tails—the lower ν is, the flatter the prior distribution. Asserting a scaled Student's t -distribution with a scale parameter, σ , provides additional flexibility in specifying the prior of μ .⁸ It can be argued that eliciting heavy-tailed prior distributions (with tails heavier than the tails of the data distribution) increases the posterior's robustness, that is, lowers the sensitivity of the posterior to the prior specification.

Conjugate Prior Distributions

In many situations, the choice of a prior distribution is governed by the desire to obtain analytically tractable and convenient posterior distribution. Thus, if one assumes that the data have been generated by a certain class of distributions, employing the class of the so-called "conjugate prior distributions" guarantees that the posterior distribution is of the same class as the prior distribution.⁹ Although the prior and posterior distributions have the same form, their parameters differ—the parameters of the posterior distribution reflects the trade-off between prior and sample information. We now consider the case of the normal data distribution, since it is central to our discussions of financial applications.

If the data, x , are assumed to come from a normal distribution, the conjugate priors for the normal mean, μ , and variance, σ^2 , are, respectively, a normal distribution and an inverted χ^2 distribution¹⁰

$$\pi(\mu | \sigma^2) = N\left(\eta, \frac{\sigma^2}{T}\right)$$

and

$$\pi(\sigma^2) = \text{Inv} - \chi^2(\nu_0, c_0^2) \quad (6)$$

where $\text{Inv} - \chi^2(\nu, c^2)$ denotes the inverted χ^2 distribution with ν_0 degrees of freedom and a scale parameter c_0^2 . The prior parameters (hyperparameters) that need to be (subjectively)

specified in advance are η , T , ν_0 , and c_0^2 . The parameter T plays the role of a discount factor, reflecting the degree of uncertainty about the distribution of μ . Usually, T is greater than one since one naturally holds less uncertainty about the distribution of the mean, μ , (with variance σ^2/T) than the data, x (with variance σ^2).

In various financial applications, the normal distribution is often not the most appropriate assumption for a data-generation process in view of various empirical features that financial data exhibit. Alternative distributional choices most often do not have corresponding conjugate priors and the resulting posterior distributions might not be recognizable as any known distributions. Then, numerical methods are applied to compute the posteriors.

In general, eliciting conjugate priors should be preceded by an analysis of whether prior beliefs would be adequately represented by them.

Empirical Bayesian Analysis

So far, we took care to emphasize the subjective manner in which prior information is translated into a prior distribution. This involves specifying the prior hyperparameters (if an informative prior is asserted) before observing/analyzing the set of data used for model evaluation. One approach for eliciting the hyperparameters parts with this tradition—the so-called "empirical Bayesian approach." In it, sample information is used to compute the values of the hyperparameters. Here we provide an example with the natural conjugate prior for a normal data distribution.

Denote the sample of n observations by $x = (x_1, x_2, \dots, x_n)$. It can be shown that the normal likelihood function can be expressed in the following way:

$$\begin{aligned} L(\mu, \sigma^2 | x) &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} (\nu s^2 + n(\mu - \hat{\mu})^2)\right) \end{aligned} \quad (7)$$

where

$$\begin{aligned}\hat{\mu} &= \frac{\sum_{i=1}^n x_i}{n}, \quad v = n - 1, \\ \text{and} \\ s^2 &= \frac{\sum_{i=1}^n (x_i - \hat{\mu})^2}{n - 1}\end{aligned}\quad (8)$$

The quantities $\hat{\mu}$ and s^2 are, respectively, the unbiased estimators of the mean, μ , and the variance, σ^2 , of the normal distribution.¹¹ It is now easy to see that the likelihood in (7) can be viewed as the product of two distributions—a normal distribution for μ conditional on σ^2 ,

$$\mu | \sigma \sim N\left(\hat{\mu}, \frac{\sigma^2}{n}\right)$$

and an inverted χ^2 distribution for σ^2 ,

$$\sigma^2 \sim \text{Inv} - \chi^2(v, s^2)$$

which become the prior distributions under the empirical Bayesian approach. We can observe that these two distributions are, of course, the same as the ones in (6). Their parameters are functions of the two sufficient statistics for the normal distribution, instead of subjectively elicited quantities. The sample size, n , above plays the role of the discount factor, T , in (6)—the more data available, the less uncertain one is about the prior distribution of μ (its prior variance decreases).

We now turn to a discussion of the fundamentals of posterior inference. Later in this entry, we provide an illustration of the effect various prior assumptions have on the posterior distribution.

POSTERIOR INFERENCE

The posterior distribution of a parameter (vector) θ given the observed data \mathbf{x} is denoted as $p(\theta | \mathbf{x})$ and obtained by applying the Bayes' theorem given by (1). Being a combination of the data and the prior, the posterior contains all relevant information about the unknown parameter θ .

Posterior Point Estimates

Although the benefit of being able to visualize the whole posterior distribution is unquestionable, it is often more practical to report several numerical characteristics describing the posterior, especially if reporting the results to an audience used to the classical (frequentist) statistical tradition. Commonly used for this purpose are the point estimates, such as the posterior mean, the posterior median, and the posterior standard deviation.¹² When the posterior is available in closed form, these numerical summaries can also be expressed in closed form. The posterior parameters in the natural conjugate prior scenario with a normal sampling density (see (6)) are also available analytically. The mean parameter, μ , of the normal distribution has a normal posterior, conditional on σ^2 ,

$$p(\mu | \mathbf{x}, \sigma^2) = N\left(\mu^*, \frac{\sigma^2}{T + n}\right) \quad (9)$$

The posterior mean and variance of μ are given, respectively, by

$$\begin{aligned}E(\mu | \mathbf{x}, \sigma^2) &\equiv \mu^* = \hat{\mu} \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{T}{\sigma^2}} + \eta \frac{\frac{T}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{T}{\sigma^2}} \\ &= \hat{\mu} \frac{n}{n + T} + \eta \frac{T}{n + T}\end{aligned}\quad (10)$$

where $\hat{\mu}$ is the sample mean as given in (8) and

$$\text{var}(\mu | \mathbf{x}, \sigma^2) = \frac{\sigma^2}{T + n} \quad (11)$$

In practical applications, usually the emphasis is placed on obtaining the posterior distribution of μ , not least because it is more difficult to formulate prior beliefs about the variance, σ^2 (let alone the whole covariance matrix in the multivariate setting). Often, then, the variance (covariance matrix) is estimated outside of the regression model and then fed into it, as if it were the “known” variance (covariance matrix).¹³ Nevertheless, for completeness, we provide σ^2 's posterior distribution—an inverted χ^2 ,

$$p(\sigma^2 | \mathbf{x}) = \text{Inv} - \chi^2(v^*, c^{2*}) \quad (12)$$

where

$$v^* = v_0 + n, \quad (13)$$

$$c^{2*} = \frac{1}{v^*} \left(v_0 c_0^2 + (n-1)s^2 + \frac{Tn}{T+n} (\hat{\mu} - \eta)^2 \right) \quad (14)$$

and s^2 is the unbiased sample estimator of the normal variance as given in (8). Using (13) and (14), one can now compute the posterior mean and variance of σ^2 as, respectively¹⁴

$$E(\sigma^2 | x) = \frac{v^*}{v^* - 2} c^{2*} \quad (15)$$

and

$$\text{var}(\sigma^2 | x) = \frac{2v^{*2}}{(v^* - 2)^2(v^* - 4)} (c^{2*})^2 \quad (16)$$

When the posterior is not of known form and is computed numerically (through simulations), so are the posterior point estimates, as well as the distributions of any functions of these estimates (see Chapter 4 in Rachev et al., 2008).

Bayesian Intervals

The point estimate for the center of the posterior distribution is not too informative if the posterior uncertainty is significant. To assess the degree of uncertainty, a posterior $(1 - \alpha)100\%$ interval $[a, b]$, called a *credible interval*, can be constructed. The probability that the unknown parameter, θ , falls between a and b is $(1 - \alpha)100\%$,

$$P(a < \theta < b | x) = \int_a^b p(\theta | x) d\theta = 1 - \alpha$$

For reasons of convenience, the interval bounds may be determined so that an equal probability, $\alpha/2$, is left in the tails of the posterior distribution. For example, a could be chosen to be the 0.25th quantile, while b —the 0.75th quantile. The interpretation of the credible interval is often mistakenly ascribed to the classical confidence interval. In the classical setting, $(1 - \alpha)100\%$ is a coverage probability—if ar-

bitrarily many repeated samples of data are recorded, $(1 - \alpha)100\%$ of the corresponding confidence intervals will contain θ —a much less intuitive interpretation.

The credible interval is computed either analytically, by finding the theoretical quantiles of the posterior distribution (when it is of known form), or numerically, by finding the empirical quantiles using the simulations of the posterior density (see Chapter 4 in Rachev et al., 2008).¹⁵

Bayesian Hypothesis Comparison

The title of this section¹⁶ abuses the usual terminology by intentionally using “comparison” instead of “testing” in order to stress that the Bayesian framework affords one more than the mere binary reject/do-not-reject decision of the classical hypothesis testing framework. In the classical setting, the probability of a hypothesis (null or alternative) is either 0 or 1 (since frequentist statistics considers parameters as fixed, although unknown, quantities).

In contrast, in the Bayesian setting (where parameters are treated as random variables), the probability of a hypothesis can be computed (and is different from 0 or 1, in general), allowing for a true hypothesis comparison.¹⁷

Suppose one wants to compare the null hypothesis

$$H_0 : \theta \text{ is in } \Theta_0$$

with the alternative hypothesis

$$H_1 : \theta \text{ is in } \Theta_1$$

where Θ_0 and Θ_1 are disjoint sets of possible values for the unknown parameter θ . As with point estimates and credible intervals, hypothesis comparison is entirely based on θ 's posterior distribution. We compute the posterior probabilities of the null and alternative hypotheses,

$$P(\theta \text{ is in } \Theta_0 | x) = \int_{\Theta_0} p(\theta | x) d\theta \quad (17)$$

and

$$P(\theta \text{ is in } \Theta_1 | x) = \int_{\Theta_1} p(\theta | x) d\theta \quad (18)$$

respectively. These posterior hypotheses probabilities naturally reflect both the prior beliefs and the data evidence about θ . An informed decision can now be made incorporating that knowledge. For example, the posterior probabilities could be employed in scenario-generation—a tool of great importance in risk analysis.

The Posterior Odds Ratio

Although the framework outlined in the previous section is generally sufficient to make an informed decision about the relevance of hypotheses, we briefly discuss a somewhat more formal approach for Bayesian hypothesis testing. That approach consists of summarizing the posterior relevance of the two hypotheses into a single number—the posterior odds ratio. The posterior odds ratio is the ratio of the weighted likelihoods for the model parameters under the null hypothesis and under the alternative hypothesis, multiplied by the prior odds. The weights are the prior parameter distributions (thus, parameter uncertainty is taken into account).¹⁸

Denote the a priori probability of the null hypothesis by α . Then, the prior odds are the ratio $\alpha/(1 - \alpha)$. The posterior odds, denoted by PO, are simply the prior odds updated with the information contained in the data and are given by

$$\text{PO} = \frac{\alpha}{1 - \alpha} \times \frac{\int L(\theta | x, H_0) \pi(\theta) d\theta}{\int L(\theta | x, H_1) \pi(\theta) d\theta} \quad (19)$$

where $L(\theta | x, H_0)$ is the likelihood function reflecting the restrictions imposed by the null hypothesis and $L(\theta | x, H_1)$ is the likelihood function under the alternative hypothesis.

When no prior evidence in favor or against the null hypothesis exists, the prior odds is usually set equal to one. A low value of the posterior odds generally indicates evidence against the null hypothesis.

BAYESIAN PREDICTIVE INFERENCE

After performing Bayesian posterior inference about the parameters of the data-generating process, one may use the process to predict the realizations of the random variable ahead in time. The purpose of such a prediction could be to test the predictive power of the model (for example, by analyzing a metric for the distance between the model's predictions and the actual realizations) as part of a backtesting procedure or to directly use it in the decision-making process.

As in the case of posterior inference, predictive inference provides more than simply a point prediction—one has available the whole predictive distribution (either analytically or numerically) and thus increased modeling flexibility.¹⁹ The density of the predictive distribution is the sampling (data) distribution weighted by the posterior parameter density. By averaging out the parameter uncertainty (contained in the posterior), the predictive distribution provides a superior description of the model's predictive ability. In contrast, the classical approach to prediction involves computing point predictions or prediction intervals by plugging in the parameter estimates into the sampling density, treating those estimates as if they were the true parameter values.

Denoting the sampling and the posterior density by $f(x | \theta)$ and $p(\theta | x)$, respectively, the predictive density one step ahead is given by²⁰

$$f(x_{+1} | x) = \int f(x_{+1} | \theta) p(\theta | x) d\theta \quad (20)$$

where x_{+1} denotes the one-step-ahead realization. Notice that since we integrate (average) over the values of θ , the predictive distribution is independent of θ and depends only on the past realizations of the random variable X —it describes the process we assume has generated the data. The predictive density could be used to obtain a point prediction (for example, the predictive mean) or an interval prediction (similar in spirit to the Bayesian interval

discussed above) or to perform a hypotheses comparison.

ILLUSTRATION: POSTERIOR TRADE-OFF AND THE NORMAL MEAN PARAMETER

Using an illustration, we show the effects prior distributions have on posterior inference. For simplicity, we look at the case of a normal data distribution with a known variance, $\sigma^2 = 1$. That is, we need to elicit a prior distribution of the mean parameter, μ , only. We investigate the following prior assumptions:

1. A noninformative, improper prior (Jeffreys' prior): $\pi(\mu) \propto 1$.
2. A noninformative, proper prior: $\pi(\mu) = N(\eta, \tau^2)$, where $\eta = 0$ and $\tau = 10^6$.
3. An informative conjugate prior with subjectively determined hyperparameters: $\pi(\mu) = N(\eta, \tau^2)$, where $\eta = 0.02$ and $\tau = 0.1$.

As mentioned earlier in the entry, the relative strengths of the prior and the sampling distribution determine the degree of trade-off of prior and data information in the posterior. When the amount of available data is large, the sampling distribution dominates the prior in the posterior inference. (In the limit, as the number of observations grows indefinitely, only the sampling distribution plays a role in determining posterior results.²¹) To illustrate this sample-size effect, we consider the following two samples of data:

1. The monthly return on the S&P 500 stock index for the period January 1999 through December 2005 (a total of 192 returns).
2. The monthly return on the S&P 500 stock index for the period January 2005 through December 2005 (a total of 12 returns).

Let us denote the return data by the $n \times 1$ vector $\mathbf{r} = (r_1, r_2, \dots, r_n)$, where $n = 192$ or

$n = 12$. We assume that the sampling (data) distribution is normal, $R \sim N(\mu, \sigma^2)$. Combining the normal likelihood and the noninformative improper prior, we obtain for the posterior distribution of μ

$$\begin{aligned} p(\mu | \mathbf{r}, \sigma^2 = 1) &\propto (2\pi)^{-n/2} \exp\left(-\frac{\sum_{i=1}^n (r_i - \mu)^2}{2}\right) \\ &\propto \exp\left(-\frac{n(\mu - \hat{\mu})^2}{2}\right) \end{aligned} \quad (21)$$

where $\hat{\mu}$ is the sample mean as given in (8). Therefore, the posterior of μ is a normal distribution with mean $\hat{\mu}$ and variance $1/n$. As expected, the data completely determine the posterior distributions for both data samples, since we assumed prior ignorance about μ .

When a normal prior for μ , $N(\eta, \tau^2)$, is asserted, the posterior can be shown to be normal as well. In the generic case, for an arbitrary data variance σ^2 , we have

$$\begin{aligned} p(\mu | \mathbf{r}, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^n (r_i - \mu)^2}{\sigma^2}\right) \\ &\quad \times (2\pi\tau^2)^{-1/2} \exp\left(-\frac{(\mu - \eta)^2}{2\tau^2}\right) \\ &\propto \exp\left(-\frac{(\mu - \mu^*)^2}{2\tau^{2*}}\right) \end{aligned} \quad (22)$$

where the posterior mean, μ^* , is

$$\mu^* = \hat{\mu} \frac{\frac{n}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} + \eta \frac{\frac{1}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \quad (23)$$

and the posterior variance, τ^{2*} , is

$$\tau^{2*} = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \quad (24)$$

Notice that the posterior mean is a weighted average of the sample mean, $\hat{\mu}$, and the prior mean, η . The quantities $1/\sigma^2$ and $1/\tau^2$ have self-explanatory names: *data precision* and *prior precision*, respectively. The higher the precision, the more concentrated the distribution around its mean value.²² Let us see how the information trade-off between the data and the prior is reflected in the values of the posterior parameters.

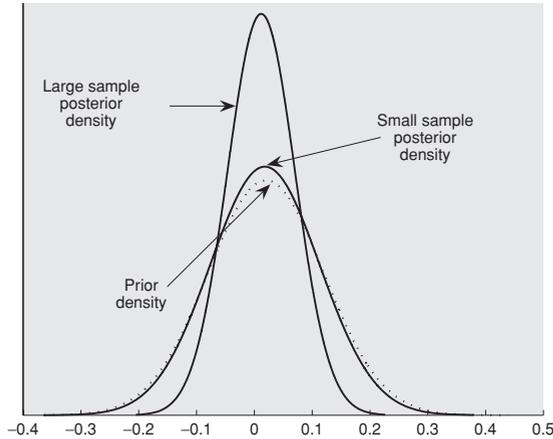


Figure 1 Sample Size and Posterior Trade-Off for the Normal Mean Parameter: The Case of Informative Prior

In the case of the noninformative, proper prior, $\tau = 10^6$. The rightmost term in (23) is then negligibly small and the posterior mean is very close to the sample mean: $\mu^* \approx \hat{\mu}$, while the posterior variance in (24) is approximately equal to $1/n$ (substituting in $\sigma^2 = 1$). That is, for both data samples, the noninformative proper prior produced posteriors almost the same as in the case of the noninformative improper prior, as expected.

Consider how the posterior is affected when informativeness of the prior is increased, as in the third prior scenario. Figure 1 helps visualize the posterior trade-off for the long and short data samples, respectively. The smaller the amount of observed data, the larger the influence of the prior on the posterior (the “closer” the posterior to the prior).

KEY POINTS

- The degree of posterior information trade-off has two determinants: strength of the prior information and amount of historical data available.
- Informative prior beliefs can modify substantially the information content of the observed data.

- Informative prior elicitation most commonly involves two steps: selecting the form of the prior distribution (usually, an analytically convenient one) and specifying its parameters (the hyperparameters) to reflect the prior beliefs.
- Noninformative priors help account for estimation uncertainty without substantially influencing the posterior parameter inference.
- A conjugate prior distribution guarantees that the resulting posterior distribution is of the same form as the prior.
- The posterior distribution can be summarized with point estimates, such as posterior mean, posterior median, posterior standard deviation, and posterior quantiles, as well as interval estimates.
- As in the case of posterior inference, when forecasting, one has available the whole predictive distribution of the random variable(s).

NOTES

1. See Chapter 3 in Berger (1985), Chapter 3 in Leonard and Hsu (1999), Berger (1990, 2006), and Garthwaite, Kadane, and O’Hagan (2005), among others.
2. The median is a measure of the center of a distribution alternative to the mean, defined as the value of the random variable, which divides the probability mass in halves. The median is the typical value the random variable takes. It is a more robust measure than the mean as it is not affected by the presence of extreme observations and, unless the distribution is symmetric, is not equal to the mean.
3. A random variable, $X \sim N(\mu, \sigma^2)$, is transformed into a standard normal random variable, $Z \sim N(0, 1)$, by subtracting the mean and dividing by its standard deviation:

$$Z = \frac{X - \mu}{\sigma}$$

4. Reference priors are another class of noninformative priors developed by Berger and

Bernardo (1992); see also Bernardo and Smith (1994). Their derivation is somewhat involved and applications in the field of finance are rare. One exception is Aguilar and West (2000).

5. Suppose a density has the form $f(x - \mu)$. The parameter μ is called the *location parameter* if it only appears within the expression $(x - \mu)$. The density, f , is then called a location density. For example, the normal density, $N(\mu, \sigma^{2*})$, is a location density when σ^{2*} is fixed.
6. Suppose a density has the form $\frac{1}{\sigma} f(\frac{x}{\sigma})$. The parameter σ is the *scale parameter*. For example, the normal density, $N(\mu^*, \sigma^2)$, is a scale density when the mean is fixed at some μ^* .
7. See Jeffreys (1961). In general, Jeffreys' prior of a parameter (vector), θ , is given by

$$\pi(\theta) = |I(\theta)|^{1/2}$$

where $I(\theta)$ is the so-called Fisher's information matrix for θ , given by

$$I(\theta) = -E \left(\frac{\partial^2 \log f(x|\theta)}{\partial \theta \partial \theta'} \right)$$

and the expectation is with respect to the random variable X , whose density function is $f(x|\theta)$. Notice that applying the expression for $\pi(\theta)$ to, for example, the normal distribution, one obtains the joint prior $\pi(\mu, \sigma) \propto 1/\sigma^2$, instead of the one in (5). Nevertheless, Jeffreys advocated the use of (5) since he assumed independence of the location and scale parameters.

8. The Student's t -distribution has heavier tails than the normal distribution. For values of ν less than 2, its variance is not defined.
9. Technically speaking, for the parameters of all distributions belonging to the exponential family there are conjugate prior distributions.
10. Notice that μ and σ^2 are not independent in (6). This prior scenario is the so-called natu-

ral conjugate prior scenario. Natural conjugate priors are priors whose functional form is the same as the likelihood's. The joint prior density of μ and σ^2 , $\pi(\mu, \sigma^2)$ can be represented as the product of a conditional and a marginal density: $\pi(\mu, \sigma^2) = \pi(\mu|\sigma^2)\pi(\sigma^2)$. If the dependence of the normal mean and variance is deemed inappropriate for the particular application, it is possible to make them independent and still benefit from the convenience of their functional forms—by eliciting a prior for μ as in (2).

11. An unbiased estimator of a parameter θ is a function of the data (a statistic), whose expected value is θ . The statistics $\hat{\mu}$ and s^2 are the so-called sufficient statistics for the normal distribution—knowing them is sufficient to uniquely determine the normal distribution that generated the data. In empirical Bayesian analysis, the hyperparameters are usually functions of the sufficient statistics of the sampling distribution.
12. In decision theory, loss functions are used to assess the impact of an action. In the context of parameter inference, if θ^* is the true parameter value, the loss associated with employing the estimate $\hat{\theta}$ instead of θ^* is represented by the loss function $L(\theta^*, \hat{\theta})$. One approach to estimating θ is to determine the value that minimizes the expected resulting loss. In Bayesian analysis, we minimize the expected posterior loss: its expectation is computed with respect to θ 's posterior distribution. It can be shown that the estimate of central tendency that minimizes the expected, posterior, squared-error loss function, $L(\theta^*, \hat{\theta}) = (\theta^* - \hat{\theta})^2$, is the posterior mean, while the estimate that minimizes the expected, posterior, absolute-error loss function, $L(\theta^*, \hat{\theta}) = |\theta^* - \hat{\theta}|$, is the posterior median.
13. One example for such an approach is the Black-Litterman model. See Black and Litterman (1991).

14. These are the expressions for expected value and variance of a random variable with the inverted χ^2 distribution.
15. A special type of Bayesian interval is the highest posterior density (HPD) interval. It is built so as to include the values of θ that have the highest posterior probability (the most likely values). When the posterior is symmetric and has a single peak (is unimodal), credible and HPD intervals coincide. With very skewed posterior distributions, however, the two intervals look very different. A disadvantage of HPD intervals is that they could be disjoint when the posterior has more than one peak (is multimodal). In unimodal settings, the Bayesian HPD interval obtained under the assumptions of a noninformative prior corresponds to the classical confidence interval.
16. In this section, we emphasize a practical approach to Bayesian hypothesis testing. For a rigorous description of Bayesian hypothesis testing, see, for example, Zellner (1971).
17. In the classical setting, the decision to reject or not the null hypothesis is made on the basis of the realization of a test statistic—a function of the data—whose distribution is known. The p -value of the hypothesis test is the probability of obtaining a value of the statistic as extreme or more extreme than the one observed. The p -value is compared to the test's significance level, which represents the predetermined probability of rejecting the null hypothesis falsely. If the p -value is sufficiently small (smaller than the significance level), the null hypothesis is rejected. The p -value is often mistakenly given the interpretation of a posterior probability of the null hypothesis. It has been suggested that a low p -value, interpreted by many as strong evidence against the null hypothesis, could be in fact quite a misleading signal about evidence strength. See, for example, Berger (1985) and Stambaugh (1999).
18. The posterior odds ratio bears similarity to the likelihood ratio which is at the center of most frequentist hypothesis tests. As its name suggests, the likelihood ratio is the ratio of the (maximized) likelihoods under the null and the alternative hypotheses.
19. The predictive density is usually of known (closed) form under conjugate prior assumptions.
20. Here, we assume that θ is continuous, which is the case in most financial applications.
21. This statement is valid only if one assumes that the data-generating process remains unchanged through time.
22. The posterior mean is an example for the shrinkage effect that combining prior and data information has.

REFERENCES

- Aguilar, O., and West, M. (2000). Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics* 18(3): 338–357.
- Berger, J. (1985). *Statistical Theory and Bayesian Analysis*. Berlin: Springer-Verlag.
- Berger, J. (1990). Robust Bayesian analysis: Sensitivity to the prior. *Journal of Statistical Planning and Inference* 25(3): 303–328.
- Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis* 1(3): 385–402.
- Berger, J., and Bernardo, J. (1992). On the development of reference priors. In J. Bernardo, J. Berger, A. David, and A. Smith (eds.), *Bayesian Statistics*, Vol. 4. Oxford: Clarendon Press, pp. 35–60.
- Bernardo, J. (2006). Noninformative priors do not exist: A discussion with José M. Bernardo. University of Valencia. Available at <http://www.uv.es/~bernardo/Dialogue.pdf>
- Bernardo, J., and Smith, A. (1994). *Bayesian Theory*. New York: John Wiley & Sons.
- Black F. and Litterman R. (1991), Asset allocation: combining investors views with market equilibrium, *Journal of Fixed Income*, September, pp. 7–18.
- Garthwaite, P., Kadane, J., and O'Hagan, A. (2005). Statistical methods for eliciting probability

- distributions. *Journal of the American Statistical Association* 100(470): 680–701.
- Jeffreys, H. (1961). *Theory of Probability*. New York: Oxford University Press.
- Leonard, T., and Hsu, J. (1999). *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*. Cambridge: Cambridge University Press.
- Rachev, S. T., Hsu, J. S. J., Bagasheva, B. S., and Fabozzi, F. J. (2008). *Bayesian Methods in Finance*. Hoboken, NJ: John Wiley & Sons.
- Stambaugh, R. (1999). Predictive regressions. *Journal of Financial Economics* 54: 375–421.
- Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley & Sons.

Bayesian Linear Regression Model

BILIANA S. GÜNER, PhD

Assistant Professor of Statistics and Econometrics, Ozyegin University, Turkey

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

JOHN S. J. HSU, PhD

Professor of Statistics and Applied Probability, University of California, Santa Barbara

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Linear regression is the “workhorse” of financial modeling. Cornerstone applications, such as asset pricing models, as well as time series models, are built around linear regression’s methods and tools. Casting the linear regression methodology in a Bayesian setting helps account for estimation uncertainty, allows for integration of prior information, and makes accessible the Bayesian numerical simulation framework.

In this entry, we lay the foundations of *Bayesian linear regression* estimation. We start with a univariate model with Gaussian innovations and consider two cases for prior distributional assumptions—diffuse and informative. Then, we show how one could incorporate knowledge that the sample is not homogeneous with respect to the variance, for example, due to a structural break. Finally, multivariate regression estimation is discussed.

THE UNIVARIATE LINEAR REGRESSION MODEL

The univariate linear regression model attempts to explain the variability in one variable (called the dependent variable) with the help of one or more other variables (called explanatory or independent variables) by asserting a linear relationship between them. We write the model as

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{K-1} X_{K-1} + \epsilon \quad (1)$$

where

- Y = dependent variable;
- X_k = independent (explanatory) variables,
 $k = 1, \dots, K - 1$;
- α = regression intercept;
- β_k = regression (slope) coefficients, $k =$
 $1, \dots, K - 1$, representing the effect a
unit change in X_k , $k = 1, \dots, K - 1$, has
on Y , keeping the remaining independ-
ent variables, X_j , $j \neq k$, fixed;
- ϵ = regression disturbance.

The regression disturbance is the source of randomness about the linear (deterministic) relationship between the dependent and independent variables. Whereas $\alpha + \beta_1 X_1 + \dots + \beta_{K-1} X_{K-1}$ represents the part of Y 's variability explained by X_k , $k = 1, \dots, K - 1$, ϵ represents the portion of Y 's variability left unexplained. It is usually assumed that the independent variables are fixed (nonstochastic).

Suppose that we have n observations of the dependent and the independent variables available. These data are then described by

$$\begin{aligned} y_i &= \alpha + \beta_1 x_{1,i} + \dots + \beta_{K-1} x_{K-1,i} + \epsilon_i \\ i &= 1, \dots, n \end{aligned} \quad (2)$$

The subscript i , $i = 1, \dots, n$, refers to the i th observation of the respective random variable. To describe the source of randomness, ϵ , one needs to make a distributional assumption about it. For simplicity, assume that ϵ_i , $i = 1, \dots, n$, are independently and identically distributed (IID) with the normal distribution and have zero means and (equal) variances, σ^2 . Then, the dependent variable, Y , has a normal distribution as well,

$$y_i \sim N(\mu_i, \sigma^2) \quad (3)$$

where $\mu_i = \alpha + \beta_1 x_{1,i} + \dots + \beta_{K-1} x_{K-1,i}$. Notice that the constant-variance assumption in (3) is quite restrictive. We come back to this issue later in the entry.

The expression in (2) is often written in the following compact form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (4)$$

where \mathbf{y} is a $n \times 1$ vector,

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$\boldsymbol{\beta}$ is a $(K) \times 1$ vector,

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_{K-1} \end{pmatrix}$$

\mathbf{X} is an $n \times (K)$ matrix whose first column consists of ones,

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{K-1,1} \\ 1 & x_{1,2} & \cdots & x_{K-1,2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,n} & \cdots & x_{K-1,n} \end{pmatrix}$$

and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector,

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

We write the normal distributional assumption for the regression disturbances in compact form as

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$$

where I_n is an $(n \times n)$ identity matrix. The parameters in (4) we need to estimate are $\boldsymbol{\beta}$ and σ^2 . Assuming normally distributed disturbances, we write the likelihood function for the model parameters as

$$\begin{aligned} L(\alpha, \beta_1, \dots, \beta_{K-1}, \sigma \mid \mathbf{y}, \mathbf{X}) \\ = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha \right. \\ \left. - \beta_1 x_{1,i} - \dots - \beta_{K-1} x_{K-1,i})^2 \right\} \end{aligned}$$

Or, in matrix notation, we have the likelihood function for the parameters of a multivariate normal distribution,

$$L(\beta, \sigma | \mathbf{y}, \mathbf{X}) = (2\pi\sigma^2)^{-n/2} \times \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)' (\mathbf{y} - \mathbf{X}\beta) \right\} \quad (5)$$

Bayesian Estimation of the Univariate Regression Model

In the classical setting, the regression parameters are usually estimated by maximizing the model's likelihood with respect to β and σ^2 , for instance, the likelihood in (5) if the normal distribution is assumed. When disturbances are assumed to be normally distributed, the maximum likelihood and the ordinary least squares (OLS) methods produce identical parameter estimates. It can be shown that the OLS estimator of the regression coefficients vector, β , is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (6)$$

where the prime symbol ($'$) denotes a matrix transpose.¹ The estimator of σ^2 is²

$$\hat{\sigma}^2 = \frac{1}{n-K} (\mathbf{y} - \mathbf{X}\hat{\beta})' (\mathbf{y} - \mathbf{X}\hat{\beta}) \quad (7)$$

To account for the parameters' estimation risk and to incorporate prior information, regression estimation can be cast in a Bayesian setting. We consider two prior scenarios—a *diffuse improper prior* and an *informative conjugate prior* for the regression parameter vector, (β, σ^2) .

Diffuse Improper Prior

The joint diffuse improper prior for β and σ^2 is given by

$$\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2} \quad (8)$$

where the regression coefficients can take any real value, $-\infty < \beta_k < \infty$, for $k = 1, \dots, K$,

and the disturbance variance is positive, $\sigma^2 > 0$.

Combining the likelihood in (5) and the prior above, we obtain the posteriors of the model parameters as follows:

- The posterior distribution of β conditional on σ^2 is (multivariate) normal:

$$p(\beta | \mathbf{y}, \mathbf{X}, \sigma^2) = N(\hat{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2) \quad (9)$$

where $\hat{\beta}$ is the OLS estimate in (6) and $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ is the covariance matrix of $\hat{\beta}$.

- The posterior distribution of σ^2 is inverted- χ^2 :

$$p(\sigma^2 | \mathbf{y}, \mathbf{X}) = \text{Inv-}\chi^2(n-K, \hat{\sigma}^2) \quad (10)$$

where $\hat{\sigma}^2$ is the estimator of σ^2 in (7).

It could be useful to obtain the marginal (unconditional) distribution of β in order to characterize it independently of σ^2 (as in practical applications, the variance is an unknown parameter).³ It can be shown, by integrating the joint posterior distribution

$$p(\beta, \sigma^2 | \mathbf{y}, \mathbf{X}) = p(\beta | \mathbf{y}, \mathbf{X}, \sigma^2) p(\sigma^2 | \mathbf{y}, \mathbf{X})$$

with respect to σ^2 , that β 's unconditional posterior distribution is a multivariate Student's t distribution with a kernel given by⁴

$$p(\beta | \mathbf{y}, \mathbf{X}) \propto \left((n-K) + (\beta - \hat{\beta})' \frac{\mathbf{X}'\mathbf{X}}{\hat{\sigma}^2} (\beta - \hat{\beta}) \right)^{-n/2} \quad (11)$$

Notice that integrating σ^2 out makes β 's distribution more heavy-tailed, duly reflecting the uncertainty about σ^2 's true value. Although β 's mean vector is unchanged, its variance increased (on average) by the term $\nu/(\nu-2)$:

$$\Sigma_{\beta} = \hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1} \frac{\nu}{\nu-2}$$

where $\nu = n - K$ is the degrees of freedom parameter of the multivariate Student's t distribution above.

In conclusion of our discussion of the posteriors in the diffuse improper prior scenario, suppose we are interested particularly in one

of the regression coefficients, say β_k . For example, β_k could be the return on a factor (size, value, momentum, etc.) in a multifactor model of stock returns. It can be shown that the standardized β_k has a Student's t distribution with $n - K$ degrees of freedom as its marginal posterior distribution,

$$\frac{\beta_k - \widehat{\beta}_k}{(h_{k,k})^{1/2}} \mid \mathbf{y}, \mathbf{X} \sim t_{n-K} \quad (12)$$

where $h_{k,k}$ is the k th diagonal element of $\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ and $\widehat{\beta}_k$ is the OLS estimate of β_k (the corresponding component of $\widehat{\beta}$). Bayesian intervals for β_k can then be constructed analytically.

Informative Prior

Under the normality assumption for the regression errors in (4), one can make use of the natural conjugate framework to reflect the existing prior knowledge and to obtain convenient analytical posterior results. Thus, let us assume that the regression coefficients vector, β , has a normal prior distribution (conditional on σ^2) and σ^2 —an inverted- χ^2 prior distribution:

$$\beta \mid \sigma \sim N(\beta_0, \sigma^2 \mathbf{A}) \quad (13)$$

and

$$\sigma^2 \sim \text{Inv-}\chi^2(v_0, c_0^2) \quad (14)$$

Four parameters have to be determined a priori: β_0 , \mathbf{A} , v_0 , and c_0^2 . The scale matrix \mathbf{A} is often chosen to be $\tau^{-1}(\mathbf{X}'\mathbf{X})^{-1}$ in order to obtain a prior covariance the same as the covariance matrix of the OLS estimator of β up to a scaling constant. Varying the (scale) parameter, τ , allows one to adjust the degree of confidence one has that β 's mean is β_0 —the smaller the value of τ , the greater the degree of uncertainty about β .

The easiest way to assert the prior mean, β_0 , is to fix it at some default value (such as $\mathbf{0}$, depending on the estimation context), unless more specific prior information is available, or to set it equal to the OLS estimate, $\widehat{\beta}$, obtained from

running the regression (4) on a prior sample of data.⁵

The parameters of the inverted- χ^2 distribution could be asserted using a prior sample of data as follows:

$$\begin{aligned} v_0 &= n_0 - K \\ c_0^2 &= \frac{1}{v_0} (\mathbf{y}_0 - \mathbf{X}_0 \widehat{\beta}_0)' (\mathbf{y}_0 - \mathbf{X}_0 \widehat{\beta}_0) \end{aligned}$$

where the subscript, 0, refers to the prior data sample. If no prior data sample is available, the inverted- χ^2 hyperparameters could be specified indirectly, by expressing beliefs about the prior mean and variance of σ^2 .⁶

The posterior distributions for the model parameters, β and σ^2 have the same form as the prior distributions, however, their parameters are updated to reflect the data information, along with the prior beliefs.

- The posterior for β is

$$p(\beta \mid \mathbf{y}, \mathbf{X}, \sigma^2) = N(\beta^*, \Sigma_\beta) \quad (15)$$

where the posterior mean and covariance matrix of β are given by

$$\beta^* = (\mathbf{A}^{-1} + \mathbf{X}'\mathbf{X})^{-1} (\mathbf{A}^{-1}\beta_0 + \mathbf{X}'\mathbf{X}\widehat{\beta}) \quad (16)$$

and

$$\Sigma_\beta = \sigma^2 (\mathbf{A}^{-1} + \mathbf{X}'\mathbf{X})^{-1} \quad (17)$$

We can observe that the posterior mean is a weighted average of the prior mean and the OLS estimator of β , as noted earlier in the entry as well.⁷

- The inverted- χ^2 posterior distribution of σ^2 is

$$p(\sigma^2 \mid \mathbf{y}, \mathbf{X}) = \text{Inv-}\chi^2(v^*, c^{2*}) \quad (18)$$

The parameters of σ^2 's posterior distribution are given by

$$v^* = v_0 + n \quad (19)$$

and

$$v^* c^{2*} = (n - K) \widehat{\sigma}^2 + (\beta_0 - \widehat{\beta})' \mathbf{H} (\beta_0 - \widehat{\beta}) + v_0 c_0^2 \quad (20)$$

where $\mathbf{H} = ((\mathbf{X}'\mathbf{X})^{-1} + \mathbf{A})^{-1}$

As earlier, we can derive the marginal posterior distribution of β by integrating σ^2 out of the joint posterior distribution. We obtain again a multivariate Student's t distribution, $t(v^*, \beta^*, Q)$,

$$p(\beta | \mathbf{y}, \mathbf{X}) \propto (v^* + (\beta - \beta^*)' \mathbf{Q}(\beta - \beta^*))^{-v^*/2} \quad (21)$$

where $\mathbf{Q} = (\mathbf{A}^{-1} + \mathbf{X}'\mathbf{X})/c^{2*}$

The mean of β remains the same, β^* (as it is independent of σ^2), while its unconditional (with respect to σ^2) covariance matrix is equal to $Q^{-1}v^*/(v^* - 2)$. The marginal posterior distribution for a single regression coefficient, β_k , can be shown to be

$$\frac{\beta_k - \beta_k^*}{(q_{k,k})^{1/2}} | \mathbf{y}, \mathbf{X} \sim t_{v_0+n-K} \quad (22)$$

where $q_{k,k}$ is the k th diagonal element of \mathbf{Q}^{-1} and β_k^* is the k th component of β^* .

Prediction

Suppose that we would like to predict the dependent variable, Y , p steps ahead in time and denote by the $p \times 1$ vector $\tilde{\mathbf{y}} = (y_{T+1}, y_{T+2}, \dots, y_{T+p})$ these future observations. We assume that the future observations of the independent variables are known and given by $\tilde{\mathbf{X}}$. The predictive density in the linear regression context can be expressed as,⁸

$$p(\tilde{\mathbf{y}} | \mathbf{y}, \tilde{\mathbf{X}}, \mathbf{X}) = \iint p(\tilde{\mathbf{y}} | \beta, \sigma^2, \tilde{\mathbf{X}}) \times p(\beta, \sigma^2 | \mathbf{y}, \mathbf{X}) d\beta, \sigma^2 \quad (23)$$

where $p(\beta, \sigma^2 | \mathbf{y}, \mathbf{X})$ is the joint posterior distribution of β and σ^2 .

It can be shown that the predictive distribution is multivariate Student's t . Under the diffuse improper prior scenario, the predictive distribution is

$$p(\tilde{\mathbf{y}} | \mathbf{y}, \tilde{\mathbf{X}}, \mathbf{X}) = t(n - K, \tilde{\mathbf{X}}\hat{\beta}, \mathbf{S}) \quad (24)$$

where $\mathbf{S} = \hat{\sigma}^2(I_p + \tilde{\mathbf{X}}(\mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{X}}')$ and $\hat{\beta}$ is the posterior mean of β under the diffuse improper scenario. In the case of the informative prior,

the predictive distribution of $\tilde{\mathbf{y}}$ is

$$p(\tilde{\mathbf{y}} | \mathbf{y}, \tilde{\mathbf{X}}, \mathbf{X}) = t(v_0 + n, \tilde{\mathbf{X}}\beta^*, \mathbf{V}) \quad (25)$$

where $\mathbf{V} = c^{2*}(I_p + \tilde{\mathbf{X}}(\mathbf{A}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\tilde{\mathbf{X}}')$ and β^* is the posterior mean of β in (16).

Certainly, it is again possible to derive the distribution for the predictive distribution for a single component of $\tilde{\mathbf{y}}$ —a univariate Student's t distribution—in the two scenarios, respectively,

$$\frac{\tilde{y}_k - \tilde{\mathbf{X}}^k\hat{\beta}_k}{s_{k,k}^{1/2}} \sim t_{n-K} \quad (26)$$

where $\tilde{\mathbf{X}}^k$ is the k th row of $\tilde{\mathbf{X}}$ (the observations of the independent variables pertaining to the k th future period), and $s_{k,k}$ is the k th diagonal element of the scale matrix, \mathbf{S} , in (24), and

$$\frac{\tilde{y}_k - \tilde{\mathbf{X}}^k\beta_k^*}{v_{k,k}^{1/2}} \sim t_{v_0+n-K} \quad (27)$$

where $v_{k,k}$ is the k th diagonal element of the scale matrix, \mathbf{V} , in (25).

The Case of Unequal Variances

We mentioned earlier in the entry that the equal-variance assumption in (3) might be somewhat restrictive. Two examples would help clarify what that means. First, suppose that the n observations of Y are collected through time. It is a common practice in statistical estimation to use the longest available data record, likely spanning many years. Changes in the underlying economic or financial paradigms, the way data are recorded, and so on, that might have occurred during the sample period might have caused the variance of the random variable (as well as its mean, for that matter) to shift.⁹ The equal-variance assumption would then lead to variance overestimation in the low-variance period(s) and variance underestimation in the high-variance period(s). When the variance (and/or mean) shifts permanently, the so-called “structural-break” models can be employed to reflect it.¹⁰

Second, if our estimation problem is based on observations recorded at a particular point in time (producing a cross-sectional sample), the equal-variance assumption might be violated again. All units in our sample could potentially have different variances, so that $\text{var}(y_i) = \sigma_i^2$, instead of $\text{var}(y_i) = \sigma^2$ as in (3), for $i = 1, \dots, n$. Estimation would then be severely hampered because this would imply a greater number of unknown parameters (variances and regression coefficients) than available data points.

In practice one would perhaps be able to identify groups of homogeneous sample units that can be assumed to have equal variances. Suppose, for instance, that the cross-sectional sample consists of small-cap and large-cap stock returns. One could then expect that the return variances (volatilities) across the two groups differ but assume that companies within each group have equal return volatilities. More generally, one could assume some form of functional relation among the unknown variances—this would serve to reduce the number of unknown parameters to estimate. We now provide one possible way to address the variance inequality in the case when the sample observations can be divided into two homogeneous (with respect to their variances) groups or when a structural break (whose timing we know) is present in the sample.¹¹

Denote the observations from the two groups by $\mathbf{y}_1 = (y_{1,1}, y_{1,2}, \dots, y_{1,n_1})$ and $\mathbf{y}_2 = (y_{2,1}, y_{2,2}, \dots, y_{2,n_2})$, so that $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$ and $n_1 + n_2 = n$. The univariate regression setup in (1) is modified as

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{X}_1\beta + \epsilon_1 \\ \mathbf{y}_2 &= \mathbf{X}_2\beta + \epsilon_2 \end{aligned} \quad (28)$$

where \mathbf{X}_1 and \mathbf{X}_2 are, respectively, $(n_1 \times K)$ and $(n_2 \times K)$ matrices of observations of the independent variables. The disturbances are assumed to be independent and distributed as

$$\begin{aligned} \epsilon_1 &\sim N(\mathbf{0}, \sigma_1^2 I_{n_1}) \\ \epsilon_2 &\sim N(\mathbf{0}, \sigma_2^2 I_{n_2}) \end{aligned} \quad (29)$$

where $\sigma_1^2 \neq \sigma_2^2$. The likelihood function for the model parameters, β , σ_1^2 , and σ_2^2 is given by

$$\begin{aligned} L(\beta, \sigma_1^2, \sigma_2^2 | \mathbf{y}, \mathbf{X}_1, \mathbf{X}_2) &\propto (\sigma_1^2)^{-\frac{n_1}{2}} (\sigma_2^2)^{-\frac{n_2}{2}} \\ &\times \exp\left(-\frac{1}{2\sigma_1^2}(\mathbf{y}_1 - \mathbf{X}_1\beta)'(\mathbf{y}_1 - \mathbf{X}_1\beta) \right. \\ &\quad \left. - \frac{1}{2\sigma_2^2}(\mathbf{y}_2 - \mathbf{X}_2\beta)'(\mathbf{y}_2 - \mathbf{X}_2\beta)\right) \end{aligned} \quad (30)$$

A noninformative diffuse prior can be asserted, as in (3.5), by assuming that the parameters are independent. The prior is written, then, as

$$\pi(\beta, \sigma_1, \sigma_2) \propto \frac{1}{\sigma_1\sigma_2}$$

It is straightforward to write out the joint posterior density of β , σ_1^2 , and σ_2^2 , which can be integrated with respect to the two variances to obtain the marginal posterior distribution of the regression coefficients vector. Zellner (1971) shows that the marginal posterior of β is the product of two multivariate Student's t densities.

$$p(\beta | \mathbf{y}, \mathbf{X}_1, \mathbf{X}_2) \propto t(v_1, \widehat{\beta}_1, S_1) \times t(v_2, \widehat{\beta}_2, S_2)$$

where, for $i = 1, 2$, $\widehat{\beta}_i$ is the OLS estimator of β in the two expressions in (28) viewed as separate regressions,

$$v_i = n_i - K, \quad S_i = \widehat{s}_i^2(\mathbf{X}_i' \mathbf{X}_i)$$

and

$$\widehat{s}_i^2 = \frac{1}{n_i - K} (\mathbf{y}_i - \mathbf{X}_i \widehat{\beta}_i)' (\mathbf{y}_i - \mathbf{X}_i \widehat{\beta}_i).$$

Zellner shows that the marginal posterior of β above can be approximated with a normal distribution (through a series of asymptotic expansions).

Illustration: The Univariate Linear Regression Model

We now provide an example to illustrate the posterior and predictive inference in a univariate linear regression model. We restrict our attention to the diffuse noninformative prior

and the informative prior discussed above, in order to take advantage of their analytical convenience.¹²

Our data consist of the monthly returns on 25 portfolios; the companies in each portfolio are ranked according to market capitalization and book-to-market (BM) ratios. The returns we use for model estimation span the period from January 1995 to December 2005 (a total of 132 time periods). We extract the factors that best explain the variability of returns of the 25 portfolios using principal components analysis. The first five factors explain around 95% of the variability and we use their returns as the independent variables in our linear regression model, making up the matrix X (the first column is a column of ones). The return on the portfolio consisting of the companies with the smallest size and BM ratios is the dependent variable, y . In addition, returns recorded for the months from January 1990 to December 1994 (a total

of 60 time periods) are employed to compute the hyperparameters of the informative prior distributions, in the manner explained in the previous section. Our interest centers primarily on the posterior inference for the regression coefficients, $\beta_k, k = 1, \dots, 6$ —the intercept and the five factor exposures (in the terminology of multifactor models).

Posterior Distributions

The prior and posterior parameter values for β are given in Table 1. Part A of the table presents the results under the diffuse improper prior assumption and Part B under the informative prior assumption. In parentheses are the posterior standard deviations of the regression coefficients.¹³ The OLS estimates of the regression coefficients are given by the posterior means in the diffuse prior scenario. Notice how the posterior mean of β under the informative prior is shrunk

Table 1 Posterior Inference for β

		β_1	β_2	β_3	β_4	β_5	β_6
		Intercept	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
A.	Prior Mean	-	-	-	-	-	-
	Posterior Mean	0.0048	-0.3108	-0.3997	0.0648	-0.4132	-0.0042
	Posterior Standard Deviation	(0.0011)	(0.0048)	(0.0103)	(0.0202)	(0.0297)	(0.0410)
	$b_{0.01}$	0.0021	-0.3219	-0.4238	0.0174	-0.4826	-0.1000
	$b_{0.05}$	0.0029	-0.3187	-0.4168	0.0312	-0.4624	-0.0721
	$b_{0.25}$	0.0040	-0.314	-0.4067	0.0511	-0.4333	-0.0319
	$b_{0.75}$	0.0055	-0.3075	-0.3928	0.0784	-0.3931	0.0235
	$b_{0.95}$	0.0067	-0.3029	-0.3827	0.0983	-0.364	0.0636
	$b_{0.99}$	0.0075	-0.2996	-0.3757	0.1121	-0.3438	0.0915
	B.	Prior Mean	0.0037	-0.2952	-0.4217	0.038	-0.2784
Posterior Mean		0.0042	-0.303	-0.4107	0.0514	-0.3458	0.0510
Posterior Standard Deviation		(0.0008)	(0.0033)	(0.0072)	(0.0142)	(0.0208)	(0.0287)
$b_{0.01}$		0.0024	-0.3108	-0.4276	0.0182	-0.3945	-0.0162
$b_{0.05}$		0.0029	-0.3085	-0.4226	0.0280	-0.3801	0.0038
$b_{0.25}$		0.0037	-0.3052	-0.4156	0.0418	-0.3598	0.0318
$b_{0.75}$		0.0048	-0.3007	-0.4059	0.0609	-0.3318	0.0703
$b_{0.95}$		0.0056	-0.2975	-0.3986	0.0747	-0.3115	0.0983
$b_{0.99}$		0.0061	-0.2952	-0.3939	0.0844	-0.2972	0.1180

Notes: Part A contains posterior results under the diffuse improper prior; Part B contains posterior results under the informative prior.

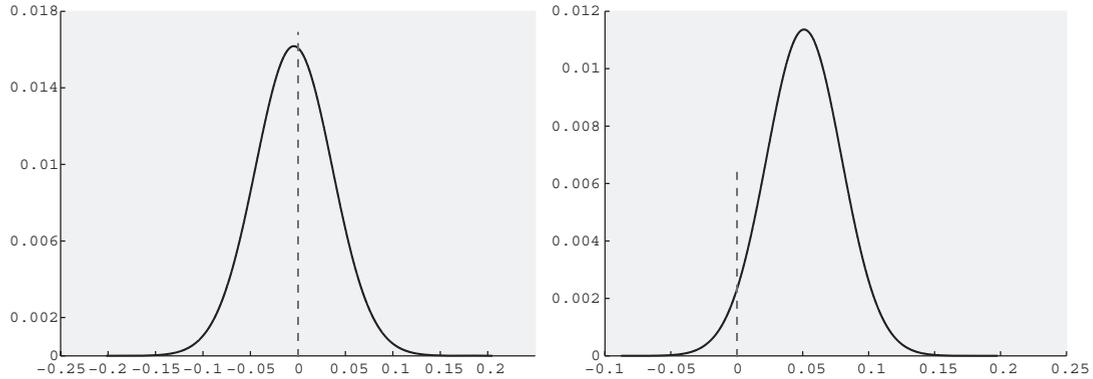


Figure 1 Posterior Densities of β_6 under the Two Prior Scenarios

Notes: The plot on the left refers to the diffuse improper prior; the plot on the right—to the informative prior.

away from the the OLS estimate and towards the prior value, for the chosen value of $\tau = 1$. We could introduce more uncertainty into the prior distribution of β (make it less informative) by choosing a smaller value of τ —the posterior mean of β would then be closer to the OLS estimate. Conversely, the stronger our prior belief about the mean of β , the closer the posterior mean would be to the prior mean.

Credible Intervals

Since the marginal posterior distribution of β_k , $k = 1, \dots, 6$, is of known form (Student's t), we can compute analytically the Bayesian confidence intervals for the regression coefficients. We provide the values of several quantiles of the posterior distribution of each β_k . For example, under the diffuse improper prior, the 95% (symmetric) Bayesian interval for β_2 is $(-0.3187, -0.3029)$, while, under the informative prior, the 99% (symmetric) Bayesian interval for β_6 is $(-0.0162, 0.1180)$.¹⁴

Hypothesis Comparison

In the frequentist regression tradition, testing the significance of the regression coefficients is of great interest—the validity of the null hypothesis $\beta_k = 0$ is examined. In the Bayesian

setting, we could evaluate and compare the posterior probabilities, $P(\beta_k > 0 | \mathbf{y}, \mathbf{X})$ and $P(\beta_k < 0 | \mathbf{y}, \mathbf{X})$ (given in Table 1 for each factor exposure). We could safely conclude that the exposures on Factor 1 through Factor 4 are different from zero—the mass of their posterior distributions is concentrated on either positive or negative values. For the exposure on Factor 5, the picture is less than clear-cut. Under the diffuse, improper prior, a bit over 50% of the posterior mass is below zero and the rest—above zero. Therefore, one would perhaps take the pertinence of this factor for explaining the variability of the return on the small-cap/small-BM portfolio with a grain of salt. Notice, however, how the situation changes in the informative-prior case. More than 95% of the posterior mass is above zero. The strong prior beliefs about a positive mean of β_6 lead to the conclusion that the exposure of the portfolio returns to Factor 5 is not zero. Figure 1 further illustrates these observations.

THE MULTIVARIATE LINEAR REGRESSION MODEL

Quite often in finance, and especially in investment management, one is faced with modeling data consisting of many assets whose returns

or other attributes are not independent. Casting the problem in a multivariate framework is one way to tackle dependencies between assets.¹⁵ In this section, we outline the basics of multivariate regression estimation within the Bayesian setting.¹⁶

Suppose that T observations are available on N dependent variables. We arrange these in the $T \times N$ matrix, \mathbf{y} ,

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_t \\ \vdots \\ \mathbf{y}_T \end{pmatrix} = \begin{pmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,N} \\ \cdots & \cdots & \cdots & \cdots \\ y_{t,1} & y_{t,2} & \cdots & y_{t,N} \\ \cdots & \cdots & \cdots & \cdots \\ y_{T,1} & y_{T,2} & \cdots & y_{T,N} \end{pmatrix}$$

The multivariate linear regression is written as

$$\mathbf{y} = \mathbf{X}\mathbf{B} + \mathbf{U} \quad (31)$$

where

$\mathbf{X} = T \times K$ matrix of observations of the K independent variables,

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_t \\ \vdots \\ \mathbf{x}_T \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,K} \\ \cdots & \cdots & \cdots & \cdots \\ x_{t,1} & x_{t,2} & \cdots & x_{t,K} \\ \cdots & \cdots & \cdots & \cdots \\ x_{T,1} & x_{T,2} & \cdots & x_{T,K} \end{pmatrix}$$

$\mathbf{B} = K \times N$ matrix of regression coefficients,

$$\mathbf{B} = \begin{pmatrix} \alpha \\ \beta_1 \\ \cdots \\ \beta_K \end{pmatrix} = \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_N \\ \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,N} \\ \cdots & \cdots & \cdots & \cdots \\ \beta_{K,1} & \beta_{K,2} & \cdots & \beta_{K,N} \end{pmatrix}$$

$\mathbf{U} = T \times N$ matrix of regression disturbances,

$$\mathbf{U} = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_t \\ \vdots \\ \mathbf{u}_T \end{pmatrix} = \begin{pmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,N} \\ \cdots & \cdots & \cdots & \cdots \\ u_{t,1} & u_{t,2} & \cdots & u_{t,N} \\ \cdots & \cdots & \cdots & \cdots \\ u_{T,1} & u_{T,2} & \cdots & u_{T,N} \end{pmatrix}$$

The first column of \mathbf{X} usually consists of ones to reflect the presence of an intercept. In the multivariate setting, the usual linear regression assumption that the disturbances are IID

means that each row of \mathbf{U} is an independent realization from the same N -dimensional multivariate distribution. We assume that this distribution is multivariate normal with zero mean and covariance matrix, Σ ,

$$\mathbf{u}_t \sim N(\mathbf{0}, \Sigma) \quad (32)$$

for $t = 1, \dots, T$. The off-diagonal elements of Σ are nonzero, as we assume the dependent variables are correlated, and the covariance matrix contains N variances and $N(N-1)/2$ distinct covariances.

Using the expression for the density of the multivariate normal distribution, we write the likelihood function for the unknown model parameters, \mathbf{B} and Σ , as¹⁷

$$L(\mathbf{B}, \Sigma | \mathbf{Y}, \mathbf{X}) \propto |\Sigma|^{-T/2} \times \exp\left(-\frac{1}{2} \sum_{t=1}^T (\mathbf{y}_t - \mathbf{x}_t \mathbf{B}) \Sigma^{-1} (\mathbf{y}_t - \mathbf{x}_t \mathbf{B})'\right) \quad (33)$$

where $|\Sigma|$ is the determinant of the covariance matrix. We now turn to specifying the prior distributional assumptions for \mathbf{B} and Σ .

Diffuse Improper Prior

The lack of specific prior knowledge about the elements of \mathbf{B} and Σ can be reflected by employing the Jeffreys' prior, which, in the multivariate setting, takes the form¹⁸

$$\pi(\mathbf{B}, \Sigma) \propto |\Sigma|^{-\frac{N+1}{2}} \quad (34)$$

The posterior distributions parallel those in the univariate case. With the risk of stating the obvious, note that \mathbf{B} is a random matrix; therefore, its posterior distribution, conditional on Σ , will be a generalization of the multivariate normal posterior distribution in (9). To describe it, we first vectorize (expand column-wise) the matrix of regression coefficients, \mathbf{B} , and denote the

resulting $KN \times 1$ vector by β ,

$$\beta = \text{vec}(\mathbf{B}) = \begin{pmatrix} \alpha' \\ \beta'_1 \\ \vdots \\ \beta'_K \end{pmatrix}$$

by stacking vertically the columns of \mathbf{B}' . It can be shown that β 's posterior distribution, conditional on Σ , is a multivariate normal given by

$$p(\beta | \mathbf{Y}, \mathbf{X}, \Sigma) = N(\hat{\beta}, \Sigma \otimes (\mathbf{X}'\mathbf{X})^{-1}) \quad (35)$$

where $\hat{\beta} = \text{vec}(\hat{\mathbf{B}}) = \text{vec}((\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}))$ is the vectorized OLS estimator of \mathbf{B} and " \otimes " denotes the Kronecker product.¹⁹

The posterior distribution of Σ can be shown to be the inverted-Wishart distribution (the multivariate analog of the inverted-gamma distribution),

$$p(\Sigma | \mathbf{y}, \mathbf{X}) = IW(v^*, \mathbf{S}) \quad (36)$$

where the degrees of freedom parameter is $v^* = T - K + N + 1$ and the scale matrix is $\mathbf{S} = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})$.

A full Bayesian informative prior approach to estimation of the multivariate linear regression model would involve specifying a prior distribution for the regression coefficients, β , and the covariance matrix, Σ . The conjugate prior scenario is invariably the scenario of choice, so as to keep the estimation within analytically manageable boundaries. That scenario consists of a multivariate normal prior for β and inverted-Wishart for Σ .²⁰

KEY POINTS

- To account for estimation risk and to incorporate prior information, regression estimation can be cast in a Bayesian setting.
- Depending on the amount of prior information, diffuse or informative priors can be selected for the regression parameters.
- Under the assumption that the regression innovations are distributed with the normal

distribution, the natural conjugate priors for the regression coefficients and variance are Gaussian and inverted- χ^2 distributions, respectively.

- The case of unequal variances is easily incorporated into the linear regression. Unequal variances may be due to reasons such as structural breaks in time series data or nonhomogeneity in cross-sectional data.

NOTES

1. In order for the inverse matrix in (6) to exist, it is necessary that $\mathbf{X}'\mathbf{X}$ be nonsingular, that is, that the $n \times K$ matrix \mathbf{X} have a rank K (all its columns be linearly independent).

2. The MLE of σ^2 is in fact

$$\hat{\sigma}_{MLE}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})$$

However, as it is not unbiased, the estimator in (7) is more often employed.

3. In fact, it is possible to describe fully the distribution of β even without knowing its unconditional distribution, by employing a numerical simulation method such as the *Gibbs sampler*, for example, and making inferences on the basis of samples drawn from β 's and σ^2 's posterior distributions.
4. We denote the multivariate scaled, noncentral Student's t distribution with degrees of freedom ν , location parameter vector μ , and scale matrix \mathbf{S} by $t(\nu, \mu, \mathbf{S})$. Its mean and covariance matrix are given, respectively, by μ and $\mathbf{S}^{-1}\nu/(\nu - 2)$.
5. There are two contrasting approaches to prior parameter assertion. The full Bayesian approach calls for specifying the hyperprior parameters independently of the data used for model estimation. The empirical Bayesian approach would use the OLS estimate, $\hat{\beta}$, obtained from the data sample used for estimation, as the value for the hyperprior parameter.
6. The mean and variance of a random variable X distributed with the inverted- χ^2 distribution with parameters ν and c are given,

respectively, by

$$E(X) = \frac{\nu}{\nu - 2}c \quad \text{var}(X) = \frac{2\nu^2}{(\nu - 2)^2(\nu - 4)}c^2$$

7. See Chapter 6 in Rachev et al. (2008) for more details on this shrinkage effect.
8. Denoting the sampling and posterior densities by $f(x|\theta)$ and $p(\theta|x)$, respectively, the predictive density one step ahead is defined as

$$f(x_{+1}|x) = \int f(x_{+1}|\theta)p(\theta|x)d\theta$$

where x is the observed data, θ is the sampling distribution's parameter, and x_{+1} denotes the one-step-ahead realization.

9. Returns on interest-rate instruments and foreign exchange are particularly likely to exhibit structural breaks.
10. See, for example, Wang and Zivot (2000). Chapter 11 in Rachev et al. (2008) discusses the so-called "regime switching" models, in which parameters are allowed to change values according to the state of the world prevailing in a particular period in time.
11. See Chapter 4 in Zellner (1971).
12. See Chapter 5 in Rachev et al. (2008) for details on how to employ numerical simulation methods to tackle inference when no analytical results are available.
13. The standard deviation of the univariate Student's t distribution with degrees-of-freedom parameter ν and scale parameter σ is given by $\sigma\sqrt{\nu/(\nu - 2)}$.
14. Notice that, since the Student's t distribution is unimodal, these (symmetric) intervals are also the highest posterior density intervals.
15. Although the multivariate normal distribution is usually assumed because of its analytical tractability, dependencies among asset returns could be somewhat more complex than what the class of elliptical distributions (to which the normal distribution belongs) is able to describe. Alternative distributional assumptions could be made

at the expense of analytical convenience and occasional substantial estimation problems (especially in high-dimensional settings). A more flexible way of dependence modeling is provided through the use of copulas. Some types of copulas could also suffer from estimation problems, especially in large-scale applications.

16. For applications to portfolio construction, see Chapters 6 through 9 in Rachev et al. (2008).
17. The expression in the exponent in (33) could also be written as

$$-\frac{1}{2}\text{tr}(\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB})\Sigma^{-1},$$

where "tr" denotes the trace operator, which sums the diagonal elements of a square matrix.

18. As in the univariate case, we assume independence between (the elements of) \mathbf{B} and Σ .
19. The Kronecker product is an operator for direct multiplication of matrices (which are not necessarily compatible). For two matrices, A of size $m \times n$ and B of size $p \times q$, the Kronecker product is defined as

$$A \otimes B = \begin{pmatrix} a_{1,1}B & a_{1,2}B & \dots & a_{1,n}B \\ \dots & \dots & \dots & \dots \\ a_{m,1}B & a_{m,2}B & \dots & a_{m,n}B \end{pmatrix}$$

resulting in an $mp \times nq$ block matrix.

20. See Chapters 6 and 7 of Rachev et al. (2008) for further details in the context of portfolio selection.

REFERENCES

- Rachev, S., Hsu, J., Bagasheva, B., and Fabozzi, F. (2008). *Bayesian Methods in Finance*. Hoboken, NJ: Wiley & Sons.
- Wang, J., and Zivot, E. (2000). A Bayesian time series model of multiple structural changes in level, trend, and variance. *Journal of Business and Economic Statistics* 18(3): 374–386.
- Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley & Sons.

Bayesian Estimation of ARCH-Type Volatility Models

BILIANA S. GÜNER, PhD

Assistant Professor of Statistics and Econometrics, Ozyegin University, Turkey

SVETLOZAR RACHEV, PhD, Dr Sci

Frey Family Foundation Chair Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

JOHN S. J. HSU, PhD

Professor of Statistics and Applied Probability, University of California, Santa Barbara

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Empirical evidence abounds that asset returns exhibit characteristics such as volatility clustering, asymmetry, and heavy-tailedness. Volatility clustering describes the tendency of returns to alternate between periods of high volatility and low volatility. In addition, volatility responds asymmetrically to positive and negative return shocks—it tends to be higher when the market falls than when it rises. The nonconstancy of volatility has been suggested as an underlying reason for returns' fat tails. Volatility models attempt to systematically explain these stylized facts about asset returns. The Bayesian methodology offers distinct advantages over the classical framework in estimating volatility models. Parameter restrictions, such as stationarity restriction, are notoriously difficult to handle within the frequentist setting and straightforward to implement in the Bayesian one. The MCMC numerical simulation methods facilitate greatly the estimation of complex volatility models, such as Markov-switching volatility models.

Generalized autoregressive conditional heteroskedastic (GARCH) models are used in financial modeling to provide a measure of volatility that could be employed in portfolio selection, risk management, and derivatives pricing. In this entry, we focus on the Bayesian treatment of

GARCH model estimation. Our discussion of prior distributions' choice and posterior analysis is developed around an example where the data are assumed to follow the Student's t distribution. We then introduce a Bayesian approach to Markov-switching GARCH models

and explain in detail the steps one could use to estimate this important extension of the simple GARCH model.

BAYESIAN ESTIMATION OF THE GARCH(1,1) MODEL

Volatility is a forward-looking concept. It is the variance of the yet unrealized asset return, conditional on all relevant, available information. Denote by r_t the asset return at time t and by F_{t-1} the set of information available up to time $t - 1$. The information set includes, for example, past asset returns and past trading volume. The return's dynamics can be described as follows:

$$r_t = \mu_{t|t-1} + \sigma_{t|t-1}\epsilon_t \quad (1)$$

where

- $\mu_{t|t-1}$ is the return's conditional expectation at time t ,
- $\sigma_{t|t-1}$ is the return's conditional volatility at time t ,
- ϵ_t is a white noise process (a sequence of independent and identically distributed random variables with zero mean and variance of one).

The aim of volatility models is to specify the dynamics of $\sigma_{t|t-1}$. Autoregressive conditional heteroskedastic (ARCH)-type models describe the conditional volatility at time t as a deterministic function of (attribute of) past squared returns. That is, volatility at time t can be uniquely determined at time $t - 1$. The volatility updating expression of a GARCH(1,1) process is given by

$$\sigma_{t|t-1}^2 = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1|t-2}^2 \quad (2)$$

where $u_t = \sigma_{t|t-1}\epsilon_t$. The model parameters are restricted to be nonnegative— $\omega > 0$, $\alpha \geq 0$, and $\beta \geq 0$ —in order to ensure that the conditional variance is positive for all values of the white noise process, ϵ_t . Additionally, the requirement for stationarity imposes the constraint that the sum $\alpha + \beta$ is smaller than one.

Estimation of the model parameters is usually performed by likelihood maximization. Since the return at time t , r_t , depends on $\sigma_{t|t-1}$ and through it on the conditional volatilities in all previous periods, the unconditional density of the return is not available in closed form (it is a mixture of densities depending on the dynamics of $\sigma_{t|t-1}^2$). Therefore, the likelihood function of the GARCH(1,1) model is expressed as the product of the conditional densities of r_t for each period t , $t = 1, 2, \dots, T$.

Given F_0 , the likelihood function $L(\boldsymbol{\theta} | r_1, r_2, \dots, r_T, F_0)$ is written as¹

$$L(\boldsymbol{\theta} | \mathbf{r}, F_0) = f(r_1 | \boldsymbol{\theta}, F_0) f(r_2 | \boldsymbol{\theta}, F_1) \dots f(r_T | \boldsymbol{\theta}, F_{T-1}) \quad (3)$$

where $\mathbf{r} = (r_1, r_2, \dots, r_T)$. Due to the form of the likelihood function, posterior estimation is performed, without exception, numerically. This, on the other hand, implies that few, if any, restrictions exist on the choice of prior distributions, when estimation is cast in a Bayesian setting.

In this entry, our focus is on the Student's t distributional assumption for the return disturbances, in an attempt to reflect the empirically observed heavy-tailedness of returns. This comes at the expense of only a marginal increase in complexity (compared to estimation of a model with normally distributed disturbances). The two numerical simulation methods we employ to simulate the posterior distribution of the vector of model parameters, $\boldsymbol{\theta}$, are the Metropolis-Hastings algorithm and the Gibbs sampler.²

Our focus is the model of returns in (1) with a modification. We assume that the return mean is unconditional and equal to zero. That is, we define our parameter vector as $\boldsymbol{\theta} = (\omega, \alpha, \beta, \nu)$

Distributional Setup

Next, we outline the general setup we use in our Bayesian estimation of the GARCH(1,1) model. We modify this setup in the second half of the entry, where we discuss regime switching.

Likelihood Function

Assuming that ϵ_t is distributed with a Student's t distribution with ν degrees of freedom, we write the likelihood function for the model's parameters as

$$L(\boldsymbol{\theta} | \mathbf{r}, F_0) \propto \prod_{t=1}^T \left[(\sigma_{t|t-1}^2)^{-1} \left(1 + \frac{1}{\nu} \frac{r_t^2}{\sigma_{t|t-1}^2} \right)^{-\frac{\nu+1}{2}} \right] \quad (4)$$

where σ_0^2 is considered as a known constant, for simplicity. Under the Student's t assumption for ϵ_t , the conditional volatility at time t is given by

$$\frac{\nu}{\nu - 2} \sigma_t^2$$

for ν greater than 2.

Prior Distributions

For simplicity, assume that the conditional variance parameters have uninformative diffuse prior distributions over their respective ranges,³

$$\pi(\omega, \alpha, \beta) \propto 1 I_{\{\theta_G\}} \quad (5)$$

where $I_{\{\theta_G\}}$ is an indicator function reflecting the constraints on the conditional variance parameters,

$$I_{\{\theta_G\}} = \begin{cases} 1 & \text{if } \omega > 0, \alpha > 0, \text{ and } \beta > 0, \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

The choice of prior distribution for the degrees-of-freedom parameter, ν , requires more care. Bauwens and Lubrano (1998) show that if a diffuse prior for ν is asserted on the interval $[0, \infty)$, the posterior distribution of ν is not proper (its right tail does not decay quickly enough, so that the posterior does not integrate to 1). Therefore, the prior for ν needs to be proper. Geweke (1993a) advocates the use of an exponential prior distribution with density given by

$$\pi(\nu) = \lambda \exp(-\nu\lambda) \quad (7)$$

The mean of the exponential distribution is given by $1/\lambda$. The parameter λ can thus be uniquely determined from the prior intuition about ν 's mean. Another prior option for ν is a uniform prior over an interval $[0, M]$, where

M is some finite number. Empirical research indicates that the degrees-of-freedom parameter calibrated from financial returns data (especially of daily and higher frequency) is usually less than 20, so the upper bound, M , of ν 's range could be fixed at 20, for instance. Bauwens and Lubrano propose a third prior for ν —the upper half of a Cauchy distribution centered around zero. In our discussion, we adopt the exponential prior distribution for ν in (7).

Posterior Distributions

Given the distributional assumptions above, the posterior distribution of $\boldsymbol{\theta}$ is written as

$$\begin{aligned} p(\boldsymbol{\theta} | \mathbf{r}, F_0) &\propto \prod_{t=1}^T \left[(\sigma_{t|t-1}^2)^{-1} \left(1 + \frac{1}{\nu} \frac{r_t^2}{\sigma_{t|t-1}^2} \right)^{-\frac{\nu+1}{2}} \right] \\ &\times \exp(-\nu\lambda) \\ &\times I_{\{\theta_G\}} \end{aligned} \quad (8)$$

The restrictions on ω , α , and β are enforced during the sampling procedure by rejecting the draws that violate them. Stationarity can also be imposed and dealt with in the same way.

As evident from the expression in (8), the joint posterior density does not have a closed form. Posterior numerical simulations are facilitated if one employs a specific representation of the Student's t distribution—a scale mixture of normal distributions. We explain this representation before we move on to the discussion of sampling algorithms.

Mixture of Normals Representation of the Student's t Distribution

Suppose that return r_t is distributed with the Student's t distribution with ν degrees of freedom, scale parameter σ , and location parameter μ . This distributional assumption can be represented as a scale mixture of normal distributions, given by⁴

$$r_t | \mu_t, \sigma_t, \eta_t \sim N \left(\mu_t, \frac{\sigma_t}{\eta_t} \right) \quad (9)$$

where η , the so-called “mixing variable,” has the gamma distribution,

$$\eta_t | \nu \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \quad \text{for } t = 1, \dots, T \quad (10)$$

The benefit of employing this representation is increased tractability of the posterior distribution because the nonlinear expression for the model’s likelihood in (4) is linearized. Sampling from the conditional distributions of the remaining parameters is thus greatly facilitated. This comes at the expense of T additional model parameters, $\boldsymbol{\eta} = (\eta_1, \dots, \eta_T)'$, whose conditional posterior distribution needs to be simulated as well.⁵

Under this Student’s t representation, the parameter vector, $\boldsymbol{\theta}$, is transformed to⁶

$$\boldsymbol{\theta} = (\omega, \alpha, \beta, \nu, \boldsymbol{\eta}') \quad (11)$$

The log-likelihood function for $\boldsymbol{\theta}$ is simply the normal log-likelihood,

$$\begin{aligned} \log(L(\boldsymbol{\theta} | \mathbf{r}, F_0)) &= \text{const} - \frac{1}{2} \sum_{t=1}^T \\ &\times \left[\log(\sigma_{t|t-1}^2) - \log(\eta_t) + \frac{\eta_t r_t^2}{\sigma_{t|t-1}^2} \right] \end{aligned} \quad (12)$$

The posterior distribution of $\boldsymbol{\theta}$ has an additional term reflecting the mixing variables’ distribution. The log-posterior distribution is written as

$$\begin{aligned} \log(p(\boldsymbol{\theta} | \mathbf{r}, F_0)) &= \text{const} - \frac{1}{2} \sum_{t=1}^T \\ &\times \left[\log(\sigma_{t|t-1}^2) - \log(\eta_t) + \frac{\eta_t r_t^2}{\sigma_{t|t-1}^2} \right] \\ &+ \frac{T\nu}{2} \log\left(\frac{\nu}{2}\right) - T \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) \\ &+ \left(\frac{\nu}{2} - 1\right) \sum_{t=1}^T \log(\eta_t) \\ &- \frac{\nu}{2} \sum_{t=1}^T (\eta_t) - \nu\lambda \end{aligned} \quad (13)$$

$$\text{for } \omega > 0, \quad \alpha \geq 0, \quad \text{and } \beta \geq 0$$

Next, we discuss some strategies for simulating the posterior in (13).

Posterior Simulations with the Metropolis-Hastings Algorithm

The Metropolis-Hastings (M-H) algorithm could be implemented in two ways. The first way is by sampling the whole parameter vector, $\boldsymbol{\theta}$, from a proposal distribution (usually a multivariate Student’s t distribution) centered on the posterior mode and scaled by the negative inverse Hessian (evaluated at the posterior mode).⁷ The second way is by employing a sampling scheme in which the parameter vector is updated component by component. Here, we focus on the latter M-H implementation.

Consider the decomposition of the parameter vector $\boldsymbol{\theta}$ into three components, $\boldsymbol{\theta} = (\boldsymbol{\theta}_G, \nu, \boldsymbol{\eta}')$, where $\boldsymbol{\theta}_G = (\omega, \alpha, \beta)$. We would like to employ a scheme of sampling consecutively from the conditional posterior distributions of the components, given, respectively, by $p(\boldsymbol{\theta}_G | \boldsymbol{\eta}, \nu, \mathbf{r}, F_0)$, $p(\nu | \boldsymbol{\theta}_G, \boldsymbol{\eta}, \mathbf{r}, F_0)$, and $p(\boldsymbol{\eta} | \boldsymbol{\theta}_G, \nu, \mathbf{r}, F_0)$. The scale mixture of normals representation of a Student’s t distribution allows us to recognize the conditional posterior distribution of the last component, $\boldsymbol{\eta}$, as a standard distribution. For the first two components, $\boldsymbol{\theta}_G$ and ν , whose posterior distributions are not of standard form, we offer two posterior simulation approaches and mention alternatives that have been suggested in the literature.

Conditional Posterior Distribution for $\boldsymbol{\eta}$

The full conditional posterior distribution for the (independently-distributed) mixing parameters, η_t , $t = 1, \dots, T$, can be shown to be a gamma distribution,

$$\begin{aligned} p(\eta_t | \boldsymbol{\theta}_G, \nu, \mathbf{r}, F_0) \\ = \text{Gamma}\left(\frac{\nu + 1}{2}, \frac{r_t^2}{2\sigma_{t|t-1}^2} + \frac{\nu}{2}\right) \end{aligned} \quad (14)$$

Conditional Posterior Distribution for ν

It can be seen from (13) that the conditional posterior distribution of the degrees-of-freedom parameter, ν , does not have a standard form.

The kernel of the posterior distribution is given by the expression,

$$p(v | \theta_G, \eta, r, F_0) \propto \Gamma\left(\frac{v}{2}\right)^{-T} \left(\frac{v}{2}\right)^{\frac{Tv}{2}} \exp(v\lambda^*) \quad (15)$$

where

$$\lambda^* = \frac{1}{2} \sum_{t=1}^T (\log(\eta_t) - \eta_t) - \lambda \quad (16)$$

Geweke (1993b) describes a rejection sampling approach that could be employed to simulate draws from the conditional posterior distribution of v in (15). In this entry, we employ a sampling algorithm called the griddy Gibbs sampler. The appendix provides details on it.

Proposal Distribution for θ_G

The kernel of θ_G 's log-posterior distribution is given by the expression,

$$\begin{aligned} &\log(p(\theta_G | \theta_{-\theta_G}, r, F_0)) \\ &= \text{const} - \frac{1}{2} \sum_{t=1}^T \left[\log(\sigma_{t|t-1}^2) + \frac{\eta_t r_t^2}{\sigma_{t|t-1}^2} \right] \\ &\text{for } \omega > 0, \quad \alpha \geq 0, \quad \text{and } \beta \geq 0 \end{aligned}$$

where $\sigma_{t|t-1}^2, t = 1, \dots, T$, is a function of θ_G .

We specify a Student's t proposal distribution for θ_G , centered on the posterior mode of θ_G and scaled by the negative inverse Hessian of the posterior kernel, evaluated at the posterior mode. Other approaches for posterior simulation, for example, the griddy Gibbs sampler, could be employed as well. (In this case, the components of θ_G would be sampled separately.)

Having determined the full conditional posterior distribution η , as well as a proposal distribution for θ_G and a sampling scheme for v , implementing a hybrid M-H algorithm is straightforward. Its steps are as follows. At iteration m of the algorithm,

- Draw an observation, θ_G^* , of the vector of conditional variance parameters, θ_G , from its proposal distribution.

- Check whether the positivity (and stationarity) parameter restrictions on the components of θ_G are satisfied. If not, draw θ_G^* repeatedly until they are satisfied.
- Compute the acceptance probability

$$\begin{aligned} &a(\theta_G^*, \theta_G^{(t-1)}) \\ &= \min \left\{ 1, \frac{p(\theta_G^* | \mathbf{y}) / q(\theta_G^* | \theta_G^{(t-1)})}{p(\theta_G^{(t-1)} | \mathbf{y}) / q(\theta_G^{(t-1)} | \theta_G^*)} \right\} \quad (17) \end{aligned}$$

where $p(\theta_G | \mathbf{y})$ is θ_G 's posterior distribution and $q(\theta_G | \cdot)$ is θ_G 's proposal distribution. The previous draw of the parameter vector is given by θ_G^{t-1} . Accept or reject the candidate draw θ_G^* with probability $a(\theta_G^*, \theta_G^{(t-1)})$.

- Draw an observation, $\eta^{(m)}$, from the full conditional posterior distribution, $p(\eta_t | \theta_G^{(m)}, r, F_0)$, in (14)
- Draw an observation, $v^{(m)}$, from its conditional posterior distribution with kernel in (15) using the griddy Gibbs sampler as explained in the appendix.

At each iteration of the sampling algorithm, the sampling strategy described above produces a large output consisting of the draws from the model parameters and the T mixing variables, η . However, since the role of the mixing parameters is only auxiliary and their conditional distribution is of no interest, at any iteration of the algorithm above one needs to store only the latest draw of η (as well as the draws of v and θ_G , of course).

In the simple GARCH model discussed now, it is implicitly assumed that expression (2) describes the volatility process during the whole sample period and (at least) in the short run after the end of the sample. That is, the parameters of the model are unchanged throughout. It is not inconceivable, however, that the volatility dynamics differ in different periods. Then, volatility forecasts produced by a simple (single-regime) model are likely to overestimate volatility during periods of low volatility and underestimate it during periods of

high volatility. In the next section, we discuss a class of models extending the simple GARCH(1,1) model, which could potentially provide more accurate volatility forecasting power. Regime-switching models incorporate the possibility that the dynamics of the volatility process evolves through different states of nature, which we call regimes.

MARKOV-SWITCHING GARCH MODELS

The *Markov-switching* (MS) models, introduced by Hamilton (1989), provide maximal flexibility in modeling transitions of the volatility dynamics across regimes. They form the class of the so-called endogenous regime-switching models in which transitions between states of nature are governed by parameters estimated within the model; the number of transitions is not specified a priori, unlike the number of states. Each volatility state could be revisited multiple times.⁸ In our discussion below, we use the terms “state” and “regime” interchangeably.

Different approaches to introducing regime changes in the GARCH process have been proposed in the empirical finance literature. Hamilton and Susmel (1994) incorporate a regime-dependent parameter, g_{S_t} , into the standard deviation (scale) of the returns process,

$$r_t = \mu_{t|t-1} + \sqrt{g_{S_t}} \sigma_{t|t-1} \epsilon_t$$

where S_t denotes period t 's regime. Another option, pursued by Cai (1994), is to include a regime-dependent parameter as part of the constant in the conditional variance equation,

$$\sigma_{t|t-1}^2 = (\omega + g_{S_t}) + \sum_{p=1}^P \alpha_p u_{t-p}^2$$

Both Hamilton and Susmel (1994) and Cai (1994) model the dynamics of the conditional variance with an ARCH process. The reason, as explained further below, is that when GARCH term(s) are present in the process, the regime-

dependence makes the likelihood function analytically intractable.

The most flexible approach to introducing regime-dependence is to allow all parameters of the conditional variance equation to vary across regimes. That approach is suggested by Henneke, Rachev, Fabozzi, and Nikolov (2011) who model jointly the conditional mean as an ARMA(1,1) process in a Bayesian estimation setting.⁹ The implication for the dynamics of the conditional variance is that the manner in which the variance responds to past return shocks and volatility levels changes across regimes. For example, high-volatility regimes could be characterized by “hyper-sensitivity” of asset returns to return shocks and high volatility in one period could have a more lasting effect on future volatilities compared to low-volatility regimes. This would call for a different relationship between the parameters α and β in different regimes.

In this section, we discuss the estimation method of Henneke, Rachev, Fabozzi, and Nikolov (2011), with some modifications.

Preliminaries

Suppose that there are three states the conditional volatility can occupy, denoted by $i, i = 1, 2, 3$. We could assign economic interpretation to them by labeling them “a low-volatility state,” “a normal-volatility state,” and “a high-volatility state.” Denote by π_{ij} the probability of a transition from state i to state j . The transition probabilities, π_{ij} , could be arranged in the transition probability matrix, Π ,

$$\Pi = \begin{pmatrix} \pi_{11} & \pi_{12} & \pi_{13} \\ \pi_{21} & \pi_{22} & \pi_{23} \\ \pi_{31} & \pi_{32} & \pi_{33} \end{pmatrix} \quad (18)$$

such that the probabilities in each row sum up to 1. The Markov property (central to model estimation, as we will see below) that lends its name to the MS models concerns the memory of the process—which volatility regime the system visits in a given period depends only on the

regime in the previous period. Analytically, the Markov property is expressed as

$$P(S_t | S_{t-1}, S_{t-2}, \dots, S_1) = P(S_t | S_{t-1}) \quad (19)$$

Each row of Π in (18) represents the three-dimensional conditional probability distribution of S_t , conditional on the regime realization in the previous period, S_{t-1} . We say that $\{S_t\}_{t=1}^T$ is a three-dimensional (discrete-time) Markov chain with transition matrix, Π .

In the regime-switching GARCH(1,1) setting, the expression for the conditional variance dynamics becomes

$$\sigma_{t|t-1}^2 = \omega(S_t) + \alpha(S_t)u_{t-1}^2 + \beta(S_t)\sigma_{t-1|t-2}^2 \quad (20)$$

For each period t ,

$$(\omega(S_t), \alpha(S_t), \beta(S_t)) = \begin{cases} (\omega_1, \alpha_1, \beta_1) & \text{if } S_t = 1, \\ (\omega_2, \alpha_2, \beta_2) & \text{if } S_t = 2, \\ (\omega_3, \alpha_3, \beta_3) & \text{if } S_t = 3 \end{cases}$$

The presence of the GARCH component in (20) complicates the model estimation substantially. To see this, notice that, via $\sigma_{t-1|t-2}^2$, the current conditional variance depends on the conditional variances from all preceding periods and, therefore, on the whole unobservable sequence of regimes up to time t . A great number of regime paths could lead to the particular conditional variance at time t (the number of possible regime combinations grows exponentially with the number of time periods), rendering classical estimation very complicated. For that reason, the early treatments of MS models include only an ARCH component in the conditional variance equation. The MCMC methodology, however, copes easily with the specification in (20), as we will see below.

We adopt the same return decomposition as in (1)—with the conditional mean set to zero—and note that, given the regime path, (20) represents the same conditional variance dynamics as a simple GARCH(1,1) process. We return to this point again further below when we discuss estimation of that MS GARCH(1,1) model.

Next, we outline the prior assumptions for the MS GARCH(1,1) model.

Prior Distributional Assumptions

The parameter vector of the MS GARCH(1,1) model, specified by (1), (20), and the Markov chain $\{S_t\}_{t=1}^T$, is given by

$$\theta = (\eta', \nu, \theta_{G,1}, \theta_{G,2}, \theta_{G,3}, \pi_1, \pi_2, \pi_3, S) \quad (21)$$

where, for $i = 1, 2, 3$,

$$\theta_{G,i} = (\omega_i, \alpha_i, \beta_i) \quad \text{and} \quad \pi_i = (\pi_{i1}, \pi_{i2}, \pi_{i3})$$

and S is the regime path for all periods,

$$S = (S_1, \dots, S_T)$$

Our prior specifications for η and ν remain unchanged from our earlier discussion: The scale-mixture-of-normals mixing parameters, η , and the degrees-of-freedom parameter, ν , are not affected by the regime specification in the MS GARCH(1,1) model. We assert prior distributions for the vector of conditional variance parameters, $\theta_{G,i}$, under each regime, i , and a prior distribution for each triple of transition probabilities π_i , $i = 1, 2, 3$.

Prior Distributions for $\theta_{G,i}$

To reflect our prior intuition about the effect the three regimes have on the conditional variance parameters, we assert proper normal priors for $\theta_{G,i}$, $i = 1, 2, 3$.

$$\theta_{G,i} \sim N(\mu_i, \Sigma_i) I_{\{\theta_{G,i}\}} \quad (22)$$

where the indicator function, $I_{\{\theta_{G,i}\}}$, is given in (6). As explained earlier in the entry, the parameter constraints are imposed during the implementation of the sampling algorithm.

Prior Distribution for π_i

A convenient prior for the probability parameter in a binomial experiment is the beta distribution.¹⁰ The analogue of the beta distribution in the multivariate case is the so-called Dirichlet distribution.¹¹ Therefore, we specify a Dirichlet prior distribution for each triple of transition probabilities, $i = 1, 2, 3$,

$$\pi_i \sim \text{Dirichlet}(a_{i1}, a_{i2}, a_{i3}) \quad (23)$$

To elicit the prior parameters, a_{ij} , $i, j = 1, 2, 3$, it is sufficient that one express prior intuition about the expected value of each of the transition probabilities in a triple, then solve the system equations for a_{ij} .

Estimation of the MS GARCH Model

The evolution of volatility in the MS GARCH model is governed by the realizations of the unobservable (latent) regime variable, S_t , $t = 1, \dots, T$. Hence, the discrete-time Markov chain, $\{S_t\}_{t=1}^T$ is also called a hidden Markov process. Earlier, we briefly discussed that the presence of the hidden Markov process creates a major estimation difficulty in the classical setting. The Bayesian methodology, in contrast, deals with the latent-variable characteristic in an easy and natural way: The latent variable is simulated together with the model parameters. In other words, the parameter space is augmented with S_t , $t = 1, \dots, T$, in much the same way as the vector of mixing variables, η , was added to the parameter space in estimating the Student's t GARCH(1,1) model. The distribution of S is a multinomial distribution,

$$\begin{aligned} p(S|\boldsymbol{\pi}) &= \prod_{t=1}^{T-1} p(S_{t+1}|S_t, \boldsymbol{\pi}) \\ &= \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \dots \pi_{32}^{n_{32}} \pi_{33}^{n_{33}} \\ &= \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} (1 - \pi_{11} - \pi_{12})^{n_{13}} \dots \\ &\quad \pi_{32}^{n_{32}} (1 - \pi_{31} - \pi_{32})^{n_{33}} \end{aligned} \quad (24)$$

where n_{ij} denotes the number of times the chain transitions from state i to state j during the span of period 1 through period T . The first equality in (24) follows from the Markov property of $\{S_t\}_{t=1}^T$.

Based on our discussion of the Student's t GARCH(1,1) model and the hidden Markov process, as well as the prior distributional assumptions for $\boldsymbol{\pi}_i$ and $\boldsymbol{\theta}_{G,i}$, $i = 1, 2, 3$, the joint log-posterior distribution of the MS GARCH(1,1) model's parameter vector $\boldsymbol{\theta}$ is

given by

$$\begin{aligned} \log(p(\boldsymbol{\theta} | \mathbf{r}, F_0)) &= \text{const} \\ &- \frac{1}{2} \sum_{t=1}^T \left[\log(\sigma_{t|t-1}^2) + \log(\eta_t) + \frac{\eta_t r_t^2}{\sigma_{t|t-1}^2} \right] \\ &- \frac{1}{2} \sum_{i=1}^3 (\boldsymbol{\theta}_{G,i} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{\theta}_{G,i} - \boldsymbol{\mu}_i) I_{\{S(t)=i\}} \\ &+ \frac{T\nu}{2} \log\left(\frac{\nu}{2}\right) - T \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) + \left(\frac{\nu}{2} - 1\right) \\ &\times \sum_{t=1}^T \log(\eta_t) - \frac{\nu}{2} \sum_{t=1}^T \eta_t - \nu\lambda \\ &+ \sum_{i=1}^3 \sum_{j=1}^3 (a_{ij} + n_{ij} - 1) \log(\pi_{ij}) \end{aligned} \quad (25)$$

for $\omega_i > 0$, $\alpha_i \geq 0$, and $\beta_i \geq 0$

Although (25) looks very similar to the joint log-posterior in (13), there is a crucial difference. The model's log-likelihood (given by the right-hand-side term in the first line of (25)) depends on the whole sequence of regimes, S . Conditional on S , however, it is the same log-likelihood as in (12). We will exploit this fact in constructing the posterior simulation algorithm as an extension of the algorithm for the Student's t GARCH(1,1) model estimation.

We now outline the posterior results for $\boldsymbol{\pi}_i$, S , and $\boldsymbol{\theta}_{G,i}$. The posterior results for the degrees-of-freedom parameter, ν , and the mixing variables, $\boldsymbol{\eta}$, remain unchanged from our earlier discussion.

Conditional Posterior Distribution of $\boldsymbol{\pi}_i$

The conditional log-posterior distribution of the vector of transition probabilities, $\boldsymbol{\pi}_i$, $i = 1, 2, 3$, is given by

$$\begin{aligned} \log(p(\boldsymbol{\pi}_i | \mathbf{r}, \boldsymbol{\theta}_{-\boldsymbol{\pi}_i})) &= \text{const} \\ &+ \sum_{j=1}^3 (a_{ij} + n_{ij} - 1) \log(\pi_{ij}) \\ &\text{for } i = 1, 2, 3 \end{aligned} \quad (26)$$

where $\boldsymbol{\theta}_{-\boldsymbol{\pi}_i}$ denotes the vector of all parameters except $\boldsymbol{\pi}_i$. The expression in (26) is readily

recognized as the logarithm of the kernel of a Dirichlet distribution with parameters $(a_{i1} + n_{i1}, a_{i2} + n_{i2}, a_{i3} + n_{i3})$. The parameters a_{ij} are specified a priori, while the parameters n_{ij} can be determined by simply counting the number of times the Markov chain, $\{S_t\}_{t=1}^T$, transitions from i to j .

Sampling from the Dirichlet distribution in (26) is accomplished easily in the following way.¹² For each $i, i = 1, 2, 3$,

(1) sample three independent observations,

$$y_{i1} \sim \chi^2_{2(a_{i1}+n_{i1})}, y_{i2} \sim \chi^2_{2(a_{i2}+n_{i2})},$$

$$y_{i3} \sim \chi^2_{2(a_{i3}+n_{i3})}$$

(2) set

$$\pi_{i1} = \frac{y_{i1}}{\sum_{k=1}^3 y_{ik}}, \quad \pi_{i2} = \frac{y_{i2}}{\sum_{k=1}^3 y_{ik}},$$

$$\pi_{i3} = \frac{y_{i3}}{\sum_{k=1}^3 y_{ik}}$$

Conditional Posterior Distribution of S

In the three-regime switching setup of this entry, the number of regime paths that could have potentially generated S_T , the regime in the final period, is 3^T . The level of complexity makes it impossible to obtain a draw of the whole $1 \times T$ vector, S , at once. Instead, its components can be drawn one at a time, in a T -step procedure. In other words, at each step, we sample from the full conditional posterior density of S_t given by

$$p(S_t = i | r, \theta_{-S}, S_{-t}) \quad (27)$$

where θ_{-S} is the parameter vector in (21) excluding S and S_{-t} is the regime path excluding the regime at time t . Applying the rules of conditional probability, $p(S_t = i | r, \theta_{-S_t})$ is written as

$$p(S_t = i | r, \theta_{-S}, S_{-t}) = \frac{p(S_t = i, S_{-t}, r | \theta_{-S})}{p(S_{-t}, r | \theta_{-S})}$$

$$= \frac{p(r | \theta_{-S}, S_{-t}, S_t = i) p(S_t = i, S_{-t} | \theta_{-S})}{p(S_{-t}, r | \theta_{-S})} \quad (28)$$

The first term in the numerator, $p(r | \theta_{-S}, S_{-t}, S_t = i)$, is simply the model's likelihood evaluated at a given regime path, in which $S_t = i$. The second term in the numerator, $p(S_t = i, S_{-t})$, is given, by the Markov property, by

$$p(S_t = i, S_{-t} | \theta_{-S}) = p(S_t = i, S_{t-1} = j, S_{t+1} = k | \theta_{-S})$$

$$= \pi_{j,i} \pi_{i,k} \quad (29)$$

while the denominator in (28) is expressed as

$$p(S_{-t}, r | \theta_{-S}) = \sum_{s=1}^3 p(S_t = s, S_{-t}, r | \theta_{-S}) \quad (30)$$

Using (28), (29), and (30), we obtain the conditional posterior distribution of S_t as

$$p(S_t = i | r, \theta_{-S}, S_{-t}) = \frac{p(r | \theta_{-S}, S_{-t}, S_t = i) \pi_{j,i} \pi_{i,k}}{\sum_{s=1}^3 p(r | \theta_{-S}, S_{-t}, S_t = s) \pi_{j,s} \pi_{s,k}} \quad (31)$$

for $i = 1, 2, 3$. An observation, S_t^* , from the conditional density in (31) is obtained in the following way:

- Compute the probability in (31) for $i = 1, 2, 3$.
- Split the interval $(0, 1)$ into three intervals of lengths proportional to the probabilities in step (1).
- Draw an observation, u , from the uniform distribution $U[0, 1]$.
- Depending on which interval u falls into, set $S_t^* = i$.

To draw the regime path, $S^{(m)}$, at the m th iteration of the posterior simulation algorithm,

- Draw $S_1^{(m)}$ from $p(S_1 | r, \theta_{-S_1})$ in (31). Update $S^{(m)}$ with $S_1^{(m)}$.
- For $t = 2, \dots, T$, draw $S_t^{(m)}$ from $p(S_t | r, \theta_{-S_t})$ in (31). Update $S^{(m)}$ with $S_t^{(m)}$.

Proposal Distribution for $\theta_{G,i}$

The posterior distribution of the vector of conditional variance parameters is not available in

closed form because of the regime dependence of the conditional variance. Since in the regime-switching setting we adopted informative prior distributions for $\theta_{G,i}$, $i = 1, 2, 3$, the kernel of the conditional log-posterior distribution is a bit different from the one in (17) and is given by

$$\begin{aligned} \log(p(\theta_{G,i} | \theta_{-\theta_{G,i}}, r, F_0)) &= \text{const} \\ &- \frac{1}{2} \sum_{t=1}^T \left[\log(\sigma_{t|t-1}^2) + \log(\eta_t) + \frac{\eta_t r_t^2}{\sigma_{t|t-1}^2} \right] \\ &- \frac{1}{2} \sum_{i=1}^3 (\theta_{G,i} - \mu_i)' \Sigma_i^{-1} (\theta_{G,i} - \mu_i) I_{\{S_t=i\}}, \end{aligned} \quad (32)$$

$$\text{for } \omega > 0, \quad \alpha \geq 0, \quad \beta \geq 0, \quad \text{and} \\ i = 1, 2, 3$$

For a given regime path, S , the only difference between the earlier posterior kernel and (32) is the term reflecting the informative prior of $\theta_{G,i}$. Therefore, specifying a proposal distribution for $\theta_{G,i}$ is in no way different from the approach in the single-regime Student's t GARCH(1,1) setting.

Sampling Algorithm for the Parameters of the MS GARCH (1,1) Model

The sampling algorithm for the MS GARCH(1,1) model parameters consists of the following steps. At iteration m ,

- Draw $\pi_i^{(m)}$ from its posterior density in (26), for $i = 1, 2, 3$.
- Draw $S^{(m)}$ from (31).
- Draw $\eta^{(m)}$ from (14).
- Draw $\nu^{(m)}$ from (15).
- Draw $\theta_{G,i}^*$, $i = 1, 2, 3$, from the proposal distribution, as explained earlier.
- Check whether the parameter restrictions on the components of $\theta_{G,i}$ are satisfied; if not, draw $\theta_{G,i}^*$ repeatedly, until they are satisfied.
- Compute the acceptance probability in (17) and accept or reject $\theta_{G,i}^*$ for $i = 1, 2, 3$.

The parameter vector, θ , is updated as new components are drawn. The steps above are repeated a large number of times until convergence of the algorithm.

APPENDIX: THE GRIDDY GIBBS SAMPLER

Implementation of the Gibbs sampler requires that parameters' conditional posterior distributions be known. Sometimes, however, the conditional posterior distributions have no closed forms. In these cases, a special form of the Gibbs sampler, called the griddy Gibbs sampler, can be employed whereby the (univariate) conditional posterior densities are evaluated on grids of parameter values. The griddy Gibbs sampler, developed by Ritter and Tanner (1992), is a combination of the ordinary Gibbs sampler and a numerical routine. In this appendix, we illustrate the griddy Gibbs sampler with the posterior distribution of the degrees-of-freedom parameter, ν .

Recall the expression for the kernel of ν 's conditional log-posterior distribution,

$$\begin{aligned} \log(p(\nu | \theta_{-\nu}, r, F_0)) &= \text{const} \\ &+ \frac{T\nu}{2} \log\left(\frac{\nu}{2}\right) - T \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) \\ &+ \left(\frac{\nu}{2} - 1\right) \sum_{t=1}^T \log(\eta_t) - \frac{\nu}{2} \sum_{t=1}^T \eta_t - \nu\lambda. \end{aligned} \quad (33)$$

The griddy Gibbs sampler approach to drawing from the conditional posterior distribution of ν is to recognize that at iteration m we can treat the latest draws of the remaining parameters as the known parameter values. Therefore, we can evaluate numerically the conditional posterior density of ν on a grid of its admissible values. The support of ν is the positive part of the real line. However, a reasonable range for the values of ν in an application to asset returns could be (2, 30).¹³

Drawing from the Conditional Posterior Distribution of ν

Denote the equally-spaced grid of values for ν by $(\nu_1, \nu_2, \dots, \nu_J)$. We outline the steps for drawing from ν 's conditional posterior distribution at iteration m of the sampling algorithm. Denote the most recent draws of the remaining model parameters by $\theta_{-\nu}^{(m-1)}$. (Note that this notation is not entirely precise since some of the parameters might have been updated last during the m th iteration of the sampler but before ν .)

- Compute the value of ν 's posterior kernel (the exponential of the expression in (33)) at each of the grid nodes and denote the resultant vector by

$$p(\nu) = (p(\nu_1), p(\nu_2), \dots, p(\nu_J)) \tag{34}$$

- Normalize $p(\nu)$ by dividing each vector component in (34) by the quantity $\sum_{j=1}^J p(\nu_j)(\nu_2 - \nu_1)$. For convenience of notation, let us redefine $p(\nu)$ to denote the vector of (normalized) posterior density values at each node of ν 's grid.
- Compute the empirical cumulative distribution function (CDF),

$$F(\nu) = \left(p(\nu_1), \sum_{j=1}^2 p(\nu_j), \dots, \sum_{j=1}^J p(\nu_j) \right) \tag{35}$$

If the grid is adequate, the first element of $F(\nu)$ should be nearly 0, while the last element of $F(\nu)$ nearly 1.

- Draw an observation from the uniform distribution ($U[0, 1]$) and denote it by u .
- Find the element of $F(\nu)$ closest to u without exceeding it.
- The grid node corresponding to the value of $F(\nu)$ in the previous step is the draw of ν from its posterior distribution.

The method above of obtaining a draw from ν 's distribution using its CDF is called the CDF inversion method.

Constructing an adequate grid is the key to efficient sampling from ν 's posterior. Since the gridy Gibbs sampling procedure relies on multiple evaluations of the posterior kernel, two desired characteristics of an adequate grid are short length and coverage of the parameter support where the posterior distribution has positive probability mass. A simple example illustrates this point. Suppose that for a given sample of observed data, the likely values of ν are in the interval (2, 15). Suppose further that we construct an equally-spaced grid of length 30, with nodes on each integer from 2 to 30. The value of the posterior kernel at the nodes corresponding to ν equal to 16 and above would be only marginally different from zero. The posterior kernel evaluations at those nodes should be avoided, if possible.

If no prior intuition exists about what the likely parameter values are, one could employ a variable grid instead of a fixed grid. At each iteration of the sampling algorithm one must analyze the distribution of posterior mass and adjust the grid, so that the majority of the grid nodes are placed in the interval of greatest probability mass. Automating this process could involve some computational effort.

KEY POINTS

- The unconditional density of the return in GARCH models is not available in closed form. Therefore, the likelihood function of the GARCH parameters is expressed as a product of the return's conditional density in each period.
- In the Bayesian setting, estimation of GARCH models is performed numerically.
- Posterior numerical simulations are facilitated if the scale mixture of normal distributions representation is adopted for the Student's t distribution.
- Markov-switching GARCH models provide maximal flexibility in modeling transitions of the volatility dynamics across regimes.

- Transitions among regimes are governed by an unobserved state variable.
- In posterior simulations, the whole path of regimes, governed by the state variable, is simulated together with the model parameters.

NOTES

1. To see that, notice that when F_t is defined as an information set consisting of lagged asset returns, $F_1 = F_0 \cup r_1$, $F_2 = F_1 \cup r_2$, etc.
2. For a discussion of numerical estimation methods, see, for example, Rachev, et al. (2008). See also Geweke (1989) for an application of importance sampling to the estimation of ARCH models.
3. It is possible to assert a prior distribution for ω , α , and β defined on the whole real line, for example, a normal distribution. To respect the positivity constraints on the parameters, such a prior would have to be truncated at the lower bound of the parameters' range. In practice, however, the constraints could also be enforced during the posterior simulation as explained further below. Alternatively, one could assert such a prior without enforcing constraints, after transforming ω , α , and β by taking their logarithms (their ranges then become the whole real line).
4. Many heavy-tailed distributions can be represented as (mean-) scale mixtures of normal distributions. Such representations make estimation based on numerical, iterative procedures easier. See, for example, Fernandez and Steel (2000) for a discussion of the Bayesian treatment of regression analysis with mixtures of normals. In continuous time, the mean and scale mixture of normals models lead to the so-called subordinated processes, widely used in mathematical and empirical finance. Rachev and Mitnik (2000) offer an extensive treatment of subordinated processes.
5. This is an example of the technique known as "data augmentation." It consists of introducing latent (unobserved) variables to help construct efficient simulation algorithms. For a (technical) review of data augmentation, see, for example, van Dyk and Meng (2001).
6. Recall that we assume that $\mu_t = 0$.
7. The Hessian matrix is the matrix of second derivatives. According to a fundamental result in maximum likelihood theory, the maximum likelihood estimator's distribution is asymptotically normal, with covariance matrix—the negative inverse Hessian matrix, evaluated at the maximum likelihood estimate. Usually, the Hessian is provided as a "by-product" of numerical optimization routines for finding the maximum-likelihood estimate. See, for example, Rachev, et al. (2008) for additional details.
8. It is certainly possible to introduce (test for) a deterministic permanent shift in a model parameter into the regime-switching model. For example, Kim and Nelson (1999) apply such a model to a Bayesian investigation of business cycle fluctuations. See also Carlin, Gelfand, and Smith (1992). Wang and Zivot (2000) consider Bayesian estimation of a heteroskedastic model with structural breaks only. The variance in that investigation, however, does not evolve according to an ARCH-type process.
9. See also Haas, Mitnik, and Paoletta (2004), Klaassen (1998), Francq and Zakoian (2001), and Ghysels, McCulloch, and Tsay (1998), among others.
10. The beta distribution is the conjugate distribution for the probability parameter in a binomial experiment.
11. A K -dimensional random variable $\mathbf{p} = (p_1, p_2, \dots, p_K)$, where $p_k \geq 0$ and $\sum_{k=1}^K p_k = 1$, distributed with a Dirichlet distribution with parameters $\mathbf{a} = (a_1, a_2, \dots, a_K)$, $a_i > 0, i = 1, \dots, K$,

has a density function

$$f(\mathbf{p} | \mathbf{a}) = \frac{\Gamma(\sum_{k=1}^K a_k)}{\prod_{k=1}^K \Gamma(a_k)} \prod_{k=1}^K p_k^{a_k-1}$$

where Γ is the gamma function. The mean and the variance of the Dirichlet distribution are given, respectively, by $E(p_k) = \frac{a_k}{a_0}$ and $\text{var}(p_k) = \frac{a_k(a_0 - a_k)}{a_0^2(a_0 + 1)}$, where $a_0 = \sum_{j=1}^K a_j$. The Dirichlet distribution is the conjugate prior distribution for the parameters of the multinomial distribution. As can be seen in our discussion on the MS GARCH (1,1) estimation, the distribution of the Markov chain, $\{S_t\}_{t=1}^T$, is, in fact, a multinomial distribution.

12. See, for example, Anderson (2003).
13. This is the typical range of the degrees-of-freedom parameter of a Student's t distribution fitted to return data. The higher the data frequency is, the more heavy-tailed returns are and the lower the value of the degrees-of-freedom parameter.

REFERENCES

- Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis*. Hoboken, NJ: John Wiley & Sons.
- Bauwens, L., and Lubrano, M. (1998). Bayesian inference on GARCH models using the Gibbs sampler. *Econometrics Journal* 1.
- Cai, J. (1994). A Markov model of switching-regime ARCH. *Journal of Business and Economic Statistics* 12(3): 309–316.
- Carlin, B., Gelfand, A., and Smith, A. (1992). Hierarchical Bayesian analysis of change point problems. *Applied Statistics* 41: 389–405.
- Francq, C., and Zakoian, J.-M. (2001). Stationarity of multivariate Markov-switching ARMA models. *Journal of Econometrics* 102: 339–364.
- Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* 57(6): 1317–1339.
- Geweke, J. (1993a). Bayesian treatment of the independent Student's t linear model. *Journal of Applied Econometrics* 8, Supplement: *Special Issue on Econometric Inference Using Simulation Techniques*, S19–S40.
- Geweke, J. (1993b). Priors for macroeconomic time series and their application, Working Paper 64. Federal Reserve Bank of Minneapolis.
- Ghysels, E., McCulloch, R., and Tsay, R. (1998). Bayesian inference for periodic regime-switching models. *Journal of Applied Econometrics* 13(2): 129–143.
- Haas, M., Mittnik, S., and Paoletta, M. (2004). A new approach to Markov-switching GARCH models. *Journal of Financial Econometrics* 2(4): 493–530.
- Hamilton, J. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57: 357–384.
- Hamilton, J., and Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics* 64: 307–333.
- Henneke, J., Rachev, S., Fabozzi, F., and Nikolov, M. (2011). MCMC-based estimation of Markov-switching ARMA-GARCH models. *Applied Economics* 43(3): 259–271.
- Kim, C., and Nelson, C. (1999). Has the U.S. economy become more stable? A Bayesian approach based on a Markov-switching model of the business cycle. *The Review of Economics and Statistics* 81(4): 608–616.
- Klaassen, F. (1998). Improving GARCH volatility forecasts. Social Science Research Network. Center Discussion Paper Series No 1998-52. Available at <http://papers.ssrn.com>.
- Rachev, S., Hsu, J., Bagasheva, B., and Fabozzi, F. (2008). *Bayesian Methods in Finance*. Hoboken, NJ: Wiley & Sons.
- Rachev, S., and Mittnik, S. (2000). *Stable Paretian Models in Finance*. New York: John Wiley & Sons.
- Ritter, C., and Tanner, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the griddy-Gibbs sampler. *Journal of the American Statistical Association* 87(419).
- van Dyk, D., and Meng, X. (2001). The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics* 10.
- Wang, J., and Zivot, E. (2000). A Bayesian time series model of multiple structural changes in level, trend, and variance. *Journal of Business and Economic Statistics* 18(3): 374–386.

Bayesian Techniques and the Black-Litterman Model

PETTER N. KOLM, PhD

Director of the Mathematics in Finance Masters Program and Clinical Associate Professor,
Courant Institute of Mathematical Sciences, New York University

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

SERGIO M. FOCARDI, PhD

Partner, The Intertek Group

Abstract: Investment policies constructed using inferior estimates, such as sample means and sample covariance matrices, typically perform very poorly in practice. Besides introducing spurious changes in portfolio weights each time the portfolio is rebalanced, this undesirable property also results in unnecessary turnover and increased transaction costs. These phenomena are not necessarily a sign that portfolio optimization does not work, but rather that the modern portfolio theory framework is very sensitive to the accuracy of inputs. There are different ways to address this issue. On the estimation side, one can try to produce more *robust estimates* of the input parameters for the optimization problems. This is most often achieved by using estimators that are less sensitive to outliers, and possibly, other sampling errors, such as Bayesian and shrinkage estimators. On the modeling side, one can constrain portfolio weights, use portfolio *resampling*, or apply robust or stochastic optimization techniques to specify scenarios or ranges of values for parameters estimated from data, thus incorporating uncertainty into the optimization process itself.

In this entry, we provide a general overview of some of the common problems encountered in mean-variance optimization before we turn our attention to shrinkage estimators for expected returns and the covariance matrix. Within the context of *Bayesian estimation*, we focus on the Black-Litterman model (see Black and Litterman, 1992). We derive

the model using so-called mixed estimation from classical econometrics. Introducing a simple cross-sectional momentum strategy, we then show how one can combine this strategy with market equilibrium using the Black-Litterman model in the mean-variance framework to rebalance the portfolio on a monthly basis.

PRACTICAL PROBLEMS ENCOUNTERED IN MEAN-VARIANCE OPTIMIZATION

The simplicity and the intuitive appeal of portfolio construction using modern portfolio theory have attracted significant attention both in academia and in practice. Yet, despite considerable effort, it took many years until portfolio managers started using modern portfolio theory for managing real money. Unfortunately, in real world applications there are many problems with it, and portfolio optimization is still considered by many practitioners to be difficult to apply. In this section we consider some of the typical problems encountered in mean-variance optimization. In particular, we elaborate on: (1) the sensitivity to estimation error; (2) the effects of uncertainty in the inputs in the optimization process; and (3) the large data requirement necessary for accurately estimating the inputs for the portfolio optimization framework. We start by considering an example illustrating the effect of estimation error.

Example: The True, Estimated, and Actual Efficient Frontiers

Broadie introduced the terms true frontier, estimated frontier, and actual frontier to refer to the efficient frontiers computed using the true expected returns (unobservable), estimated expected returns, and true expected returns of the portfolios on the estimated frontier, respectively.¹ In this example, we refer to the frontier computed using the true, but unknown, expected returns as the true frontier. Similarly, we refer to the frontier computed using estimates of the expected returns and the true covariance matrix as the estimated frontier. Finally, we define the actual frontier as follows: We take the portfolios on the estimated frontier and then calculate their expected returns using the true expected returns. Since we are using the true covariance matrix, the variance of a port-

folio on the estimated frontier is the same as the variance on the actual frontier.

From these definitions, we observe that the actual frontier will always lie below the true frontier. The estimated frontier can lie anywhere with respect to the other frontiers. However, if the errors in the expected return estimates have a mean of zero, then the estimated frontier will lie above the true frontier with extremely high probability, particularly when the investment universe is large. We look at two cases considered by Ceria and Stubbs:²

1. Using the covariance matrix and expected return vector from Idzorek (2005), they randomly generate a time series of normally distributed returns and compute the average to use as estimates of expected returns. Using the expected-return estimate calculated in this fashion and the true covariance matrix, they generate an estimated efficient frontier of risk versus expected return where the portfolios were subject to no-shorting constraints and the standard budget constraint that the sum of portfolio weights is one. Similarly, Ceria and Stubbs compute the true efficient frontier using the original covariance matrix and expected return vector. Finally, they construct the actual frontier by computing the expected return and risk of the portfolios on the estimated frontier with the true covariance and expected return values. These three frontiers are illustrated in Figure 1.
2. Using the same estimate of expected returns, Ceria and Stubbs also generate risk versus expected return where active holdings of the assets are constrained to be $\pm 3\%$ of the benchmark holding of each asset. These frontiers are illustrated in Figure 2.

We observe that the estimated frontiers significantly overestimate the expected return for any risk level in both types of frontiers. More importantly, we note that the actual frontier lies far below the true frontier in both cases. This shows that the optimal mean-variance

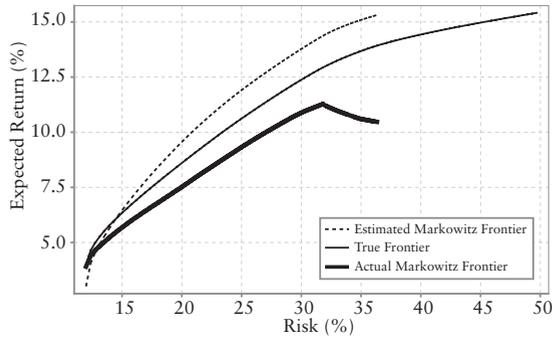


Figure 1 Markowitz Efficient Frontiers
 Source: Figure 2 in Ceria and Stubbs (2005, p. 6).
 Reprinted with the permission of Axioma, Inc.

portfolio is not necessarily a good portfolio; that is, it is not mean-variance efficient. Since the true expected return is not observable, we do not know how far the actual expected return may be from the expected return of the mean-variance optimal portfolio, and we end up holding an inferior portfolio.

Sensitivity to Estimation Error

In a portfolio optimization context, securities with large expected returns and low standard deviations will be overweighted and conversely, securities with low expected re-

turns and high standard deviations will be underweighted. Therefore, large estimation errors in expected returns and/or variances/covariances introduce errors in the optimized portfolio weights. For this reason, people often cynically refer to optimizers as error maximizers.

Uncertainty from estimation error in expected returns tends to have more influence than in the covariance matrix in a mean-variance optimization.³ The relative importance depends on the investor’s risk aversion, but as a general rule of thumb, errors in the expected returns are about 10 times more important than errors in the covariance matrix, and errors in the variances are about twice as important as errors in the covariances.⁴ As the risk tolerance increases, the relative impact of estimation errors in the expected returns becomes even more important. Conversely, as the risk tolerance decreases, the impact of errors in expected returns relative to errors in the covariance matrix becomes smaller. From this simple rule, it follows that the major focus should be on providing good estimates for the expected returns, followed by the variances. In this entry we discuss shrinkage techniques and the *Black-Litterman model* in order to mitigate estimation errors.

Constraining Portfolio Weights

Several studies have shown that the inclusion of constraints in the mean-variance optimization problem leads to better out-of-sample performance.⁵ Practitioners often use no short-selling constraints or upper and lower bounds for each security to avoid overconcentration in a few assets. Gupta and Eichhorn (1998) suggest that constraining portfolio weights may also assist in containing volatility, increase realized efficiency, and decrease downside risk or shortfall probability.

Jagannathan and Ma (2003) provide a theoretical justification for these observations. Specifically, they show that the no short-selling constraints are equivalent to reducing the

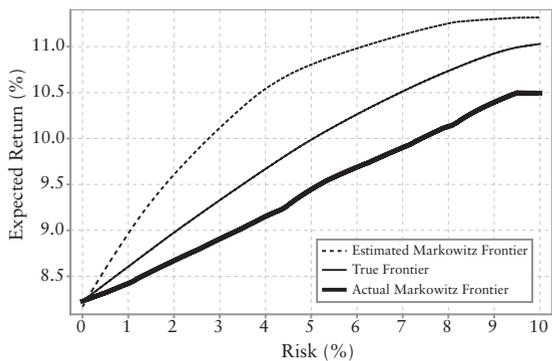


Figure 2 Markowitz Benchmark-Relative Efficient Frontiers
 Source: Figure 3 in Ceria and Stubbs (2005, p. 7).
 Reprinted with the permission of Axioma, Inc.

estimated asset covariances, whereas upper bounds are equivalent to increasing the corresponding covariances. For example, stocks that have high covariance with other stocks tend to receive negative portfolio weights. Therefore, when their covariance is decreased (which is equivalent to the effect of imposing no short-selling constraints), these negative weights disappear. Similarly, stocks that have low covariances with other stocks tend to get overweighted. Hence, by increasing the corresponding covariances the impact of these overweighted stocks decreases.

Furthermore, Monte Carlo experiments performed by Jagannathan and Ma indicate that when no-short-sell constraints are imposed, the sample covariance matrix has about the same performance (as measured by the global minimum variance (GMV) portfolio) as a covariance matrix estimator constructed from a factor structure.

Care needs to be taken when imposing constraints for robustness and stability purposes. For example, if the constraints used are too tight, they will completely determine the portfolio allocation—not the forecasts.

Instead of providing ad hoc upper and lower bounds on each security, as proposed by Bouchaud, Potters, and Aguilar (1997), one can use so-called diversification indicators that measure the concentration of the portfolio. These diversification indicators can be used as constraints in the portfolio construction phase to limit the concentration to individual securities. The authors demonstrate that these indicators are related to the information content of the portfolio in the sense of information theory.⁶ For example, a very concentrated portfolio corresponds to a large information content (as we would only choose a very concentrated allocation if our information about future price fluctuations is perfect), whereas an equally weighted portfolio would indicate low information content (as we would not put “all the eggs in one basket” if our information about future price fluctuations is poor).

Importance of Sensitivity Analysis

In practice, in order to minimize dramatic changes due to estimation error, it is advisable to perform sensitivity analysis. For example, one can study the results of small changes or perturbations to the inputs from an efficient portfolio selected from a mean-variance optimization. If the portfolio calculated from the perturbed inputs drastically differs from the first one, this might indicate a problem. The perturbation can also be performed on a security by security basis in order to identify those securities that are the most sensitive. The objective of this sensitivity analysis is to identify a set of security weights that will be close to efficient under several different sets of plausible inputs.

Issues with Highly Correlated Assets

The inclusion of highly correlated securities is another major cause for instability in the *mean-variance optimization* framework. For example, high correlation coefficients among common asset classes are one reason why real estate is popular in optimized portfolios. Real estate is one of the few asset classes that has a low correlation with other common asset classes. But real estate in general does not have the liquidity necessary in order to implement these portfolios and may therefore fail to deliver the return promised by the real estate indexes.

The problem of high correlations typically becomes worse when the correlation matrix is estimated from historical data. Specifically, when the correlation matrix is estimated over a slightly different period, correlations may change, but the impact on the new portfolio weights may be drastic. In these situations, it may be a good idea to resort to a *shrinkage estimator* or a factor model to model covariances and correlations.

Incorporating Uncertainty in the Inputs into the Portfolio Allocation Process

In the classical mean-variance optimization problem, the expected returns and the

covariance matrix of returns are uncertain and have to be estimated. After the estimation of these quantities, the portfolio optimization problem is solved as a deterministic problem—completely ignoring the uncertainty in the inputs. However, it makes sense for the uncertainty of expected returns and risk to enter into the optimization process, thus creating a more realistic model. Using point estimates of the expected returns and the covariance matrix of returns, and treating them as error-free in portfolio allocation does not necessarily correspond to prudent investor behavior.

The investor would probably be more comfortable choosing a portfolio that would perform well under a number of different scenarios, thereby also attaining some protection from estimation risk and model risk. Obviously, to have some insurance in the event of less likely but more extreme cases (e.g., scenarios that are highly unlikely under the assumption that returns are normally distributed), the investor must be willing to give up some of the upside that would result under the more likely scenarios. Such an investor seeks a robust portfolio, that is, a portfolio that is assured against some worst-case model misspecification. The estimation process can be improved through robust statistical techniques such as shrinkage and Bayesian estimators discussed later in this entry. However, jointly considering estimation risk and model risk in the financial decision-making process is becoming more important.

The estimation process frequently does not deliver a point forecast (that is, one single number), but a full distribution of expected returns. Recent approaches attempt to integrate estimation risk into the mean-variance framework by using the expected return distribution in the optimization. A simple approach is to sample from the return distribution and average the resulting portfolios (Monte Carlo approach).⁷ However, as a mean-variance problem has to be solved for each draw, this is computationally intensive for larger portfolios. In addition, the

averaging does not guarantee that the resulting portfolio weights will satisfy all constraints.

Introduced in the late 1990s by Ben-Tal and Nemirovski (1998, 1999) and El Ghaoui and Lebret (1997) the *robust optimization* framework is computationally more efficient than the Monte Carlo approach. This development in optimization technology allows for efficiently solving the robust version of the mean-variance optimization problem in about the same time as the classical mean-variance optimization problem. The technique explicitly uses the distribution from the estimation process to find a robust portfolio in one single optimization. It thereby incorporates uncertainties of inputs into a deterministic framework. The classical portfolio optimization formulations such as the mean-variance portfolio selection problem, the maximum Sharpe ratio portfolio problem, and the value-at-risk (VaR) portfolio problem all have robust counterparts that can be solved in roughly the same amount of time as the original problem.⁸

Large Data Requirements

In classical mean-variance optimization, we need to provide estimates of the expected returns and covariances of all the securities in the investment universe considered. Typically, however, portfolio managers have reliable return forecasts for only a small subset of these assets. This is probably one of the major reasons why the mean-variance framework has not been adopted by practitioners in general. It is simply unreasonable for the portfolio manager to produce good estimates of all the inputs required in classical portfolio theory.

We will see later in this entry that the Black-Litterman model provides a remedy in that it blends any views (this could be a forecast on just one or a few securities, or all of them) the investor might have with the market equilibrium. When no views are present, the resulting Black-Litterman expected returns are

just the expected returns consistent with the market equilibrium. Conversely, when the investor has views on some of the assets, the resulting expected returns deviate from market equilibrium.

SHRINKAGE ESTIMATION

It is well known since the seminal work by Stein (1956) that biased estimators often yield better parameter estimates than their generally preferred unbiased counterparts. In particular, it can be shown that if we consider the problem of estimating the mean of an N -dimensional multivariate normal variable ($N > 2$), $X \in N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with known covariance matrix $\boldsymbol{\Sigma}$, the sample mean $\hat{\boldsymbol{\mu}}$ is not the best estimator of the population mean $\boldsymbol{\mu}$ in terms of the quadratic loss function

$$L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})$$

For example, the so-called James-Stein shrinkage estimator

$$\hat{\boldsymbol{\mu}}_{JS} = (1 - w)\hat{\boldsymbol{\mu}} + w\boldsymbol{\mu}_0\boldsymbol{\iota}$$

has a lower quadratic loss than the sample mean, where

$$w = \min\left(1, \frac{N - 2}{T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\boldsymbol{\iota})' \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0\boldsymbol{\iota})}\right)$$

and $\boldsymbol{\iota} = [1, 1, \dots, 1]'$. Moreover, T is the number of observations, and $\boldsymbol{\mu}_0$ is an arbitrary number. The vector $\boldsymbol{\mu}_0\boldsymbol{\iota}$ and the weight w are referred to as the shrinkage target and the shrinkage intensity (or shrinkage factor), respectively. Although there are some choices of $\boldsymbol{\mu}_0$ that are better than others, what is surprising with this result is that it could be any number! This fact is referred to as the Stein paradox.

In effect, shrinkage is a form of averaging different estimators. The shrinkage estimator typically consists of three components: (1) an estimator with little or no structure (like the sample mean above); (2) an estimator with a lot of structure (the shrinkage target); and (3) the shrinkage intensity. The shrinkage target is

chosen with the following two requirements in mind. First, it should have only a small number of free parameters (robust and with a lot of structure). Second, it should have some of the basic properties in common with the unknown quantity being estimated. The shrinkage intensity can be chosen based on theoretical properties or simply by numerical simulation.

Probably the most well-known shrinkage estimator⁹ used to estimate expected returns in the financial literature is the one proposed by Jorion (1986) where the shrinkage target is given by $\boldsymbol{\mu}_g\boldsymbol{\iota}$ with

$$\boldsymbol{\mu}_g = \frac{\boldsymbol{\iota}' \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\mu}}}{\boldsymbol{\iota}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\iota}}$$

and

$$w = \frac{N + 2}{N + 2 + T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_g\boldsymbol{\iota})' \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_g\boldsymbol{\iota})}$$

We note that $\boldsymbol{\mu}_g$ is the return on the GMV portfolio. Several studies document that for the mean-variance framework: (1) the variability in the portfolio weights from one period to the next decrease; and (2) the out-of-sample risk-adjusted performance improves significantly when using a shrinkage estimator as compared to the sample mean.¹⁰

We can also apply the shrinkage technique for covariance matrix estimation. This involves shrinking an unstructured covariance estimator toward a more structured covariance estimator. Typically the structured covariance estimator only has a few degrees of freedom (only a few nonzero eigenvalues) as motivated by random matrix theory.

For example, as shrinkage targets, Ledoit and Wolf (2003, 2004) suggest using the covariance matrix that follows from the single-factor model developed by Sharpe (1963) or the constant correlation covariance matrix.¹¹ In practice the single-factor model and the constant correlation model yield similar results, but the constant correlation model is much easier to implement. In the case of the constant correlation model, the shrinkage estimator for the covariance matrix

takes the form

$$\hat{\Sigma}_{LW} = w\hat{\Sigma}_{CC} + (1 - w)\hat{\Sigma}$$

where $\hat{\Sigma}$ is the sample covariance matrix, and $\hat{\Sigma}_{CC}$ is the sample covariance matrix with constant correlation. The sample covariance matrix with constant correlation is computed as follows.

First, we decompose the sample covariance matrix according to

$$\hat{\Sigma} = \Lambda C \Lambda'$$

where Λ is a diagonal matrix of the volatilities of returns and C is the sample correlation matrix, that is,

$$C = \begin{bmatrix} 1 & \hat{\rho}_{12} & \cdots & \hat{\rho}_{1N} \\ \hat{\rho}_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \hat{\rho}_{N-1N} \\ \hat{\rho}_{N1} & \cdots & \hat{\rho}_{NN-1} & 1 \end{bmatrix}$$

Second, we replace the sample correlation matrix with the constant correlation matrix

$$C_{CC} = \begin{bmatrix} 1 & \hat{\rho} & \cdots & \hat{\rho} \\ \hat{\rho} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \hat{\rho} \\ \hat{\rho} & \cdots & \hat{\rho} & 1 \end{bmatrix}$$

where $\hat{\rho}$ is the average of all the sample correlations, in other words

$$\hat{\rho} = \frac{2}{(N-1)N} \sum_{i=1}^N \sum_{j=i+1}^N \hat{\rho}_{ij}$$

The optimal shrinkage intensity can be shown to be proportional to a constant divided by the length of the history, T .¹²

Ledoit and Wolf (2003, 2004) compare the empirical out-of-sample performance of their shrinkage covariance matrix estimators with other covariance matrix estimators, such as the sample covariance matrix, a statistical factor model based on the first five principal components, and a factor model based on the 48 industry factors as defined by Fama and French (1997). The results indicate that when it comes

to computing a GMV portfolio, their shrinkage estimators are superior compared to the others tested, with the constant correlation shrinkage estimator coming out slightly ahead. Interestingly enough, it turns out that the shrinkage intensity for the single-factor model (the shrinkage intensity for the constant coefficient model is not reported) is fairly constant throughout time with a value around 0.8. This suggests that there is about four times as much estimation error present in the sample covariance matrix as there is bias in the single-factor covariance matrix.

THE BLACK-LITTERMAN MODEL

In the Black-Litterman model an estimate of future expected returns is based on combining market equilibrium (e.g., the CAPM equilibrium) with an investor's views. As we will see, the Black-Litterman expected return is a shrinkage estimator where market equilibrium is the shrinkage target and the shrinkage intensity is determined by the portfolio manager's confidence in the model inputs. We will make this statement precise later in this section. Such views are expressed as absolute or relative deviations from equilibrium together with confidence levels of the views (as measured by the standard deviation of the views).

The Black-Litterman expected return is calculated as a weighted average of the market equilibrium and the investor's views. The weights depend on (1) the volatility of each asset and its correlations with the other assets and (2) the degree of confidence in each forecast. The resulting expected return, which is the mean of the posterior distribution, is then used as input in the portfolio optimization process. Portfolio weights computed in this fashion tend to be more intuitive and less sensitive to small changes in the original inputs (i.e., forecasts of market equilibrium, investor's views, and the covariance matrix).

The Black-Litterman model can be interpreted as a Bayesian model. Named after the English mathematician Thomas Bayes, the Bayesian approach is based on the subjective interpretation of probability. A probability distribution is used to represent an investor's belief on the probability that a specific event will actually occur. This probability distribution, called the prior distribution, reflects an investor's knowledge about the probability before any data are observed. After more information is provided (e.g., data observed), the investor's opinions about the probability might change. Bayes' rule is the formula for computing the new probability distribution, called the posterior distribution. The posterior distribution is based on knowledge of the prior probability distribution plus the new data. A posterior distribution of expected return is derived by combining the forecast from the empirical data with a prior distribution.

The ability to incorporate exogenous insight, such as a portfolio manager's judgment, into formal models is important: Such insight might be the most valuable input used by the model. The Bayesian framework allows forecasting systems to use such external information sources and subjective interventions (i.e., modification of the model due to judgment) in addition to traditional information sources such as market data and proprietary data.

Because portfolio managers might not be willing to give up control to a black box, incorporating exogenous insights into formal models through Bayesian techniques is one way of giving the portfolio manager better control in a quantitative framework. Forecasts are represented through probability distributions that can be modified or adjusted to incorporate other sources of information deemed relevant. The only restriction is that such additional information (i.e., the investor's views) be combined with the existing model through the laws of probability. In effect, incorporating Bayesian views into a model allows one to rationalize subjectivity within a formal, quanti-

tative framework. "[T]he rational investor is a Bayesian," as Markowitz noted (1987, p. 57).

Derivation of the Black-Litterman Model

The basic feature of the Black-Litterman model that we discuss in this and the following sections is that it combines an investor's views with the market equilibrium. Let us understand what this statement implies. In the classical mean-variance optimization framework an investor is required to provide estimates of the expected returns and covariances of all the securities in the investment universe considered. This is of course a humongous task, given the number of securities available today. Portfolio and investment managers are very unlikely to have a detailed understanding of all the securities, companies, industries, and sectors that they have at their disposal. Typically, most of them have a specific area of expertise that they focus on in order to achieve superior returns.

This is probably one of the major reasons why the mean-variance framework has not been adopted among practitioners in general. It is simply unrealistic for the portfolio manager to produce reasonable estimates (besides the additional problems of estimation error) of the inputs required in classical portfolio theory.

Furthermore, many trading strategies used today cannot easily be turned into forecasts of expected returns and covariances. In particular, not all trading strategies produce views on absolute return, but rather just provide relative rankings of securities that are predicted to outperform/underperform other securities. For example, considering two stocks, A and B, instead of the absolute view, "the one-month expected return on A and B are 1.2% and 1.7% with a standard deviation of 5% and 5.5%, respectively," a relative view may be of the form "B will outperform A with half a percent over the next month" or simply "B will outperform

A over the next month.” Clearly, it is not an easy task to translate any of these relative views into the inputs required for the modern portfolio theoretical framework. We now walk through and illustrate the usage of the Black-Litterman model in three simple steps.

Step 1: Basic Assumptions and Starting Point

One of the basic assumptions underlying the Black-Litterman model is that the expected return of a security should be consistent with market equilibrium unless the investor has a specific view on the security. In other words, an investor who does not have any views on the market should hold the market.¹³

Our starting point is the CAPM model:

$$E(R_i) - R_f = \beta_i(E(R_M) - R_f)$$

where $E(R_i)$, $E(R_M)$, and R_f are the expected return on security i , the expected return on the market portfolio, and the risk-free rate, respectively. Furthermore,

$$\beta_i = \frac{\text{cov}(R_i, R_M)}{\sigma_M^2}$$

where σ_M^2 is the variance of the market portfolio. Let us denote by $\mathbf{w}_b = (w_{b1}, \dots, w_{bN})'$ the market capitalization or benchmark weights, so that with an asset universe of N securities¹⁴ the return on the market can be written as

$$R_M = \sum_{j=1}^N w_{bj} R_j$$

Then by the CAPM, the expected excess return on asset i , $\Pi_i = E(R_i) - R_f$, becomes

$$\begin{aligned} \Pi_i &= \beta_i(E(R_M) - R_f) \\ &= \frac{\text{cov}(R_i, R_M)}{\sigma_M^2}(E(R_M) - R_f) \\ &= \frac{E(R_M) - R_f}{\sigma_M^2} \sum_{j=1}^N \text{cov}(R_i, R_j) w_{bj} \end{aligned}$$

We can also express this in matrix-vector form as

$$\mathbf{\Pi} = \delta \mathbf{\Sigma} \mathbf{w}$$

where we define the market price of risk as

$$\delta = \frac{E(R_M) - R_f}{\sigma_M^2}$$

the expected excess return vector

$$\mathbf{\Pi} = \begin{bmatrix} \Pi_1 \\ \vdots \\ \Pi_N \end{bmatrix}$$

and the covariance matrix of returns

$$\mathbf{\Sigma} = \begin{bmatrix} \text{cov}(R_1, R_1) & \cdots & \text{cov}(R_1, R_N) \\ \vdots & \ddots & \vdots \\ \text{cov}(R_N, R_1) & \cdots & \text{cov}(R_N, R_N) \end{bmatrix}$$

The true expected returns $\boldsymbol{\mu}$ of the securities are unknown. However, we assume that our previous equilibrium model serves as a reasonable estimate of the true expected returns in the sense that

$$\mathbf{\Pi} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_\Pi, \boldsymbol{\varepsilon}_\Pi \sim N(0, \tau \mathbf{\Sigma})$$

for some small parameter $\tau \ll 1$. We can think about $\tau \mathbf{\Sigma}$ as our confidence in how well we can estimate the equilibrium of expected returns. In other words, a small τ implies a high confidence in our equilibrium estimates and vice versa.

According to portfolio theory, because the market portfolio is on the efficient frontier, as a consequence of the CAPM an investor will be holding a portfolio consisting of the market portfolio and a risk-free instrument earning the risk-free rate. But let us now see what happens if an investor has a particular view on some of the securities.

Step 2: Expressing an Investor's Views

Formally, K views in the Black-Litterman model are expressed as a K -dimensional vector \mathbf{q} with

$$\mathbf{q} = \mathbf{P} \boldsymbol{\mu} + \boldsymbol{\varepsilon}_q, \boldsymbol{\varepsilon}_q \sim N(0, \boldsymbol{\Omega})$$

where \mathbf{P} is a $K \times N$ matrix (explained in the following example) and $\boldsymbol{\Omega}$ is a $K \times K$ matrix expressing the confidence in the views. In order to understand this mathematical specification better, let us take a look at an example.

Let us assume that the asset universe that we consider has five stocks ($N = 5$) and that an investor has the following two views:

1. Stock 1 will have a return of 1.5%.
2. Stock 3 will outperform Stock 2 by 4%.

We recognize that the first view is an absolute view whereas the second one is a relative view. Mathematically, we express the two views together as

$$\begin{bmatrix} 1.5\% \\ 4\% \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

The first row of the \mathbf{P} matrix represents the first view, and similarly, the second row describes the second view. In this example, we chose the weights of the second view such that they add up to zero, but other weighting schemes are also possible. For instance, the weights could also be chosen as some scaling factor times one over the market capitalizations of the stock, some scaling factor times one over the stock price, or other variations thereof. We come back to these issues later in this section when we discuss how to incorporate time-series-based strategies and cross-sectional ranking strategies.

We also remark at this point that the error terms ε_1 , ε_2 do not explicitly enter into the Black-Litterman model—but their variances do. Quite simply, these are just the variances of the different views. Although in some instances they are directly available as a by-product of the view or the strategy, in other cases they need to be estimated separately. For example,

$$\mathbf{\Omega} = \begin{bmatrix} 1\%^2 & 0 \\ 0 & 1\%^2 \end{bmatrix}$$

corresponds to a higher confidence in the views, and conversely,

$$\mathbf{\Omega} = \begin{bmatrix} 5\%^2 & 0 \\ 0 & 7\%^2 \end{bmatrix}$$

represents a much lower confidence in the views. We discuss a few different approaches in choosing the confidence levels below. The off-diagonal elements of $\mathbf{\Omega}$ are typically set to zero. The reason for this is that the error terms of the individual views are most often assumed to be independent of one another.

Step 3: Combining an Investor's Views with Market Equilibrium

Having specified the market equilibrium and an investor's views separately, we are now ready to combine the two. There are two different, but equivalent, approaches that can be used to arrive at the Black-Litterman model. We will describe a derivation that relies upon standard econometrical techniques, in particular, the so-called mixed estimation technique described by Theil (1971). The approach based on Bayesian statistics has been explained in some detail by Satchell and Scowcroft (2000).

Let us first recall the specification of market equilibrium

$$\mathbf{\Pi} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_{\Pi}, \boldsymbol{\varepsilon}_{\Pi} \sim N(0, \tau \boldsymbol{\Sigma})$$

and the one for the investor's views

$$\mathbf{q} = \mathbf{P}\boldsymbol{\mu} + \boldsymbol{\varepsilon}_{\mathbf{q}}, \boldsymbol{\varepsilon}_{\mathbf{q}} \sim N(0, \boldsymbol{\Omega})$$

We can stack these two equations together in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(0, \mathbf{V})$$

where

$$\mathbf{y} = \begin{bmatrix} \mathbf{\Pi} \\ \mathbf{q} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{I} \\ \mathbf{P} \end{bmatrix}, \mathbf{V} = \begin{bmatrix} \tau \boldsymbol{\Sigma} & \\ & \boldsymbol{\Omega} \end{bmatrix}$$

with \mathbf{I} denoting the $N \times N$ identity matrix. We observe that this is just a standard linear model for the expected returns $\boldsymbol{\mu}$. Calculating the generalized least squares (GLS) estimator for $\boldsymbol{\mu}$, we

obtain

$$\begin{aligned}
 \hat{\boldsymbol{\mu}}_{BL} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\
 &= \left(\begin{bmatrix} \mathbf{I} & \mathbf{P}' \end{bmatrix} \begin{bmatrix} (\boldsymbol{\tau}\boldsymbol{\Sigma})^{-1} & \\ & \boldsymbol{\Omega}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \boldsymbol{\Pi} \end{bmatrix} \right)^{-1} \\
 &\quad \times \begin{bmatrix} \mathbf{I} & \mathbf{P}' \end{bmatrix} \begin{bmatrix} (\boldsymbol{\tau}\boldsymbol{\Sigma})^{-1} & \\ & \boldsymbol{\Omega}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Pi} \\ \mathbf{q} \end{bmatrix} \\
 &= \left(\begin{bmatrix} \mathbf{I} & \mathbf{P}' \end{bmatrix} \begin{bmatrix} (\boldsymbol{\tau}\boldsymbol{\Sigma})^{-1} \\ \boldsymbol{\Omega}^{-1}\mathbf{P}' \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{I} & \mathbf{P}' \end{bmatrix} \begin{bmatrix} (\boldsymbol{\tau}\boldsymbol{\Sigma})^{-1}\boldsymbol{\Pi} \\ \boldsymbol{\Omega}^{-1}\mathbf{q} \end{bmatrix} \\
 &= [(\boldsymbol{\tau}\boldsymbol{\Sigma})^{-1} + \mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{P}']^{-1} [(\boldsymbol{\tau}\boldsymbol{\Sigma})^{-1}\boldsymbol{\Pi} + \mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{q}]
 \end{aligned}$$

The last line in the above formula is the Black-Litterman expected returns that blend the market equilibrium with the investor's views.

Some Remarks and Observations

Following are some comments in order to provide a better intuitive understanding of the formula. We see that if the investor has no views (that is, $\mathbf{q} = \boldsymbol{\Omega} = 0$) or the confidence in the views is zero, then the Black-Litterman expected return becomes $\hat{\boldsymbol{\mu}}_{BL} = \boldsymbol{\Pi}$. Consequently, the investor will end up holding the market portfolio as predicted by the CAPM. In other words, the optimal portfolio in the absence of views is the defined market.

If we were to plug return targets of zero or use the available cash rates, for example, into an optimizer to represent the absence of views, the result would be an optimal portfolio that looks very much different from the market. The equilibrium returns are those forecasts that in the absence of any other views will produce an optimal portfolio equal to the market portfolio. Intuitively speaking, the equilibrium returns in the Black-Litterman model are used to center the optimal portfolio around the market portfolio.

By using $\mathbf{q} = \mathbf{P}\boldsymbol{\mu} + \boldsymbol{\varepsilon}_q$, we have that the investor's views alone imply the estimate of expected returns $\hat{\boldsymbol{\mu}} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{q}$. Since $\mathbf{P}(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}' = \mathbf{I}$ where \mathbf{I} is the identity matrix, we can rewrite the Black-Litterman expected returns in the form

$$\hat{\boldsymbol{\mu}}_{BL} = [(\boldsymbol{\tau}\boldsymbol{\Sigma})^{-1} + \mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{P}']^{-1} [(\boldsymbol{\tau}\boldsymbol{\Sigma})^{-1}\boldsymbol{\Pi} + \mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{P}\hat{\boldsymbol{\mu}}]$$

Now we see that the Black-Litterman expected return is a confidence weighted linear combination of market equilibrium $\boldsymbol{\Pi}$ and the expected return $\hat{\boldsymbol{\mu}}$ implied by the investor's views. The two weighting matrices are given by

$$\begin{aligned}
 \mathbf{w}_{\boldsymbol{\Pi}} &= [(\boldsymbol{\tau}\boldsymbol{\Sigma})^{-1} + \mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{P}']^{-1} (\boldsymbol{\tau}\boldsymbol{\Sigma})^{-1} \\
 \mathbf{w}_{\mathbf{q}} &= [(\boldsymbol{\tau}\boldsymbol{\Sigma})^{-1} + \mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{P}']^{-1} \mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{P}
 \end{aligned}$$

where

$$\mathbf{w}_{\boldsymbol{\Pi}} = \mathbf{w}_{\mathbf{q}} = \mathbf{I}$$

In particular, $(\boldsymbol{\tau}\boldsymbol{\Sigma})^{-1}$ and $\mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{P}$ represent the confidence we have in our estimates of the market equilibrium and the views, respectively. Therefore, if we have low confidence in the views, the resulting expected returns will be close to the ones implied by market equilibrium. Conversely, with higher confidence in the views, the resulting expected returns will deviate from the market equilibrium implied expected returns. We say that we tilt away from market equilibrium.

It is straightforward to show that the Black-Litterman expected returns can also be written in the form

$$\hat{\boldsymbol{\mu}}_{BL} = \boldsymbol{\Pi} + \boldsymbol{\tau}\boldsymbol{\Sigma}\mathbf{P}'(\boldsymbol{\Omega} + \boldsymbol{\tau}\mathbf{P}\boldsymbol{\Sigma}\mathbf{P}')^{-1}(\mathbf{q} - \mathbf{P}\boldsymbol{\Pi})$$

where we now immediately see that we tilt away from the equilibrium with a vector proportional to $\boldsymbol{\Sigma}\mathbf{P}'(\boldsymbol{\Omega} + \boldsymbol{\tau}\mathbf{P}\boldsymbol{\Sigma}\mathbf{P}')^{-1}(\mathbf{q} - \mathbf{P}\boldsymbol{\Pi})$.

We also mention that the Black-Litterman model can be derived as a solution to the following optimization problem:

$$\begin{aligned}
 \hat{\boldsymbol{\mu}}_{BL} &= \arg \min_{\boldsymbol{\mu}} \{ (\boldsymbol{\Pi} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\Pi} - \boldsymbol{\mu}) \\
 &\quad + \boldsymbol{\tau}(\mathbf{q} - \mathbf{P}\boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{q} - \mathbf{P}\boldsymbol{\mu}) \}
 \end{aligned}$$

From this formulation we see that $\hat{\boldsymbol{\mu}}_{BL}$ is chosen such that it is simultaneously as close to $\boldsymbol{\Pi}$, and $\mathbf{P}\boldsymbol{\mu}$ is as close to \mathbf{q} as possible. The distances are determined by $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Omega}^{-1}$. Furthermore, the relative importance of the equilibrium versus the views is determined by $\boldsymbol{\tau}$. For example, for $\boldsymbol{\tau}$ large the weight of the views is increased,

whereas for τ small the weight of the equilibrium is higher. Moreover, we also see that τ is a redundant parameter as it can be absorbed into Ω .

It is straightforward to calculate the variance of the Black-Litterman combined estimator of the expected returns by the standard sandwich formula, that is,

$$\begin{aligned}\text{var}(\hat{\boldsymbol{\mu}}_{BL}) &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \\ &= [(\tau\boldsymbol{\Sigma})^{-1} + \mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{P}]^{-1}\end{aligned}$$

The most important feature of the Black-Litterman model is that it uses the mixed estimation procedure to adjust the entire market equilibrium implied expected return vector with an investor's views. Because security returns are correlated, views on just a few assets will, due to these correlations, imply changes to the expected returns on all assets. Mathematically speaking, this follows from the fact that although the vector \mathbf{q} can have dimension $K \ll N$, $\mathbf{P}'\boldsymbol{\Omega}^{-1}$ is an $N \times K$ matrix that propagates the K views into N components, $\mathbf{P}'\boldsymbol{\Omega}^{-1}\mathbf{q}$. This effect is stronger the more correlated the different securities are. In the absence of this adjustment of the expected return vector, the differences between the equilibrium expected return and an investor's forecasts will be interpreted as an arbitrage opportunity by a mean-variance optimizer and result in portfolios concentrated in just a few assets ("corner solutions"). Intuitively, any estimation errors are spread out over all assets, making the Black-Litterman expected return vector less sensitive to errors in individual views. This effect contributes to the mitigation of estimation risk and error maximization in the optimization process.

Practical Considerations and Extensions

In this subsection we discuss a few practical issues in using the Black-Litterman model. Specifically, we discuss how to incorporate factor models and cross-sectional rankings in this framework. Furthermore, we also provide some ideas on how the confidences in the views can

be estimated in cases where these are not directly available.

It is straightforward to incorporate factor models in the Black-Litterman framework. Let us assume we have a factor representation of the returns of some of the assets, that is

$$R_i = \alpha_i + \mathbf{F}\boldsymbol{\beta}_i + \varepsilon_i, i \in I$$

where $I \subset \{1, 2, \dots, N\}$. Typically, from a factor model it is easy to obtain an estimate of the residual variance, $\text{var}(\varepsilon_i)$. In this case, we set

$$q_i = \begin{cases} \alpha_i + \mathbf{F}\boldsymbol{\beta}_i, & i \in I \\ 0, & \text{otherwise} \end{cases}$$

and the corresponding confidence

$$\omega_{ii}^2 = \begin{cases} \text{var}(\varepsilon_i), & i \in I \\ 0, & \text{otherwise} \end{cases}$$

The \mathbf{P} matrix is defined by

$$\begin{aligned}p_{ii} &= \begin{cases} 1, & i \in I \\ 0, & \text{otherwise} \end{cases} \\ p_{ij} &= 0, i \neq j\end{aligned}$$

Of course in a practical implementation we would omit rows with zeros.

Many quantitative investment strategies do not a priori produce expected returns, but rather just a simple ranking of the securities. Let us consider a ranking of securities from best to worst (from an outperforming to an underperforming perspective, etc.). For example, a value manager might consider ranking securities in terms of increasing book-to-price ratio (B/P), where a low B/P would indicate an undervalued stock (potential to increase in value) and high B/P an overvalued stock (potential to decrease in value). From this ranking we form a long-short portfolio where we purchase the top half of the stocks (the group that is expected to outperform) and we sell short the second half of stocks (the group that is expected to underperform). The view \mathbf{q} in this case becomes a scalar, equal to the expected return on the long-short portfolio. The confidence of the view can be decided from backtests, as we describe next. Further, here the \mathbf{P} matrix is a $1 \times N$ matrix of

ones and minus ones. The corresponding column component is set to one if the security belongs to the outperforming group, or minus one if it belongs to the underperforming group.

In many cases we may not have a direct estimate of the expected return and confidence (variance) of the view. There are several different ways to determine the confidence level.

One of the advantages of a quantitative strategy is that it can be backtested. In the case of the long-short portfolio strategy discussed previously, we could estimate its historical variance through simulation with historical data. Of course, we cannot completely judge the performance of a strategy going forward from our backtests. Nevertheless, the backtest methodology allows us to obtain an estimate of the Black-Litterman view and confidence for a particular view/strategy.

Another approach of deriving estimates of the confidence of the view is through simple statistical assumptions. To illustrate, let us consider the second view in the preceding example: "Stock 3 will outperform Stock 2 by 4%." If we don't know its confidence, we can come up with an estimate for it from the answers to a few simple questions. We start asking ourselves with what certainty we believe the strategy will deliver a return between 3% and 5% ($4\% \pm \alpha$ where α is some constant, in this case $\alpha = 1\%$). Let us say that we believe there is a chance of two out of three that this will happen, $2/3 \approx 67\%$. If we assume normality, we can interpret this as a 67% confidence interval for the future return to be in the interval [3%, 5%]. From this confidence interval we calculate that the implied standard deviation is equal to about 0.66%. Therefore, we would set the Black-Litterman confidence equal to $(0.66\%)^2 = 0.43\%$.

Some extensions to the Black-Litterman model have been derived. For example, Satchel and Scowcroft (2000) propose a model where an investor's view on global volatility is incorporated in the prior views by assuming that τ is unknown and stochastic. Idzorek (2005) introduces a new idea for determining the

confidence level of a view. He proposes that the investor derives his confidence level indirectly by first specifying his confidence in the tilt away from equilibrium (the difference between the market capitalization weights and the weights implied by the view alone). Qian and Gorman (2001) describe a technique based on conditional distribution theory that allows an investor to incorporate his views on any or all variances.

Of course other asset classes beyond equities and bonds can be incorporated into the Black-Litterman framework.¹⁵ Some practical experiences and implementation details have been described by Bevan and Winkelmann (1998) and He and Litterman (1999). A Bayesian approach, with some similarity to the Black-Litterman model, to portfolio selection using higher moments has been proposed by Harvey et al. (2010).

KEY POINTS

- Classical mean-variance optimization is sensitive to estimation error and small changes in the inputs.
- There are four different approaches to make the classical mean-variance framework more robust: (1) improve the accuracy of the inputs; (2) use constraints for the portfolio weights; (3) use portfolio resampling to calculate the portfolio weights; and (4) apply the robust optimization framework to the portfolio allocation process.
- Typically, errors in the expected returns are about 10 times more important than errors in the covariance matrix, and errors in the variances are about twice as important as errors in the covariances.
- Estimates of expected return and covariances can be improved by using shrinkage estimation. Shrinkage is a form of averaging different estimators. The shrinkage estimator typically consists of three components: (1) an estimator with little or no structure; (2) an

estimator with a lot of structure (the shrinkage target); and (3) the shrinkage intensity.

- Jorion's shrinkage estimator for the expected return shrinks toward the return of the global minimum variance portfolio.
- The sample covariance matrix should not be used as an input to the mean-variance problem. By shrinking it toward the covariance matrix with constant correlations, its quality will be improved.
- The Black-Litterman model combines an investor's views with the market equilibrium.
- The Black-Litterman expected return is a confidence weighted linear combination of market equilibrium and the investor's views. The confidence in the views and in market equilibrium determines the relative weighting.
- Factor models as well as simple ranking models can be simultaneously incorporated into the Black-Litterman model.

NOTES

1. See Broadie (1993).
2. We are grateful to Axioma Inc. for providing us with this example. Previously, it has appeared in Ceria and Stubbs (2005).
3. See Best and Grauer (1991, 1992).
4. See Chopra and Ziemba (1993) and Kallberg and Ziemba (1984).
5. See, for example, Frost and Savarino (1988), Chopra (1991), and Grauer and Shen (2000).
6. The relationship to information theory is based upon the premise that the diversification indicators are generalized entropies. See Curado and Tsallis (1991).
7. See, for example, Michaud (1998), Jorion (1992), and Scherer (2002).
8. See Goldfarb and Iyengar (2003).
9. Many similar approaches have been proposed. For example, see Jobson and Korkie (1981) and Frost and Savarino (1986).
10. See, for example, Michaud (1998), Jorion (1986), and Larsen and Resnick (2001).
11. Elton, Gruber, and Urich (1978) proposed the single factor model for purposes of co-

variance estimation. They show that this approach leads to (1) better forecasts of the covariance matrix; (2) more stable portfolio allocations over time; and (3) more diversified portfolios. They also find that the average correlation coefficient is a good forecast of the future correlation matrix.

12. Although straightforward to implement, the optimal shrinkage intensity, w , is a bit tedious to write down mathematically. Let us denote by $r_{i,t}$ the return on security i during period t , $1 \leq i \leq N$, $1 \leq t \leq T$,

$$\bar{r}_i = \frac{1}{T} \sum_{t=1}^T r_{i,t} \quad \text{and}$$

$$\hat{\sigma}_{ij} = \frac{1}{T-1} \sum_{t=1}^T (r_{i,t} - \bar{r}_i)(r_{j,t} - \bar{r}_j)$$

Then the optimal shrinkage intensity is given by the formula

$$w = \max \left\{ 0, \min \left\{ \frac{\hat{\kappa}}{T}, 1 \right\} \right\}$$

where

$$\hat{\kappa} = \frac{\hat{\pi} - \hat{c}}{\hat{\gamma}}$$

and the parameters, $\hat{\pi}$, \hat{c} , $\hat{\gamma}$, are computed as follows. First, $\hat{\pi}$ is given by

$$\hat{\pi} = \sum_{i,j=1}^N \hat{\pi}_{ij}$$

where

$$\hat{\pi}_{ij} = \frac{1}{T} \sum_{t=1}^T ((r_{i,t} - \bar{r}_i)(r_{j,t} - \bar{r}_j) - \hat{\sigma}_{ij})^2$$

Second, \hat{c} is given by

$$\hat{c} = \sum_{i=1}^N \hat{\pi}_{ii} + \sum_{\substack{i=1 \\ i \neq j}}^N \frac{\hat{\rho}}{2} \left(\sqrt{\hat{\rho}_{ii}/\hat{\rho}_{ii}} \hat{\nu}_{i,jj} + \sqrt{\hat{\rho}_{ii}/\hat{\rho}_{ii}} \hat{\nu}_{jj,ii} \right)$$

where

$$\vartheta_{ii,jj} = \frac{1}{T} \sum_{t=1}^T [((r_{i,t} - \bar{r}_i)^2 - \hat{\sigma}_{ii}) \times ((r_{i,t} - \bar{r}_i)(r_{j,t} - \bar{r}_j) - \hat{\sigma}_{ij})]$$

Finally, $\hat{\gamma}$ is given by

$$\hat{\gamma} = \|C - C_{CC}\|_F^2$$

where $\|\cdot\|_F$ denotes the Frobenius norm defined by

$$\|A\|_F = \sqrt{\sum_{i,j=1}^N a_{ij}^2}$$

13. A predecessor to the Black-Litterman model is the so-called Treynor-Black model. In this model, an investor's portfolio is shown to consist of two parts: (1) a passive portfolio/positions held purely for the purpose of mimicking the market portfolio, and (2) an active portfolio/positions based on the investor's return/risk expectations. This somewhat simpler model relies on the assumption that returns of all securities are related only through the variation of the market portfolio (Sharpe's diagonal model). See Treynor and Black (1973).
14. For simplicity, we consider only equity securities. Extending this model to other assets classes such as bonds and currencies is fairly straightforward.

Two comments about the above two relationships are of importance:

1. As it may be difficult to accurately estimate expected returns, practitioners use other techniques. One is that of reverse optimization, also referred to as the technique of implied expected returns. The technique simply uses the expression $\Pi = \delta \Sigma \mathbf{w}$ to calculate the expected return vector given the market price of risk δ , the covariance matrix Σ , and the market capitalization weights \mathbf{w} . The technique was first introduced by Sharpe (1974) and Fisher (1975) and is an impor-

tant component of the Black-Litterman model.

2. We note that $E(R_M) - R_f$ is the market risk premium (or the equity premium) of the universe of assets considered. As pointed out by Herold (2005) and Idzorek (2005), using a market proxy with different risk-return characteristics than the market capitalization weighted portfolio for determining the market risk premium may lead to nonintuitive expected returns. For example, using a market risk premium based on the S&P 500 for calculating the implied equilibrium return vector for the NASDAQ 100 should be avoided.
15. See, for example, Black and Litterman (1992) and Litterman (2003).

REFERENCES

- Ben-Tal, A., and Nemirovski, A. S. (1998). Robust convex optimization. *Mathematics of Operations Research* 23: 769–805.
- Ben-Tal, A., and Nemirovski, A. S. (1999). Robust solutions to uncertain linear programs. *Operations Research Letters* 25: 1–13.
- Best, M. J., and Grauer, R. R. (1991). On the sensitivity of mean-variance-efficient portfolios to changes in assets means: Some analytical and computational results. *Review of Financial Studies* 4: 315–342.
- Best, M. J., and Grauer, R. R. (1992). The analytics of sensitivity analysis for mean-variance portfolio problems. *International Review of Financial Analysis* 1: 17–37.
- Black, F., and Litterman, R. (1990). *Asset Allocation: Combining Investor Views with Market Equilibrium*. *Fixed Income Research*, Goldman Sachs.
- Black, F., and Litterman, R. (1992). Global portfolio optimization. *Financial Analysts Journal* 48: 28–43.
- Bevan, A., and Winkelmann, K. (1998). Using the Black-Litterman global asset allocation model: Three years of practical experience. *Fixed Income Research*, Goldman Sachs.
- Bouchaud, J-P., Potters, M., and Aguilar, J-P. (1997). Missing information and asset allocation. Working Paper, Science & Finance. Capital Fund Management.

- Broadie, M. (1993). Computing efficient frontiers using estimated parameters. *Annals of Operations Research: Special Issue on Financial Engineering* 45: 21–58.
- Ceria, S., and Stubbs, R. A. (2005). Incorporating estimation errors into portfolio selection: Robust portfolio construction. Axioma, Inc.
- Chopra, V. K. (1991). Mean-variance revisited: Near-optimal portfolios and sensitivity to input variations. Russell Research Commentary.
- Chopra, V. K., and Ziemba, W. T. (1993). The effect of errors in means, variances, and covariances on optimal portfolio choice. *Journal of Portfolio Management*, 19: 6–11.
- Curado, E. M. F., and Tsallis, C. (1991). Generalized statistical mechanics: Connection with thermodynamics. *Journal of Physics A: Mathematical and General* 24: L69–L72.
- El Ghaoui, L., and Lebret, H. (1997). Robust solutions to least-squares problems with uncertain data. *SIAM Journal Matrix Analysis with Applications* 18: 1035–1064.
- Elton, E. J., Gruber, M. J., and Urich, T. J. (1978). Are betas best? *Journal of Finance* 33: 1375–1384.
- Fama, E. F., and French, K. R. (1997). Industry costs of equity. *Journal of Financial Economics* 43: 153–193.
- Fisher, L. (1975). Using modern portfolio theory to maintain an efficiently diversified portfolio. *Financial Analysts Journal* 31: 73–85.
- Frost, P. A., and Savarino, J. E. (1988). For better performance: Constrain portfolio weights. *Journal of Portfolio Management* 15: 29–34.
- Frost, P. A., and Savarino, J. E. (1986). An empirical Bayes approach to efficient portfolio selection. *Journal of Financial and Quantitative Analysis* 21: 293–305.
- Goldfarb, D., and Iyengar, G. (2003). Robust portfolio selection problems. *Mathematics of Operations Research* 28: 1–38.
- Grauer, R. R., and Shen, F. C. (2000). Do constraints improve portfolio performance? *Journal of Banking and Finance* 24: 1253–1274.
- Gupta, F., and Eichhorn, D. (1998). Mean-variance optimization for practitioners of asset allocation. Chapter 4 in Frank J. Fabozzi (ed.), *Handbook of Portfolio Management*. Hoboken, NJ: John Wiley & Sons.
- Harvey, C. R., Liechty, J. C., Liechty, M. W., and Mueller, P. (2010). Portfolio selection with higher moments. *Quantitative Finance* 10: 469–485.
- He, G., and Litterman, R. (1999). The intuition behind Black-Litterman model portfolios. *Investment Management Division*, Goldman Sachs.
- Herold, U. (2005). Computing implied returns in a meaningful way. *Journal of Asset Management* 6: 53–64.
- Idzorek, T. M. (2005). A step-by-step guide to the Black-Litterman model: Incorporating user-specified confidence levels. Research Paper, Ibbotson Associates, Chicago.
- Jagannathan, R., and Ma, T. (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Journal of Finance* 58: 1651–1683.
- Jobson, J. D., and Korkie, B. M. (1981). Putting Markowitz theory to work. *Journal of Portfolio Management* 7: 70–74.
- Jorion, P. (1986). Bayes-Stein estimation for portfolio analysis. *Journal of Financial and Quantitative Analysis* 21: 279–292.
- Jorion, P. (1992). Portfolio optimization in practice. *Financial Analysts Journal* 48: 68–74.
- Kallberg, J. G., and Ziemba, W. T. (1984). Misspecification in portfolio selection problems. In G. Bamberg and K. Spremann (eds.), *Risk and Capital: Lecture Notes in Economics and Mathematical Systems*. New York: Springer.
- Larsen, G. Jr., and Resnick, B. (2001). Parameter estimation techniques, optimization frequency, and portfolio return enhancement. *Journal of Portfolio Management* 27: 27–34.
- Ledoit, O., and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* 10: 603–621.
- Ledoit, O., and Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management* 30: 110–119.
- Litterman, R. (2003). *Modern Investment Management: An Equilibrium Approach*. Hoboken, NJ: John Wiley & Sons.
- Markowitz, H. M. (1987). *Mean-Variance Analysis in Portfolio Choice and Capital Markets*. Cambridge, MA: Basil Blackwell.
- Michaud, R. O. (1998). *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation*. Oxford: Oxford University Press.
- Qian, E., and Gorman, S. (2001). Conditional distribution in portfolio theory. *Financial Analysts Journal* 57: 44–51.
- Satchell, S., and Scowcroft, A. (2000). A demystification of the Black-Litterman model: Managing

- quantitative and traditional portfolio construction. *Journal of Asset Management* 1: 138–150.
- Scherer, B. (2002). Portfolio resampling: Review and critique. *Financial Analysts Journal* 58: 98–109.
- Scherer, B. (2007). How different is robust optimization really? *Journal of Asset Management* 7: 374–387.
- Sharpe, W. F. (1963). A simplified model for portfolio analysis. *Management Science* 9: 277–293.
- Sharpe, W. F. (1974). Imputing expected returns from portfolio composition. *Journal of Financial and Quantitative Analysis* 9: 463–472.
- Stein, C. (1956). Inadmissibility of the usual estimator for the mean of multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 1: 197–206.
- Theil, H. (1971). *Principles of Econometrics*. New York: John Wiley & Sons.
- Treynor, J. L., and Black, F. (1973). How to use security analysis to improve portfolio selection. *Journal of Business* 46: 66–86.

Bond Valuation

Basics of Bond Valuation

FRANK J. FABOZZI, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

STEVEN V. MANN, PhD
Professor of Finance, Moore School of Business, University of South Carolina

Abstract: The value of any financial asset is the present value of its expected future cash flows. To value a bond, one must be able estimate the bond's remaining cash flows and identify the appropriate discount rate(s). The traditional approach to bond valuation is to discount every cash flow with the same discount rate. Simply put, the relevant yield curve used in valuation is assumed to be flat. This approach permits opportunities for arbitrage. Alternatively, the arbitrage-free valuation approach starts with the premise that a bond should be viewed as a portfolio or package of zero-coupon bonds. Moreover, each of the bond's cash flows is valued using a unique discount rate that depends on the shape of the yield curve and when the cash flow is delivered in time. The relevant set of discount rates (that is, spot rates) is derived from the Treasury yield curve and when used to value risky bonds augmented with a spread.

Valuation is the process of determining the fair value of a financial asset. In this entry, we will explain the general principles of bond valuation. Our focus will be on how to value *option-free bonds* (that is, bonds that are not callable, puttable, or convertible). A special analytical framework is required to value more complex bond structures such as bonds that are callable or puttable and mortgage-backed and certain asset-backed securities.

GENERAL PRINCIPLES OF BOND VALUATION

The fundamental principle of valuation is that the value of any financial asset is equal to the present value of its expected future cash flows.

This principle holds for any financial asset from zero-coupon bonds to interest rate swaps. Thus, the valuation of a financial asset involves the following three steps:

- Step 1:* Estimate the expected future cash flows.
- Step 2:* Determine the appropriate interest rate or interest rates that should be used to discount the cash flows.
- Step 3:* Calculate the present value of the expected future cash flows found in Step 1 by using the appropriate interest rate or interest rates determined in Step 2.

Estimating Cash Flows

Cash flow is simply the cash that is expected to be received in the future from owning a

financial asset. For a fixed income security, it does not matter whether the cash flow is interest income or repayment of principal. A security's cash flows represent the sum of each period's expected cash flow. Even if we disregard default, the cash flows for only a few fixed income securities are simple to forecast accurately. U.S. Treasury securities possess this feature since they have known cash flows. While the probability of default of the U.S. government is not zero, it is close enough to that threshold to be safely ignored. Besides, if the U.S. government ever does default, we will have other things to worry about than valuing bonds. For Treasury coupon securities, the cash flows consist of the coupon interest payments every six months up to and including the maturity date and the principal repayment at the maturity date.

Many fixed income securities have features that make estimating their cash flows problematic. These features may include one or more of the following:

1. The issuer or the investor has the option to change the contractual due date of the repayment of the principal.
2. The coupon and/or principal payment is reset periodically based on a formula that depends on one or more market variables (e.g., interest rates, inflation rates, exchange rates, etc.).
3. The investor has the choice to convert or exchange the security into common stock or some other financial asset.

Callable bonds, putable bonds, mortgage-backed securities, and asset-backed securities are examples of (1). Floating-rate securities and Treasury Inflation Protected Securities (TIPS) are examples of (2). Convertible bonds and exchangeable bonds are examples of (3).¹

For securities that fall into the first category, a key factor determining whether the owner of the option (either the issuer of the security or the investor) will exercise the option to alter the security's cash flows is the level of interest rates

in the future relative to the security's coupon rate. In order to estimate the cash flows for these types of securities, we must determine how the size and timing of their expected cash flows will change in the future. For example, when estimating the future cash flows of a callable bond, we must account for the fact that when interest rates change, the expected cash flows change. This introduces an additional layer of complexity to the valuation process. For bonds with embedded options, estimating cash flows is accomplished by introducing a parameter that reflects the expected volatility of interest rates.

Determining the Appropriate Interest Rate or Rates

Once we estimate the cash flows for a fixed income security, the next step is to determine the appropriate interest rate for discounting each cash flow. Before proceeding, we pause here to note that we will use the terms "interest rate," "discount rate," and "required yield" interchangeably throughout this entry. The interest rate used to discount a particular security's cash flows will depend on three basic factors: (1) the level of benchmark interest rates (that is, U.S. Treasury rates); (2) the risks that the market perceives the securityholder is exposed to; and (3) the compensation the market expects to receive for these risks.

The minimum interest rate that an investor should require is the yield available in the marketplace on a default-free cash flow. For bonds with dollar-denominated cash flows, yields on U.S. Treasury securities serve as benchmarks for default-free interest rates. For now, we can think of the minimum interest rate that investors require as the yield on a comparable maturity Treasury security.

The additional compensation or spread over the yield on the Treasury issue that investors will require reflects the additional risks the investor faces by acquiring a security that is not issued by the U.S. government. These risks include default risk, liquidity risk, and the risks

associated with any embedded options. These yield spreads will depend not only on the risks an individual issue is exposed to but also on the level of Treasury yields, the market's risk aversion, the business cycle, and so forth.

For each cash flow estimated, the same interest rate can be used to calculate the present value. This is the traditional approach to valuation and it serves as a useful starting point for our discussion. We discuss the traditional approach in the next section and use a single interest rate to determine present values. By doing this, however, we are implicitly assuming that the yield curve is flat. Since the yield curve is almost never flat and a coupon bond can be thought of as a package of zero-coupon bonds, it is more appropriate to value each cash flow using an interest rate specific to that cash flow. After the traditional approach to valuation is discussed, we will explain the proper approach to valuation using multiple interest rates and demonstrate why this must be the case.

Discounting the Expected Cash Flows

Once the expected (estimated) cash flows and the appropriate interest rate or interest rates that should be used to discount the cash flows are determined, the final step in the valuation process is to value the cash flows. The present value of an expected cash flow to be received t years from now using a discount rate i is:

$$\text{Present value}_t = \frac{\text{Expected cash flow in period}^t}{(1 + i)^t}$$

The value of a financial asset is then the sum of the present value of all the expected cash flows. Specifically, assuming that there are N expected cash flows:

$$\text{Value} = \text{Present value}_1 + \text{Present value}_2 + \dots + \text{Present value}_N$$

Determining a Bond's Value

Determining a bond's value involves computing the present value of the expected future cash flows using a discount rate that reflects market interest rates and the bond's risks. A bond's cash flows come in two forms—coupon interest payments and the repayment of principal at maturity. In practice, many bonds deliver semiannual cash flows. Fortunately, this does not introduce any complexities into the calculation. Two simple adjustments are needed. First, we adjust the coupon payments by dividing the annual coupon payment by 2. Second, we adjust the discount rate by dividing the annual discount rate by 2. The time period t in the present value expression is treated in terms of 6-month periods as opposed to years.

To illustrate the process, let's value a 4-year, 6% coupon bond with a maturity value of \$100. The coupon payments are \$3 ($0.06 \times \$100/2$) every six months for the next eight periods. In addition, on the maturity date, the investor receives the repayment of principal (\$100). The value of a nonamortizing bond can be divided in two components: (1) the present value of the coupon payments (that is, an annuity) and (2) the present value of the maturity value (that is, a lump sum). Therefore, when a single discount rate is employed, a bond's value can be thought of as the sum of two present values—an annuity and a lump sum.

The adjustment for the discount rate is easy to accomplish but tricky to interpret. For example, if an annual discount rate of 6% is used, how do we obtain the semiannual discount rate? We will simply use one-half the annual rate, 3.0% ($6\%/2$). How can this be? A 3.0% semiannual rate is not a 6% effective annual rate. As we will see later in this entry, the convention in the bond market is to quote annual interest rates that are just double the semiannual rates. This convention will be explained more fully later when we discuss yield to maturity. For now, accept on faith that one-half the discount rate is used as a semiannual discount rate in the balance of the entry.

We now have everything in place to value a semiannual coupon-paying bond. The present value of an annuity is equal to:

$$\text{Annuity payment} \times \left[\frac{1 - \frac{1}{(1+r)^{\text{no. of years}}}}{r} \right]$$

where r is the *annual* discount rate.

Applying this formula to a semiannual-pay bond, the annuity payment is one half the annual coupon payment and the number of periods is double the number of years to maturity. Accordingly, the present value of the coupon payments can be expressed as:

$$\begin{aligned} &\text{Semiannual coupon payment} \\ &\times \left[\frac{1 - \frac{1}{(1+i)^{\text{no. of years} \times 2}}}{i} \right] \end{aligned}$$

where i is the semiannual discount rate ($r/2$). Notice that in the formula, for the number of periods we use the number of years multiplied by 2 since a period in our illustration is six months.

The present value of the maturity value is just the present value of a lump sum and is equal to:

$$\begin{aligned} &\text{Present value of the maturity value} \\ &= \frac{\$100}{(1+i)^{\text{No. of years} \times 2}} \end{aligned}$$

We will illustrate the calculation by valuing our 4-year, 6% coupon bond assuming that the relevant discount rate is 7%. The data are summarized below:

Semiannual coupon payment = \$3 (per \$100 of par value)

Semiannual discount rate (i) = 3.5% ($7\%/2$)

Number of years to maturity = 4

The present value of the coupon payments is:

$$\$3 \times \left[\frac{1 - \frac{1}{(1.035)^{4 \times 2}}}{0.035} \right] = \$20.6219$$

This number tells us that the coupon payments contribute \$20.6219 to the bond's value.

The present value of the maturity value is:

$$\begin{aligned} \text{Present value of the maturity value} &= \frac{\$100}{(1.035)^{4 \times 2}} \\ &= \$75.9412 \end{aligned}$$

This number (\$75.9412) tells us how much the maturity value contributes to the bond's value. The bond's value is then \$96.5631 (\$20.6219 + \$75.9412). The price is less than par value and the bond is said to be trading at a discount. This will occur when the fixed coupon rate a bond offers (6%) is less than the required yield demanded by the market (the 7% discount rate). A discount bond has an inferior coupon rate relative to new comparable bonds being issued at par so its price must drop so as to offer the required yield of 7%. If the discount bond is held to maturity, the investor will experience a capital gain that just offsets the lower current coupon rate so that it appears equally attractive to new comparable bonds issued at par.

Suppose instead of a 7% discount rate, a 5% discount rate is used. This discount rate is less than the coupon rate on the bond (6%). It can be shown that the present value of the coupon payments is \$21.5104 and the present value of the maturity value is \$82.0747. Thus, the bond's value in this case is \$103.5851. That is, the price is greater than par value and the bond is said to be trading at a premium. This will occur when the fixed coupon rate a bond offers (6%) is greater than the required yield demanded by the market (the 5% discount rate). Accordingly, a premium bond carries a higher coupon rate than new bonds (otherwise the same) being issued today at par so the price will be bid up and the required yield will fall until it equals 5%. If the premium bond is held to maturity, the investor will experience a capital loss that just offsets the benefits of the higher coupon rate so that it will appear equally attractive to new comparable bonds issued at par.

Finally, let's suppose that the discount rate is equal to the coupon rate. That is, suppose that

the discount rate is 6%. It can be shown that the present value of the coupon payments is \$21.0591 and the present value of the maturity value is \$78.9409. Thus, the bond's value in this case is \$100 or par value. Thus, when a bond's coupon rate is equal to the discount rate, the bond will trade at par value. Note that the preceding statement is strictly true only when a bond is valued on its coupon payment dates.

Valuing a Zero-Coupon Bond

For a zero-coupon bond, there is only one cash flow—the repayment of principal at maturity. The value of a zero-coupon bond that matures N years from now is:

$$\frac{\text{Maturity value}}{(1 + i)^{N \times 2}}$$

where i is the semiannual discount rate.

The expression presented above states that the price of a zero-coupon bond is simply the present value of the maturity value. In the present value computation, why is the number of periods used for discounting rather than the number of years to the bond's maturity when there are no semiannual coupon payments? We do this in order to make the valuation of a zero-coupon bond consistent with the valuation of a coupon bond. In other words, both coupon and zero-coupon bonds are valued using semiannual discounting rates.

To illustrate, the value of a 10-year zero-coupon bond with a maturity value of \$100 discounted at a 6.4% interest rate is \$53.2606, as presented below:

$$\begin{aligned} i &= 0.032 = (0.064/2) \\ N &= 10 \\ \frac{\$100}{(1.032)^{10 \times 2}} &= \$53.2606 \end{aligned}$$

Valuing a Bond between Coupon Payments

In our discussion of bond valuation to this point, we have assumed that the bonds are valued on their coupon payment dates (that is, the next coupon payment is one full period away).

For bonds with semiannual coupon payments, this occurs only twice a year. Our task now is to describe how bonds are valued on the other 363 days (or 364 days) of the year.

In order to value a bond with a settlement date between coupon payments, we must answer three questions. First, how many days are there until the next coupon payment date? The answer depends on the day count convention for the bond being valued. Second, how should we compute the present value of the cash flows received over the fractional period? Third, how much must the buyer compensate the seller for the coupon earned over the fractional period? This amount is accrued interest. We will answer these three questions in order to determine the full price and the clean price of a coupon bond. For a more detailed discussion of these issues for not only U.S. bonds but bonds traded in other countries, see Krgin (2002).

Computing the Full Price When valuing a bond purchased with a settlement date between coupon payment dates, the first step is to determine the fractional periods between the settlement date and the next coupon date. Using the appropriate day count convention, this is determined as follows:

$$w \text{ periods} = \frac{\text{Days between settlement date and next coupon payment date}}{\text{Days in the coupon period}}$$

Then the present value of each expected future cash flow to be received t periods from now using a discount rate i assuming the next coupon payment is w periods from now (settlement date) is:

$$\text{Present value}_t = \frac{\text{Expected cash flow}}{(1 + i)^{t-1+w}}$$

Note for the first coupon payment subsequent to the settlement date, $t = 1$ so the exponent is just w . This procedure for calculating the present value when a bond is purchased between coupon payments is called the "Street

method." In the Street method, as can be seen in the previous expression, coupon interest is compounded over the fractional period w .²

To illustrate the calculation, suppose that a U.S. Treasury note maturing on December 31, 2007, was purchased with a settlement date of November 22, 2006. This note's coupon rate was 4.375 and it had coupon payment dates of June 30 and December 31. As a result, the next coupon payment was December 31, 2006, while the previous coupon payment was paid on June 30, 2006. There were three cash flows remaining and they were to be delivered on December 31, 2006, June 30, 2007, and December 31, 2007. The final cash flow represented the last coupon payment and the maturity value of \$100. Also assume the following:

1. Actual/actual day count convention
2. 39 days between the settlement date and the next coupon payment date
3. 184 days in the coupon period

Then w is 0.2120 periods (39/184). The present value of each cash flow assuming that each is discounted at a 4.9% annual discount rate is

$$\begin{aligned} \text{Period 1: Present value}_1 &= \frac{\$2.1875}{(1.0245)^{0.2120}} \\ &= \$2.1761 \end{aligned}$$

$$\begin{aligned} \text{Period 2: Present value}_2 &= \frac{\$2.1875}{(1.0245)^{1.2120}} \\ &= \$2.1243 \end{aligned}$$

$$\begin{aligned} \text{Period 3: Present value}_3 &= \frac{\$102.1875}{(1.0245)^{2.2120}} \\ &= \$96.8498 \end{aligned}$$

The sum of the present values of the cash flows is \$101.1502. This price is referred to as the *full price* (or the *dirty price*).

It is the full price the bond's buyer pays the seller at delivery. However, the very next cash flow received and included in the present value calculation was not earned by the bond's buyer. A portion of the next coupon payment is the accrued interest. Accrued interest is the portion of a bond's next coupon payment that the bond's

seller is entitled to depending on the amount of time the bond was held by the seller. Recall that the buyer recovers the accrued interest when the next coupon payment is delivered.

Computing the Accrued Interest and the Clean Price The last step in this process is to find the bond's value without accrued interest (called the *clean price* or simply price). To do this, the accrued interest must be computed. The first step is to determine the number of days in the accrued interest period (that is, the number of days between the last coupon payment date and the settlement date) using the appropriate day count convention. For ease of exposition, we will assume in the example that follows that the actual/actual calendar is used. We will also assume there are only two bondholders in a given coupon period—the buyer and the seller.

As an illustration, we return to the previous example with the 4.375% coupon Treasury note. Since there were 184 days in the coupon period and 39 days from the settlement date to the next coupon period, there were 145 days (184–39) in the accrued interest period. Therefore, the percentage of the next coupon payment that is accrued interest is:

$$\frac{145}{184} = 0.7880 = 78.80\%$$

Of course, this is the same percentage found by simply subtracting w from 1. In our example, w was 0.2120. Then, $1 - 0.2120 = 0.7880$.

Given the value of w , the amount of accrued interest (AI) is equal to:

$$\text{AI} = \text{Semiannual coupon payment} \times (1 - w)$$

Accordingly, using a 4.375 Treasury note with a settlement date of November 22, 2006, the portion of the next coupon payment that was accrued interest was:

$$\$2.1875 \times (1 - 0.7880) = \$1.7238 \text{ (per \$100 of par value)}$$

Once we know the full price and the accrued interest, we can determine the clean price. The

clean price is the price quoted in the market and represents the bond's value to the new bondholder. The clean price is computed as follows:

$$\text{Clean price} = \text{Full price} - \text{Accrued interest}$$

In our illustration, the clean price is:

$$\$99.43 = \$101.1502 - \$1.7238$$

Note that in computing the full price, the present value of the next coupon payment is computed. However, the buyer pays the seller the accrued interest now despite the fact that it will not be recovered until the next coupon payment date. To make this concrete, suppose one sells a bond such that the settlement date is halfway between the coupon payment dates. In this case $w = 0.50$. Accordingly, the seller will be entitled to one-half of the next coupon payment which would not otherwise be received for another three months. Thus, when calculating the clean price, we subtract "too much" accrued interest—one-half the coupon payment rather than the present value of one-half the coupon payment. Of course, this is the market convention for calculating accrued interest but it does introduce a curious twist in bond valuation.

The Price/Discount Rate Relationship

An important general property of present value is that the higher (lower) the discount rate, the lower (higher) the present value. Since the value of a security is the present value of the expected future cash flows, this property carries over to the value of a security: The higher (lower) the discount rate, the lower (higher) a security's value. We can summarize the relationship between the coupon rate, the required market yield, and the bond's price relative to its par value as follows:

$$\begin{aligned} \text{Coupon rate} &= \text{Yield required by market} \\ &\Rightarrow \text{Price} = \text{Par value} \end{aligned}$$

$$\begin{aligned} \text{Coupon rate} &< \text{Yield required by market} \\ &\Rightarrow \text{Price} < \text{Par value (discount)} \end{aligned}$$

$$\begin{aligned} \text{Coupon rate} &> \text{Yield required by market} \\ &\Rightarrow \text{Price} > \text{Par value (premium)} \end{aligned}$$

This agrees with what we found for the 4-year, 6% coupon bond:

Coupon Rate	Yield Required by Market	Price	Bond Trading at
6%	7%	\$96.5631	Discount
6%	5%	\$103.5851	Premium
6%	6%	\$100.0000	Par

Figure 1 depicts this inverse relationship between an option-free bond's price and its discount rate (that is, required yield). There are two things to infer from the price/discount rate relationship depicted in the figure. First, the relationship is downward sloping. This is simply the inverse relationship between present values and discount rates at work. Second, the relationship is represented as a curve rather than a straight line. In fact, the shape of the curve in Figure 1 is referred to as convex. By convex, it simply means the curve is "bowed in" relative to the origin. This second observation raises two questions about the convex or curved shape of the price/discount rate relationship. First, why is it curved? Second, what is the import of the curvature?

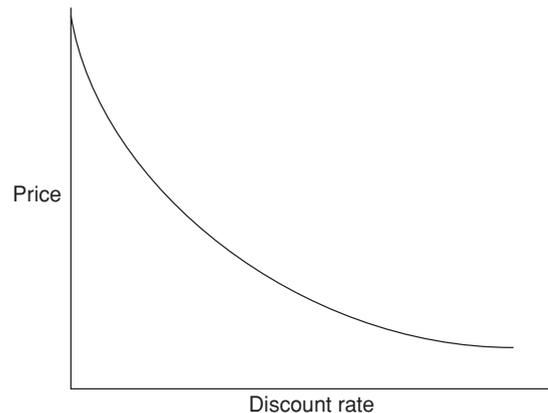


Figure 1 Price/Discount Rate Relationship for an Option-Free Bond

The answer to the first question is mathematical. The answer lies in the denominator of the bond pricing formula. Since we are raising one plus the discount rate to powers greater than one, it should not be surprising that the relationship between the level of the price and the level of the discount rate is not linear.

As for the importance of the curvature to bond investors, let's consider what happens to bond prices in both falling and rising interest rate environments. First, what happens to bond prices as interest rates fall? The answer is obvious—bond prices rise. How about the rate at which they rise? If the price/discount rate relationship was linear, as interest rates fell, bond prices would rise at a constant rate. However, the relationship is not linear, it is curved and curved inward. Accordingly, when interest rates fall, bond prices increase at an increasing rate. Now, let's consider what happens when interest rates rise. Of course, bond prices fall. How about the rate at which bond prices fall? Once again, if the price/discount rate relationship were linear, as interest rates rose, bond prices would fall at a constant rate. Since it curved inward, when interest rates rise, bond prices decrease at a decreasing rate.

Time Path of Bond

As a bond moves towards its maturity date, its value changes. More specifically, assuming that the discount rate does not change, a bond's value:

1. Decreases over time if the bond is selling at a premium.
2. Increases over time if the bond is selling at a discount.
3. Is unchanged if the bond is selling at par value.

With respect to the last property, we are assuming the bond is valued on its coupon anniversary dates.

At the maturity date, the bond's value is equal to its par or maturity value. So, as a bond's

maturity approaches, the price of a discount bond will rise to its par value and a premium bond will fall to its par value—a characteristic sometimes referred to as *pull to par value*.

ARBITRAGE-FREE BOND VALUATION

The traditional approach to valuation is to discount every cash flow of a fixed income security using the same interest or discount rate. The fundamental flaw of this approach is that it views each security as the same package of cash flows. For example, consider a 5-year U.S. Treasury note with a 6% coupon rate. The cash flows per \$100 of par value would be 9 payments of \$3 every six months and \$103 ten 6-month periods from now. The traditional practice would discount every cash flow using the same discount rate regardless of when the cash flows are delivered in time and the shape of the yield curve. Finance theory tells us that any security should be thought of as a package or portfolio of zero-coupon bonds.

The proper way to view the 5-year 6% coupon Treasury note is as a package of zero-coupon instruments whose maturity value is the amount of the cash flow and whose maturity date coincides with the date the cash flow is to be received. Thus, the 5-year 6% coupon Treasury issue should be viewed as a package of 10 zero-coupon instruments that mature every six months for the next five years. This approach to valuation does not allow a market participant to realize an arbitrage profit by breaking apart or "stripping" a bond and selling the individual cash flows (that is, stripped securities) at a higher aggregate value than it would cost to purchase the security in the market. Simply put, arbitrage profits are possible when the sum of the parts is worth more than the whole or vice versa. Because this approach to valuation precludes arbitrage profits, we refer to it as the *arbitrage-free valuation approach*.

By viewing any security as a package of zero-coupon bonds, a consistent valuation

framework can be developed. Viewing a security as a package of zero-coupon bonds means that two bonds with the same maturity and different coupon rates are viewed as different packages of zero-coupon bonds and valued accordingly. Moreover, two cash flows that have identical risk delivered at the same time will be valued using the same discount rate even though they are attached to two different bonds.

To implement the arbitrage-free approach it is necessary to determine the theoretical rate that the U.S. Treasury would have to pay on a zero-coupon Treasury security for each maturity. We say “theoretical” because other than U.S. Treasury bills, the Treasury does not issue zero-coupon bonds. Zero-coupon Treasuries are, however, created by dealer firms. The name given to the zero-coupon Treasury rate is the (Treasury) *spot rate*. Our next task is to explain how the Treasury spot rate can be calculated.

Theoretical Spot Rates

The theoretical spot rates for Treasury securities represent the appropriate set of interest or discount rates that should be used to value default-free cash flows. A default-free theoretical spot rate can be constructed from the observed Treasury yield curve or par curve. We will begin our quest of how to estimate spot rates with the par curve.

Par Rates

The raw material for all yield curve analysis is the set of yields on the most recently issued (that is, on-the-run) Treasury securities. The U.S. Treasury routinely issues 10 securities—the 1-month, 3-month, 6-month, and 1-year bills and the 2-, 3-, 5-, 7-, and 10-year notes, and a 30-year bond. These on-the-run Treasury issues are default risk-free and trade in one of the most liquid and efficient secondary markets in the world. Because of these characteristics, historically Treasury yields serve as a reference benchmark for risk-free rates which are used for

pricing other securities. However, other benchmarks such as the swap curve are now used but the principles of valuation remain unchanged.

In practice, however, the observed yields for the on-the-run Treasury coupon issues are not usually used directly. Instead, the coupon rate is adjusted so that the price of the issue would be the par value. Accordingly, the par yield curve is the adjusted on-the-run Treasury yield curve where coupon issues are at par value and the coupon rate is therefore equal to the yield to maturity. The exception is for the 6-month Treasury bills; the bond-equivalent yield for this issue is already the spot rate.

Deriving a par curve from a set of points starting with the yield on the 6-month bill and ending the yield on the 30-year bond is not a trivial matter. The end result is a curve that tells us “if the Treasury were to issue a security today with a maturity equal to say 12 years, what coupon rate would the security have to pay in order to sell at par?” Some analysts contend that estimating the par curve with only the yields of the on-the-run Treasuries uses too little information that is available from the market. In particular, one must estimate the back-end of the yield curve with only one security, that is, the 30-year bond. Some analysts prefer to use the on-the-run Treasuries and selected off-the-run Treasuries.

In summary, a par rate is the average discount rate of many cash flows (those of a par bond) over many periods. This begs the question, “the average of what?” As we will see, par rates are complicated averages of the implied spot rates. Thus, in order to uncover the spot rates, we must find a method to “break apart” the par rates. There are several approaches that are used in practice.³ The approach that we describe below for creating a theoretical *spot rate curve* is called *bootstrapping*.

Bootstrapping the Spot Rate Curve

Bootstrapping begins with the par curve. To illustrate bootstrapping, we will use the Treasury

Table 1 Hypothetical Treasury Par Yield Curve

Period	Years	Annual Yield to Maturity (BEY) (%)*	Price	Spot Rate (BEY) (%)*
1	0.5	3.00	—	3.0000
2	1.0	3.30	—	3.3000
3	1.5	3.50	100.00	3.5053
4	2.0	3.90	100.00	3.9164
5	2.5	4.40	100.00	4.4376
6	3.0	4.70	100.00	4.7520
7	3.5	4.90	100.00	4.9622
8	4.0	5.00	100.00	5.0650
9	4.5	5.10	100.00	5.1701
10	5.0	5.20	100.00	5.2772
11	5.5	5.30	100.00	5.3864
12	6.0	5.40	100.00	5.4976
13	6.5	5.50	100.00	5.6108
14	7.0	5.55	100.00	5.6643
15	7.5	5.60	100.00	5.7193
16	8.0	5.65	100.00	5.7755
17	8.5	5.70	100.00	5.8331
18	9.0	5.80	100.00	5.9584
19	9.5	5.90	100.00	6.0863
20	10.0	6.00	100.00	6.2169

*The yield to maturity and the spot rate are annual rates. They are reported as bond-equivalent yields. To obtain the semiannual yield or rate, one half the annual yield or annual rate is used.

par curve shown in Table 1. The par yield curve shown extends only out to 10 years. Our objective is to show how the values in the last column of the table (labeled “Spot Rate”) are obtained. Throughout the analysis and illustrations to come, it is important to remember the basic principle is that the value of the Treasury coupon security should be equal to the value of the package of zero-coupon Treasury securities that duplicates the coupon bond’s cash flows.

The key to this process is the existence of the Treasury strips market. A government securities dealer has the ability to take apart the cash flows of a Treasury coupon security (that is, strip the security) and create zero-coupon securities. These zero-coupon securities, which are called Treasury strips, can be sold to investors. At what interest rate or yield can these Treasury strips be sold to investors? The answer is

they can be sold at the Treasury spot rates. If the market price of a Treasury security is less than its value after discounting with spot rates (that is, the sum of the parts is worth more than the whole), then a dealer can buy the Treasury security, strip it, and sell off the Treasury strips so as to generate greater proceeds than the cost of purchasing the Treasury security. The resulting profit is an arbitrage profit.

Before we proceed to our illustration of bootstrapping, a very sensible question must be addressed. Specifically, if Treasury strips are in effect zero-coupon Treasury securities, why not use strip rates (that is, the rates on Treasury strips) as our spot rates? In other words, why must we estimate theoretical spot rates via bootstrapping using yields from Treasury bills, notes, and bonds when we already have strip rates conveniently available? There are three major reasons. First, although Treasury strips are actively traded, they are not as liquid as on-the-run Treasury bills, notes, and bonds. As a result, Treasury strips have some liquidity risk for which investors will demand some compensation in the form of higher yields. Second, the tax treatment of strips is different from that of Treasury coupon securities. Specifically, the accrued interest on strips is taxed even though no cash is received by the investor. Thus, they are negative cash flow securities to taxable entities, and, as a result, their yield reflects this tax disadvantage. Finally, there are maturity sectors where non-U.S. investors find it advantageous to trade off yield for tax advantages associated with a strip. Specifically, certain non-U.S. tax authorities allow their citizens to treat the difference between the maturity value and the purchase price as a capital gain and tax this gain at a favorable tax rate. Some will grant this favorable treatment only when the strip is created from the principal rather than the coupon. For this reason, those who use Treasury strips to represent theoretical spot rates restrict the issues included to coupon strips.

Now let’s see how to generate the spot rates. Consider the 6-month Treasury security in

Table 1. This security is a Treasury bill and is issued as a zero-coupon instrument. Therefore, the annualized bond-equivalent yield (not the bank discount yield) of 3.00% for the 6-month Treasury security is equal to the 6-month spot rate. Using the yield on the 1-year bill, we use 3.3% as the 1-year spot rate. Given these two spot rates, we can compute the spot rate for a theoretical 1.5-year zero-coupon Treasury. The value of a theoretical 1.5-year Treasury should equal the present value of the three cash flows from the 1.5-year coupon Treasury, where the yield used for discounting is the spot rate corresponding to the time of receipt of the cash flow. Since all the coupon bonds are selling at par, as explained in the previous section, the yield to maturity for each bond is the coupon rate. Using \$100 as par, the cash flows for the 1.5-year coupon Treasury are:

0.5 year	$0.035 \times \$100 \times 0.5$	= \$1.75
1.0 year	$0.035 \times \$100 \times 0.5$	= \$1.75
1.5 years	$0.035 \times \$100 \times 0.5 + 100$	= \$101.75

The present value of the cash flows is then:

$$\frac{1.75}{(1+z_1)^1} + \frac{1.75}{(1+z_2)^2} + \frac{101.75}{(1+z_3)^3}$$

where

- z_1 = one-half the annualized 6-month theoretical spot rate
- z_2 = one-half the 1-year theoretical spot rate
- z_3 = one-half the 1.5-year theoretical spot rate

Since the 6-month spot rate is 3% and the 1-year spot rate is 3.30%, we know that:

$$z_1 = 0.0150 \quad \text{and} \quad z_2 = 0.0165$$

We can compute the present value of the 1.5-year coupon Treasury security as:

$$\begin{aligned} \frac{1.75}{(1+z_1)^1} + \frac{1.75}{(1+z_2)^2} + \frac{101.75}{(1+z_3)^3} &= \frac{1.75}{(1.015)^1} \\ &+ \frac{1.75}{(1.0165)^2} + \frac{101.75}{(1+z_3)^3} \end{aligned}$$

Since the price of the 1.5-year coupon Treasury security is equal to its par value (see

Table 1), the following relationship must hold

$$\frac{1.75}{(1.015)^1} + \frac{1.75}{(1.0165)^2} + \frac{101.75}{(1+z_3)^3} = 100$$

If we had not been working with a par yield curve, the equation would have been set to the market price for the 1.5-year issue rather than par value.

Note we are treating the 1.5 year par bond as if it were a portfolio of three zero-coupon bonds. Moreover, each cash flow has its own discount rate that depends on when the cash flow is delivered in the future and the shape of the yield curve. This is in sharp contrast to the traditional valuation approach that forces each cash flow to have the same discount rate.

We can solve for the theoretical 1.5-year spot rate as follows:

$$\begin{aligned} 1.7241 + 1.6936 + \frac{101.75}{(1+z_3)^3} &= 100 \\ \frac{101.75}{(1+z_3)^3} &= 96.5822 \\ (1+z_3)^3 &= \frac{101.75}{96.5822} \\ (1+z_3)^3 &= 1.05351 \\ z_3 &= 0.017527 \\ &= 1.7527\% \end{aligned}$$

Doubling this yield we obtain the bond-equivalent yield of 3.5053%, which is the theoretical 1.5-year spot rate. This is the rate that the market would apply to a 1.5-year zero-coupon Treasury security if, in fact, such a security existed. In other words, all Treasury cash flows to be received 1.5 years from now should be valued (that is, discounted) at 3.5053%.

Given the theoretical 1.5-year spot rate, we can obtain the theoretical 2-year spot rate. The cash flows for the 2-year coupon Treasury in Table 1 are:

0.5 year	$0.039 \times \$100 \times 0.5$	= \$1.95
1.0 year	$0.039 \times \$100 \times 0.5$	= \$1.95
1.5 years	$0.039 \times \$100 \times 0.5$	= \$1.95
2.0 years	$0.039 \times \$100 \times 0.5 + 100$	= \$101.95

The present value of the cash flows is then:

$$\frac{1.95}{(1+z_1)^1} + \frac{1.95}{(1+z_2)^2} + \frac{1.95}{(1+z_3)^3} + \frac{101.95}{(1+z_4)^4}$$

where z_4 = one-half of the 2-year theoretical spot rate.

Since the 6-month spot rate, 1-year spot rate, and 1.5-year spot rate are 3.00%, 3.30%, and 3.5053%, respectively, then:

$$z_1 = 0.0150 \quad z_2 = 0.0165 \quad z_3 = 0.017527$$

Therefore, the present value of the 2-year coupon Treasury security is:

$$\begin{aligned} & \frac{1.95}{(1.0150)^1} + \frac{1.95}{(1.0165)^2} + \frac{1.95}{(1.017527)^3} \\ & + \frac{101.95}{(1+z_4)^4} = 100 \end{aligned}$$

Since the price of the 2-year coupon Treasury security is equal to par, the following relationship must hold:

$$\begin{aligned} & \frac{1.95}{(1.0150)^1} + \frac{1.95}{(1.0165)^2} + \frac{1.95}{(1.017527)^3} \\ & + \frac{101.95}{(1+z_4)^4} = 100 \end{aligned}$$

We can solve for the theoretical 2-year spot rate as follows:

$$\begin{aligned} \frac{101.95}{(1+z_4)^4} &= 94.3407 \\ (1+z_4)^4 &= \frac{101.95}{94.3407} \\ z_4 &= 0.019582 = 1.9582\% \end{aligned}$$

Doubling this yield, we obtain the theoretical 2-year spot rate bond-equivalent yield of 3.9164%.

One can follow this approach sequentially to derive the theoretical 2.5-year spot rate from the calculated values of z_1 , z_2 , z_3 , and z_4 (the 6-month, 1-year, 1.5-year, and 2-year rates), and the price and coupon of the 2.5-year bond in Table 1. Further, one could derive theoretical spot rates for the remaining 15 half-yearly rates. The spot rates thus obtained are shown in the last column of Table 1. They represent the term structure of default-free spot rate for maturities up to 10 years at the particular time to which the bond price quotations refer.

Let us summarize to this point. We started with the par curve which is constructed using the adjusted yields from the on-the-run Treasuries. A par rate is the average discount rate of many cash flows over many periods. Specifically, par rates are complicated averages of spot rates. The spot rates are uncovered from par rates via bootstrapping. A spot rate is the average discount rate of a single cash flow over many periods. It appears that spot rates are also averages. Spot rates are averages of one or more forward rates.

Valuation Using Treasury Spot Rates

To illustrate how Treasury spot rates are used to compute the arbitrage-free value of a Treasury security, we will use the hypothetical Treasury spot rates shown in the fourth column of Table 2 to value an 8%, 10-year Treasury security. The present value of each period's cash flow is shown in the fifth column. The sum of the present values is the arbitrage-free value for the Treasury security. For the 8%, 10-year Treasury it is \$107.0018.

Reason for Using Treasury Spot Rates

Thus far, we have simply asserted that the value of a Treasury security should be based on discounting each cash flow using the corresponding Treasury spot rate. But what if market participants value a security using just the yield for the on-the-run Treasury with a maturity equal to the maturity of the Treasury security being valued? Let's see why the value of a Treasury security should trade close to its arbitrage-free value.

Stripping and Arbitrage-Free Valuation

The key to the arbitrage-free valuation approach is the existence of the Treasury strips market. A dealer has the ability to take apart the cash flows of a Treasury coupon security (that is, strip the security) and create zero-coupon

securities. These zero-coupon securities, called Treasury strips, can be sold to investors. At what interest rate or yield can these Treasury strips be sold to investors? They can be sold at the Treasury spot rates. If the market price of a Treasury security is less than its value using the arbitrage-free valuation approach, then a dealer can buy the Treasury security, strip it, and sell off the individual Treasury strips so as to generate greater proceeds than the cost of purchasing the Treasury security. The resulting profit is an arbitrage profit. Since as we will see, the value determined by using the Treasury spot rates does not allow for the generation of an arbitrage profit, this is referred to as an “arbitrage-free” approach.

To illustrate this, suppose that the yield for the on-the-run 10-year Treasury issue is 7.08%. Suppose that the 8% coupon 10-year Treasury issue is valued using the traditional approach based on 7.08%. The value based on discounting all the cash flows at 7.08% is \$106.5141 as shown in the next-to-the-last column in Table 2.

Consider what would happen if the market priced the security at \$106.5141 and that the spot rates are those shown in the fourth column of Table 2. The value based on the Treasury spot rates is \$107.0018 as shown in the fifth column of Table 2. What can the dealer do? The dealer can buy the 8% 10-year issue for \$106.5141, strip it, and sell the Treasury strips at the spot rates shown in Table 2. By doing so, the proceeds that will be received by the dealer are \$107.0018. This results in an arbitrage profit (ignoring transaction costs) of \$0.4877 ($= \$107.0018 - \106.5141). Dealers recognizing this arbitrage opportunity will bid up the price of the 8% 10-year Treasury issue in order to acquire it and strip it. The arbitrage profit will be eliminated when the security is priced at \$107.0018, the value that we said is the arbitrage-free value.

To understand in more detail where this arbitrage profit is coming from, look at the last three columns in Table 2. The sixth column shows how much each cash flow can be sold for by the dealer if it is stripped. The values in this

Table 2 Determination of the Arbitrage-Free Value of an 8% 10-Year Treasury and Arbitrage Opportunity

Period	Years	Arbitrage-Free Value			Arbitrage Opportunity		
		Cash Flow (\$)	Spot Rate (%)	Present Value (\$)	Sell for	Buy for	Arbitrage Profit
1	0.5	4	6.05	3.8826	3.8826	3.8632	0.0193
2	1.0	4	6.15	3.7649	3.7649	3.7312	0.0337
3	1.5	4	6.21	3.6494	3.6494	3.6036	0.0458
4	2.0	4	6.26	3.5361	3.5361	3.4804	0.0557
5	2.5	4	6.29	3.4263	3.4263	3.3614	0.0648
6	3.0	4	6.37	3.3141	3.3141	3.2465	0.0676
7	3.5	4	6.38	3.2107	3.3107	3.1355	0.0752
8	4.0	4	6.40	3.1090	3.1090	3.0283	0.0807
9	4.5	4	6.41	3.0113	3.0113	2.9247	0.0866
10	5.0	4	6.48	2.9079	2.9079	2.8247	0.0832
11	5.5	4	6.49	2.8151	2.8151	2.7282	0.0867
12	6.0	4	6.53	2.7203	2.7203	2.6349	0.0854
13	6.5	4	6.63	2.6178	2.6178	2.5448	0.0730
14	7.0	4	6.78	2.5082	2.5082	2.4578	0.0504
15	7.5	4	6.79	2.4242	2.4242	2.3738	0.0504
16	8.0	4	6.81	2.3410	2.3410	2.2926	0.0484
17	8.5	4	6.84	2.2583	2.2583	2.2142	0.0441
18	9.0	4	6.93	2.1666	2.1666	2.1385	0.0281
19	9.5	4	7.05	2.0711	2.0711	2.0654	0.0057
20	10.0	104	<u>7.20</u>	<u>51.2670</u>	<u>51.2670</u>	<u>51.8645</u>	<u>-0.5975</u>
			Total	107.0018	107.0018	106.5141	0.4877

column are just those in the fifth column. The next-to-last column shows how much the dealer is effectively purchasing the cash flow for if each cash flow is discounted at 7.08%. The sum of the arbitrage profit from each stripped cash flow is the total arbitrage profit and is contained in the last column.

We have just demonstrated how coupon stripping of a Treasury issue will force the market value to be close to the value as determined by the arbitrage-free valuation approach when the market price is less than the arbitrage-free value (that is, the whole is worth less than the sum of the parts). What happens when a Treasury issue's market price is greater than the arbitrage-free value? Obviously, a dealer will not want to strip the Treasury issue since the proceeds generated from stripping will be less than the cost of purchasing the issue.

When such situations occur, the dealer can purchase a package of Treasury strips so as to create a synthetic Treasury coupon security that is worth more than the same maturity and same coupon Treasury issue. This process is called reconstitution.

The process of stripping and reconstituting ensures that the price of a Treasury issue will not depart materially (depending on transaction costs) from its arbitrage-free value.

Credit Spreads and the Valuation of Non-Treasury Securities

The Treasury spot rates can be used to value any default-free security. For a non-Treasury security, the theoretical value is not as easy to determine. The value of a non-Treasury security is found by discounting the cash flows by the Treasury spot rates plus a yield spread which reflects the additional risks (e.g., default risk, liquidity risks, the risk associated with any embedded options, and so on).

The spot rate used to discount the cash flow of a non-Treasury security can be the Treasury spot rate plus a constant credit spread. For example,

suppose the 6-month Treasury spot rate is 6.05% and the 10-year Treasury spot rate is 7.20%. Also suppose that a suitable credit spread is 100 basis points. Then a 7.05% spot rate is used to discount a 6-month cash flow of a non-Treasury bond and an 8.20% discount rate is used to discount a 10-year cash flow. (Remember that when each semiannual cash flow is discounted, the discount rate used is one-half the spot rate: 3.525% for the 6-month spot rate and 4.10% for the 10-year spot rate.)

The drawback of this approach is that there is no reason to expect the credit spread to be the same regardless of when the cash flow is expected to be received. Consequently, the credit spread may vary with a bond's term to maturity. In other words, there is a term structure of credit spreads. Generally, credit spreads increase with maturity. This is a typical shape for the term structure of credit spreads. Moreover, the shape of the term structure is not the same for all credit ratings. Typically, the lower the credit rating, the steeper the term structure of credit spreads.

Dealer firms typically estimate the term structure of credit spreads for each credit rating and market sector. Typically, the credit spread increases with maturity. In addition, the shape of the term structure is not the same for all credit ratings. Typically, the lower the credit rating, the steeper the term structure of credit spreads.

When the relevant credit spreads for a given credit rating and market sector are added to the Treasury spot rates, the resulting term structure is used to value the bonds of issuers with that credit rating in that market sector. This term structure is referred to as the *benchmark spot rate curve* or *benchmark zero-coupon rate curve*.

For example, Table 3 reproduces the Treasury spot rate curve in Table 2. Also shown is a hypothetical term structure of credit spreads for a non-Treasury security. The resulting benchmark spot rate curve is in the next-to-the-last column. Like before, it is this spot rate curve

Table 3 Calculation of Arbitrage-Free Value of a Hypothetical 8% 10-Year Non-Treasury Security Using Benchmark Spot Rate Curve

Period	Years	Cash Flow (\$)	Treasury Spot Rate (%)	Credit Spread (%)	Benchmark Spot (%)	Present Value (\$)
1	0.5	4	6.05	0.30	6.35	3.8769
2	1.0	4	6.15	0.33	6.48	3.7529
3	1.5	4	6.21	0.34	6.55	3.6314
4	2.0	4	6.26	0.37	6.63	3.5108
5	2.5	4	6.29	0.42	6.71	3.3916
6	3.0	4	6.37	0.43	6.80	3.2729
7	3.5	4	6.38	0.44	6.82	3.1632
8	4.0	4	6.40	0.45	6.85	3.0553
9	4.5	4	6.41	0.46	6.87	2.9516
10	5.0	4	6.48	0.52	7.00	2.8357
11	5.5	4	6.49	0.53	7.02	2.7369
12	6.0	4	6.53	0.55	7.08	2.6349
13	6.5	4	6.63	0.58	7.21	2.5241
14	7.0	4	6.78	0.59	7.37	2.4101
15	7.5	4	6.79	0.63	7.42	2.3161
16	8.0	4	6.81	0.64	7.45	2.2281
17	8.5	4	6.84	0.69	7.53	2.1340
18	9.0	4	6.93	0.73	7.66	2.0335
19	9.5	4	7.05	0.77	7.82	1.9301
20	10.0	104	7.20	0.82	<u>8.02</u>	<u>47.3731</u>
					Total	101.763

that is used to value the securities of issuers that have the same credit rating and are in the same market sector. This is done in Table 3 for a hypothetical 8% 10-year issue. The arbitrage-free value is \$101.763. Notice that the theoretical value is less than that for an otherwise comparable Treasury security. The arbitrage-free value for an 8% 10-year Treasury is \$107.0018 (see Table 3).

KEY POINTS

- A bond can be thought of as a portfolio or package of cash flows. Accordingly, the value of a bond is simply the present value of its remaining expected future cash flows.
- There is an inverse relationship between bond prices/required yields.
- The traditional approach to valuation is to discount each cash flow with the same discount rate. The weakness of the traditional

approach is its reliance on using the same discount rate for all of the bond's cash flows.

- The arbitrage-free approach allows each cash flow to be valued as a zero-coupon bond with a discount rate that depends on the shape of the yield curve and when the cash flow is delivered in time.
- The bootstrapping technique is used to derive the discount rates for discounting a bond's cash flows. These discount rates are called spot rates.
- Default-free bonds should trade at prices close to their arbitrage-free values. The process of stripping and reconstituting of Treasury securities ensures that this will occur.

NOTES

1. For a description of these securities, see Fabozzi (2012).

2. There is another method called the “Treasury method,” which treats the coupon interest over the fractional period as simple interest.
3. There is an extensive literature on estimating spot rates or what is known as term structure modeling. See Fabozzi (2002).

REFERENCES

- Fabozzi, F.J. (2012). *Bond Markets, Analysis, and Strategies: 8ed.* Hoboken, NJ: John Wiley & Sons.
- Fabozzi, F.J. (ed.) (2002). *Interest Rate, Term Structure, and Valuation Modeling.* Hoboken, NJ: John Wiley & Sons.
- Krgin, D. (2002). *Handbook of Global Fixed Income Calculations.* Hoboken, NJ: John Wiley & Sons.

Relative Value Analysis of Fixed-Income Products

STEVEN V. MANN, PhD

Professor of Finance, Moore School of Business, University of South Carolina

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Valuation of fixed-income products employs one of two basic methods—discounted cash flows and relative value. Methods using discounted cash flows require several assumptions to be used as inputs but produce a precise valuation result. The tools of relative value analysis are less ambitious. They help us discern differences in value between two similar bonds on a relative basis. Relative value analysis investors make statements such as “Bond X is cheaper than Bond Y.” Relative value tools range in complexity from yield spreads to asset swap spreads and the credit default swap basis.

There are two basic approaches to the valuation of fixed-income products. The *discounted cash flow method* seeks to value a bond given assumptions about cash flows, reference yield curves, risk premiums, and so on. Given these inputs, the bond’s value is determined. Once computed, this value is compared to the prevailing market price and a rich/cheap determination can be made. The alternative method, *relative valuation*, is less ambitious and not surprisingly more popular.

Tools of relative value analysis, when properly interpreted, give the user some clues about how similar bonds are currently valued in the market on a relative basis. This battery of tools allows us to make conjectures such as “Bond X is cheaper than Bond Y.” Yield measures are basic relative value tools. For example, one method

of measuring a risky bond’s relative value is to compute its yield spread relative to a designated benchmark. Discerning relative value is then a matter of comparing the yield spreads of two or more bonds that are otherwise the same. The bond with the largest yield spread is viewed as the cheapest and is considered the best relative value. In this entry, we will introduce yield spread measures utilizing instruments from both the cash and derivatives markets.¹

One common way fixed-income portfolio managers attempt to outperform benchmarks is through security selection. When pursuing a security selection strategy, managers attempt to overweight cheap issues and underweight rich issues to enhance the total rate of return relative to their benchmark. For this to occur, one or

more of the bond's risks must be mispriced. Active security selection to enhance performance leads to the search for effective relative value tools in bond markets.

YIELD SPREADS OVER SWAP AND TREASURY CURVES

As noted, yield spreads are a frequently used tool of relative value analysis. The computation is a simple one. A yield spread is the difference between a risky bond's yield and a benchmark yield holding maturity constant. It is critical to note the yield spread does not have any predictive power on the bondholder's realized return; the yield spread is merely a convenient way to express the price relative to the benchmark.

There are two commonly used benchmarks: the interest rate swap curve and the U.S. Treasury yield curve. A swap is a contract used to transform cash flows from one form to another. In its most basic form, in an interest rate swap two counterparties agree to exchange cash flows at designated future dates for a specified length of time. The fixed-rate payer makes payments that are determined by a fixed rate called a *swap rate*. Correspondingly, the floating-rate payer makes payments based on a reference rate, usually the London Interbank Offered Rate (LIBOR). LIBOR is the interest rate that prime banks in London are willing to pay other prime banks on certificates of deposit denominated in U.S. dollars.

Market participants quote swap rates for swaps across the maturity spectrum. The relationship between the swap rate and the swap's maturity is called the *swap curve*. Since the reference rate for a swap's floating rate payments is usually LIBOR, the swap curve is also referred to as the *LIBOR curve*.

Over time the *swap curve* has supplanted the Treasury yield curve as the benchmark of choice for computing yield spreads. Indeed, in some countries and currencies, the interest rate swap market is more liquid than the market for sovereign debt. It is important to keep in

mind that the swap curve does not represent a set of default-free interest rates. A swap rate is a rate that embodies two risks: (1) the default risk of the counterparty, and (2) liquidity risk.

As noted, in many countries, the swap curve is the benchmark of choice over a country's government securities yield curve. There are several reasons that augur use of the swap curve. First, in order to construct a government bond yield curve that is reflective of the term structure of interest rates, yields on government securities must be available across the entire maturity spectrum. In most government bond markets, however, a limited number of securities are available. For example, the U.S. Treasury issues only six securities with a maturity of two years or more (two, three, five, seven, 10, and 30 years). Conversely, in the swap market, swap rates are quoted on a wide swath of the maturity spectrum.

Second, technical factors introduce some noise into Treasury yields and preclude them from being clear signals of benchmark risk-free interest rates. Treasury securities differ on dimensions other than level of the coupon and maturity. Yields are affected when a note or bond is *cheapest to deliver* into the Treasury note or bond futures contracts. In addition, yields are also affected when the security is "on special" in the repo market. The tax treatment of bonds, especially those trading at a premium or a discount, can affect yields. Swap rates for the most part do not carry this excess baggage and are therefore more reflective of true, albeit risky, interest rates.

Lastly, because of the differences in sovereign credit risk, comparing government yields across countries is tenuous at best. The swap curve, by contrast, reflects roughly the same level of credit risk across countries. Cross-country comparisons are more meaningful.

A spread over the benchmark swap curve is simply the difference between the yield measure in question and the linearly interpolated swap rate at the same maturity. It should be a suitable yield measure such as yield to maturity,

yield to call, or cash flow yield for structured products. Because the swap rate is interpolated, the spread over the benchmark swap curve is often referred to as the *interpolated spread* or the I-spread. Interpolated spreads circumvent the problem of maturity mismatch that affects the level of the spread. This is especially true if the yield curve is steeply sloped.

To find the I-spread, consider a 5.25% coupon bond issued by General Electric (GE) that matures on December 6, 2017. For a settlement date of January 27, 2009, the I-spread was 261.6 basis points. This spread can be interpreted as the compensation the market demanded for the risk differential between the risky bond and the benchmark swap curve.

The yield spreads can also be computed using active or on-the-run Treasuries. On-the-run Treasuries are the most recently issued Treasury securities of a particular maturity. Since the yield curve is not flat, the yield spreads differ depending on the maturity of the on-the-run Treasury. Thus, even if the yield curve remains fixed, the yield spread will change as the bond rolls down the curve. Using the interpolated 8.9-year Treasury yield, suppose the yield spread for the GE bond on January 27, 2009 was 284 basis points. This yield spread can then be compared to similar bonds at the time in order to determine which bond reflects the best relative value.

ASSET SWAPS

An asset swap is a synthetic structure that transforms the nature of the bond's cash flow from one form into another. The structure is created through the combination of a bond position (fixed-rate or floating-rate) with one or more interest rate swaps. Asset swaps are used extensively by financial institutions for asset-liability management. Namely, asset swaps transform the cash flows of long-term fixed-rate assets to floating-rate cash flows, which are in a form more amenable to financial institutions' funding opportunities.

Asset Swap Mechanics

The mechanics of an asset swap are straightforward. An investor, whom we shall refer to as the asset swap buyer, does the following: (1) takes a long position in a fixed-rate coupon bond with a bullet maturity, and (2) simultaneously enters into an off-market interest rate swap with a tenor equal to the bond's remaining term to maturity. An off-market swap is one whose floating rates are determined with a nonzero spread added to the reference rate. Assume that the bond is trading at par. The asset swap buyer enters into an agreement to pay the semiannual coupon payments as the fixed-rate leg in exchange for floating-rate payments at LIBOR plus (or minus) a spread (called the asset swap spread). For simplicity, assume the frequency of the fixed-rate and floating-rate payments are the same. The spread over LIBOR that makes the net present value of the coupon payments (i.e., the fixed-rate leg) and the projected floating-rate payments equal to zero is the *asset swap spread*.² This asset swap spread is used as a measure of relative value regardless of whether the cash flows are actually swapped.

Determining the Asset Swap Spread for a Par Bond

To better understand how all the pieces fit together, let's illustrate how an asset swap spread is calculated. Consider a corporate bond issued by General Electric that matures on December 6, 2017, and pays coupon interest semiannually at an annual rate of 5.25%. Assume a position with a par value of \$1 million. Further assume that this bond sold for par for settlement on December 6, 2008. For ease of exposition, we will evaluate the asset swap on a coupon payment date to abstract some of the details of swaps.

The asset swap spread is determined using the following procedure. First, assume that a \$1 million par value position of the General Electric coupon bond was valued at a price of \$100 for settlement on December 6, 2008. (It actually traded at a large premium at the time.) The

price paid for the bond at settlement is the flat price of \$1,000,000 plus zero accrued interest such that the full price is \$1,000,000 since it is a coupon payment date. Second, assume that a long position in an interest rate swap is established with a notional principal of \$1,000,000. Third, determine the net cash difference at settlement. This amount is simply the difference between the bond's full price and the swap's principal amount plus accrued interest. By construction, this difference is zero in our illustration. Fourth, determine the spread over the reference rate (i.e., LIBOR) required to equate the present value of the swap's floating-rate payments and the present value of the fixed-rate payments (i.e., the bond's cash flows). In our illustration, a swap spread of 221.1 basis points satisfied this condition.

Our illustration is a special case for a bond selling at par, and the accrued interest on both the bond and the swap are equal to zero. The asset swap spread makes the present value of a par swap's floating payments equal the bond's payments to maturity. This is true because the net cash at settlement is equal to zero.

Par versus Market Structures

Market participants use two types of fixed-floating asset swap structures—par and market. The par structure is the most prevalent. When utilizing a *par structure*, the notional amount of the interest rate swap is equal to the bond's maturity value. The price of the bond acquired by the asset swap buyer is par regardless of its market price.³ If the bond is trading at a discount, the asset swap seller receives more for the bond than it is worth and garners an upfront "profit." Alternatively, if the bond is trading at a premium, the asset swap seller receives less for the bond than it is worth and suffers an upfront "loss." At the initiation of the asset swap, the present value of the net cash flows of both parties is zero, so any upfront profit or loss is illusory because the spread adjusts. The asset swap seller "gives up" the premium over par

at inception and in return pays a lower spread on the floating-rate cash flows. For bonds trading at a discount, the asset swap seller pays a higher spread on the floating-rate cash flows as recompense for capturing the discount at settlement.

An asset swap with a par structure is two separate transactions: (1) The asset swap buyer pays par to the asset swap seller for a bond and (2) an off-market swap. Accordingly, after the asset swap's cash flows are established, the bond's credit performance has no impact on the interest rate swap. If the bond were to default, the asset swap buyer no longer receives coupon payments or the maturity payment. The asset swap buyer's obligations imposed by the swap continue on as before until it matures or can be closed out at market value.

An alternative structure for an asset swap is called a market structure. This method differs from a par structure in four respects. First, the bond is purchased at its prevailing market price rather than at par. Second, the notional principal of the off-market swap floating-rate payments is scaled by the bond's full price. Third, at the end of the transaction's life, the asset swap buyer pays par to the asset swap seller and receives the original full price of the bond. Lastly, note also that the counterparty risk exposure is allocated differently in the two asset swap structures. If the bond in question trades at a premium, the asset swap seller bears more of the counterparty risk. Conversely, in a market structure for the same premium bond, the counterparty risk is tilted toward the asset swap buyer due to the net payment of the bond's premium at the end of the transaction. Correspondingly, if the bond in question trades at a discount, the tilt of the counterparty risk exposure is reversed for both structures.

Determining the Asset Swap Spread in the General Case

Let's introduce some real-world complications. First, we consider an asset swap with a

settlement date that falls between two coupon payment dates. Once this circumstance is considered, both the coupon-paying bond and swap will have nonzero accrued interest. Suppose an asset swap with a par structure has a settlement day that falls between the two semiannual coupon payment dates. By market convention, the asset swap buyer pays par for the bond and does not directly pay accrued interest. The asset swap buyer receives the full coupon payment at the next payment date and pays the full coupon payment as required on the fixed-rate side of the swap. The floating-rate swap payment from the asset swap seller is treated somewhat differently. Floating-rate payments are usually more frequent than fixed-rate payments (quarterly versus semiannually) and almost always use a different day count convention. The floating-rate payment is adjusted accordingly.

As an illustration, consider a 4.125% coupon bond issued by Wal-Mart that matured on February 15, 2011. This bond delivered coupon payments semiannually. Suppose an asset swap buyer took a long position in this bond that was trading at a flat price of 103.764. We will sketch the procedure for calculating the asset swap spread if it had a trade settlement date of June 23, 2008. The notional principal is set to the default of \$1 million. The asset swap spread that equates the present value of the cash flows was 75.7 basis points. As a result, the floating-rate swap payments would have been calculated with a rate of 3-month LIBOR plus 75.7 basis points. The asset swap buyer's swap payments would have been simply the five semiannual coupon payments of \$20,625 and \$1,000,000 on the maturity date of February 15, 2011. The asset swap seller's floating-rate swap payments would have depended on the value of 3-month LIBOR on each payment date. As noted, the first floating-rate payment of \$2,835.04 reflects the accrual from the settlement date on January 28, 2009, to the first payment date of February 15, 2009, using an actual/360 day count convention.

Uses of Asset Swaps

The primary reason for using an asset swap is to acquire some exposure to risks of a fixed rate while neutralizing the interest rate risk. For example, financial institutions typically fund on a floating-rate basis and unless they have a view on interest rates, management wants to invest in floating-rate assets. Financial institutions are active participants in the asset swap market by buying fixed-rate bonds and transforming the cash flow from those bonds into floating payments, which provide a better match against their liability structure. An active asset swap market tends to eliminate pricing discrepancies between fixed-rate and floating-rate products.

Asset swap spreads are often used as an indicator of relative value. If a fixed-income investor is considering five fixed-rate bonds of similar maturity and risk for inclusion in a portfolio and wants to assess their relative value, the investor would simply find the highest asset swap spreads, which represent the best relative value.

In practice, however, asset swaps are typically employed as a relative value detector in the following manner. After choosing portfolio duration (and perhaps key rate durations to control shaping risk) and after choosing a credit mix (or perhaps an average credit rating), find the constrained portfolio that produces the highest asset swap spread. This portfolio presumably represents the best relative value for a given duration target and credit target—with or without distributional constraints on durations and credit ratings.

A Miscellany of Asset Swaps

There are a handful of variations on the standard asset swap structure discussed to this point. A forward start asset swap involves taking a long position in a risky bond on a forward settlement date in combination with an interest rate swap whose asset swap spread is established today. This transaction allows an investor to gain an exposure to a risky product in the future at a known price today. Investors bear

no exposure to credit risk until the forward settlement date because the asset swap terminates if the bond defaults prior to this date.

A cross-currency asset swap is a combination of a long position in a risky bond whose cash flows are denominated in a different currency and an off-market interest rate swap. The swap transforms the fixed-rate coupon payments into floating-rate cash flows in the investor's home currency. An exchange of principal occurs at the end of the swap's life as is common with currency swaps. Moreover, the swap's cash flows are converted using a predetermined exchange rate. This asset swap variation would allow, say, a U.S. investor to take an exposure to a yen-denominated corporate bond while simultaneously mitigating the interest rate and currency risks.

Investors often use asset swaps in convertible bond arbitrage. Convertibles are ideal securities for "arbitrage" because the convertible itself, namely the underlying stock and the embedded derivatives, are traded along predictable ratios and any discrepancy or mispricing would give rise to arbitrage opportunities for hedge fund managers to exploit. The valuation of convertible bonds is driven by four primary factors: (1) interest rates, (2) credit spreads, (3) stock prices, and (4) volatility of stock prices. Convertible bond arbitrage involves taking a leveraged position (usually long) in the convertible bond to gain exposure to a mispriced factor while simultaneously hedging interest rates and small changes in stock prices.

Callable asset swaps are used to strip out equity and credit components with a structure that allows the investor to cancel the off-market swap on any call date. This ability to terminate the swap is accomplished through the purchase of Bermudan receiver swaptions.

CREDIT DEFAULT SWAPS

Credit default swaps (CDS) are contracts that enable the transfer of credit risk between the two

counterparties to the trade. CDS resemble insurance policies.⁴ Taking long/short CDS positions is referred to as buying/selling "protection." The protection buyer pays the protection seller a periodic payment (premium) for protection against a credit event experienced by a reference asset or entity. Simply put, sellers of protection are taking on credit risk for a fee while protection buyers are paying to reduce their credit risk exposure. A reference asset could refer to a single asset, and this is termed a single-name credit default swap. Alternatively, if the reference asset is a group of assets, it is referred to as a basket credit default swap. A reference entity could be a corporation or government entity (sovereign or municipal).

The payout of credit default swaps is contingent on the occurrence of a *credit event*. Definitions of credit events are published by the ISDA, the so-called "1999 Definitions." The 1999 Definitions list eight different credit events, which include: (1) bankruptcy, (2) credit event upon merger, (3) cross acceleration, (4) cross default, (5) downgrade, (6) failure to pay, (7) repudiation/moratorium, and (8) restructuring. The most controversial credit event is a restructuring. A restructuring refers to an alteration of the debt obligation's original terms in an effort to make the obligation less onerous to the borrower. Among the terms that may be offered: (1) reduction in the stated rate of interest, (2) principal reduction, (3) principal payment rescheduling or interest payment postponement, or (4) a change in the seniority level of the obligation. The inclusion of restructuring as a trigger for a credit event is desired by protection buyers because they insist it is part of their essential credit protection. Protection sellers counter that the restructuring provision is triggered by routine modifications to the debt. In April 2001, the ISDA issued the so-called "Supplement Definition" that indicates the conditions needed to qualify for a restructuring: (1) The reference obligation must have at least four bondholders, and (2) at least two-thirds of the bondholders must consent to the restructuring.

The market for single-name credit default swaps is an over-the-counter interdealer market. For credit default swaps on corporate or sovereign debt, the contract specifications are largely standardized. For example, the tenor is usually five years. Certain dealers are also willing to create customized contracts better suited to the counterparty’s risk exposure. A protection buyer makes payments (typically quarterly) that are fixed by contract until a credit event is triggered or maturity, whichever is earlier. The formula for calculating the protection buyer’s quarterly is given by the expression

$$\begin{aligned} \text{quarterly payment} &= \text{CDS spread} \\ &\times \text{notional principal} \\ &\times (\text{days in period})/360 \end{aligned}$$

Figure 1 presents these payments. If a credit event does not occur during the tenor of the CDS, the protection buyer’s fixed payments are the only payments. At inception, there is no exchange of principal between the buyer and the seller. If a credit event is triggered, there is an exchange between the protection buyer and protection seller. The protection buyer makes accrued payments up until the credit event date and then stops making quarterly payments.

What both parties must do when there is a credit event depends on the settlement terms of the CDS. The settlement terms can specify either physical settlement or cash settlement. If the CDS specifies physical delivery, the protection buyer delivers the reference obligation to the protection seller in return for the cash payment. Figure 2 illustrates this scenario. If the credit event is triggered, the seller’s pay-

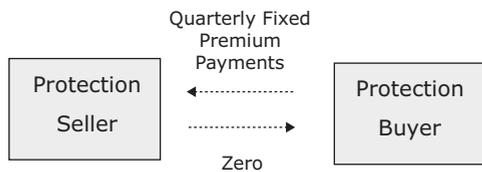


Figure 1 Premium Payments for a CDS Assuming no Credit Event

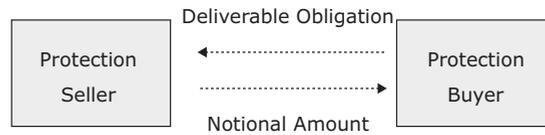


Figure 2 The Exchange if a Credit Event Occurs

ment may be a prespecified amount or it may reflect the reference obligation’s value decline. When the payment is fixed, it is based on a notional principal amount. Conversely, when the payment is based on the reference obligation’s value decline, it is usually computed using pricing information obtained by polling several CDS dealers.

Usually there is more than one obligation of the reference entity from which the protection buyer can choose. The set of all obligations that are permitted for physical delivery is called the deliverable obligations. Any obligation meeting the stated criteria (coupon, maturity, etc.) is part of this basket. Naturally, the protection buyer will choose among the deliverable obligations the one that is cheapest to deliver.

CDS are structured to replicate the experience of a default in the cash market. If a credit event occurs, the deliverable obligation should trade at a deep discount to par.

The seller’s net loss will be the difference between par and the deliverable obligation’s recovery value. Note that the CDS is a pure play in the deliverable obligation’s credit risk. A long position in the reference instrument exposes the investor to other risks.

As an illustration, consider a CDS with a reference asset being a Citigroup 6.5% coupon bond that matured on January 18, 2011. The notional principal for this contract was \$10 million. Suppose the following information was available:

Reference Entity/Asset	Citigroup 6.5% 1/18/2011
Tenor	5 years
Effective date	7/3/08
Maturity date	9/20/13
Payment frequency	Quarterly

The first coupon payment date was September 22, 2008. Suppose the deal spread was 143.5 basis points. Presented in the following table are the first four quarterly payments that the protection buyer made to the seller.

Date	Cash flow (\$)
9/22/08	\$32,287.50
12/22/08	\$36,273.61
3/20/09	\$35,077.78
6/22/09	\$37,469.44

There were 81 actual days of accrual from the effective date of 7/3/08 to the first coupon date of 9/22/08, so inserting this number along with the notional principal of \$10 million and a spread of 143.5 basis points (in decimal) gives the first quarterly payment of

$$\$32,287.50 = 0.01435 \times \$10 \text{ million} \times (81/360)$$

The remainder of the quarterly payments are computed in the same fashion. Note that while the CDS spread remains fixed, the payments will vary somewhat due to the varying number of days between coupon payment dates.

Credit Default Swap Basis

A CDS is, under certain simplifying assumptions, equivalent to a long position in an asset-swapped fixed-rate bond financed with a repurchase (repo) agreement. Accordingly, it is critical to address the linkage between asset swap spreads, CDS spreads, and credit spreads.

Practitioners access relative value by comparing CDS spreads and asset-swap spread levels. In fact, the difference between the CDS premium and the asset swap spread is referred to as the *credit default swap basis* (CDS basis).⁵ Practitioners also look at differences between CDS spreads and either the I-spread or the zero-volatility spread (Z-spread). A nonzero basis signals opportunities for investors. If the basis is negative (i.e., the CDS spread is less than the asset swap spread), this suggests that the investor buy the bond in the cash market and buy protection via a CDS. Conversely, if the basis

is positive (i.e., the CDS spread is greater than the asset swap spread), this suggests that the investor sell the bond in the cash market and sell protection via a CDS.

KEY POINTS

- There are two approaches to the valuation of fixed-income products: discounted cash flow and relative value.
- The relative value method can provide information about how similar bonds are priced on a relative basis.
- A yield spread is the difference between a risky bond's yield and a benchmark yield holding maturity constant.
- Two commonly used benchmark yield curves are the swap curve and the U.S. Treasury curve.
- An asset swap is a synthetic structure that transforms the nature of cash flows from one form into another.
- An asset swap spread is used as an indicator of relative value and is the spread over the reference rate that equates the value of the floating rate cash flows and the bond's cash flows.
- The credit default swap (CDS) basis is the difference between the CDS premium and the asset swap spread.
- A nonzero CDS basis signals opportunities for investors.

NOTES

1. For a further discussion of relative value tools, see Fabozzi and Mann (2010) and Grieves and Mann (2010).
2. For simplicity, we are ignoring any nonzero net payments at the beginning and end of the swap's life. These elements will be introduced shortly.
3. When nonpar bonds are purchased as part of an asset swap structure, tax and accounting rules create incentives to buy and sell premium/discount bonds at par through an asset swap structure.

4. More on the mechanics of CSD can be found in Anson, Fabozzi, Choudhry, and Chen (2004).
5. For a further discussion of the CDS spread, see Choudhry (2006).

REFERENCES

- Anson, M., Fabozzi, F. J., Choudhry, M., and Chen, R-R. (2004). *Credit Derivatives: Instruments, Pricing, and Applications*. Hoboken, NJ: John Wiley & Sons.
- Choudhry, M. (2006). *The Credit Default Swap Basis*. Princeton, NJ: Bloomberg Press.
- Fabozzi, F. J., and Mann, S. V. (2010). *Introduction to Fixed Income Analytics: Relative Value Analysis, Risk Measures, and Valuation*, 2nd ed. Hoboken, NJ: John Wiley & Sons.
- Grieves, R., and Mann, S. V. (2010). The search for the relative value in bonds. *Financial Markets and Portfolio Management* 25, 1: 95–106.

Yield Curves and Valuation Lattices

FRANK J. FABOZZI, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

ANDREW KALOTAY, PhD
President, Andrew Kalotay Associates

MICHAEL DORIGAN, PhD
Senior Quantitative Analyst, PNC Capital Advisors

Abstract: The complication in valuing bonds with embedded options and option-type derivatives is that cash flows depend on interest rates in the future. Academicians and practitioners have attempted to capture this interest rate uncertainty through various models, often designed as one- or two-factor processes. These models attempt to capture the stochastic behavior of rates. In practice, these elegant mathematical models must be implemented numerically in order to be useful. One such model is a single factor model that assumes a stationary variance, or volatility.

An often-used framework for the valuation of interest rate instruments with embedded options and interest rate option-type derivatives is the lattice framework. Effectively, the lattice specifies the distribution of short-term interest rates over time. The lattice holds all the information required to perform the valuation of certain option-like interest rate products. First, the lattice is used to generate the cash flows across the life of the security. Next, the interest rates on the lattice are used to compute the present value of those cash flows.

There are several interest rate models that have been used in practice to construct an *interest rate lattice*. In each case, interest rates can realize one of several possible levels when we move from one period to the next. A lattice model that

allows only two rates in the next period is called a binomial model. A lattice model that allows three possible rates in the next period is called a trinomial model. There are even more complex models that allow more than three possible rates in the next period.

Regardless of the underlying assumptions, each model shares a common restriction. In order to be “arbitrage-free,” the interest rate tree generated must produce a value for an on-the-run optionless bond that is consistent with the current par *yield curve*. In effect, the value generated by the model must be equal to the observed market price for the optionless instrument. Under these conditions the model is said to be “arbitrage free.” A lattice that produces an arbitrage-free valuation is said to be “fair.” The

lattice is used for valuation only when it has been calibrated to be fair. More on calibration below.

In this entry we will demonstrate how a lattice is used to value an option-free bond. The model is also used to value bonds with embedded options, floating-rate securities with option-type derivatives, bond options, and swaptions.¹

THE INTEREST RATE LATTICE

In our illustration, we represent the lattice as a binomial tree, the simplest lattice form. Figure 1 provides an example of a *binomial interest rate tree*, which consists of a number of “nodes” and “legs.” Each leg represents a one-year interval over time. A simplifying assumption of one-year intervals is made to illustrate the key principles. The methodology is the same for smaller time periods. In fact, in practice the selection of the length of the time period is critical, but we need not be concerned with this nuance here.

The distribution of future interest rates is represented on the tree by the nodes at each point in time. Each node is labeled as “N” and has

a subscript, a combination of L’s and H’s. The subscript indicates whether the node is lower or higher on the tree, respectively, relative to the other nodes. Thus, node N_{HH} is reached when the 1-year rate realized in the first year is the higher of the two rates for that period, then the highest of the rates in the second year.

The root of the tree is N , the only point in time at which we know the interest rate with certainty. The 1-year rate today (that is, at N) is the current 1-year spot rate, which we denote by r_0 .

We must make an assumption concerning the probability of reaching one rate at a point in time. For ease of illustration, we have assumed that rates at any point in time have the same probability of occurring. In other words, the probability is 50% on each leg.

The interest rate model we will use to construct the binomial tree assumes that the 1-year rate evolves over time based on a lognormal random walk with a known (stationary) volatility. Technically, the tree represents a *one-factor model*. Under the distributional assumption, the relationship between any two adjacent rates at a point in time is calculated via the following equation:

$$r_{1,H} = r_{1,L} e^{2\sigma\sqrt{t}}$$

where σ is the assumed volatility of the 1-year rate, t is the length of the time period in years, and e is the base of the natural logarithm. Since we assume a 1-year interval, that is, $t = 1$, we can disregard the calculation of the square root of t in the exponent.

For example, suppose that $r_{1,L}$ is 4.4448% and σ is 10% per year, then:

$$\begin{aligned} r_{1,H} &= 4.4448\%(e^{2 \times 0.10}) = 4.4448\%(1.2214) \\ &= 5.4289\% \end{aligned}$$

In the second year, there are three possible values for the 1-year rate. The relationship between $r_{2,LL}$ and the other two 1-year rates is as follows:

$$r_{2,HH} = r_{2,LL}(e^{4\sigma}) \quad \text{and} \quad r_{2,HL} = r_{2,LL}(e^{2\sigma})$$

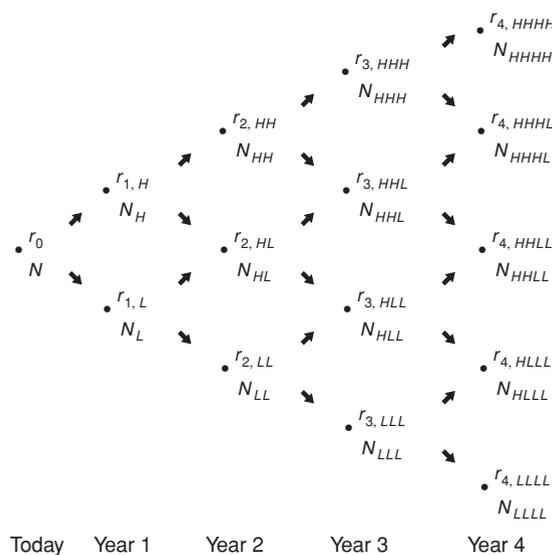


Figure 1 Four-Year Binomial Interest Rate Tree

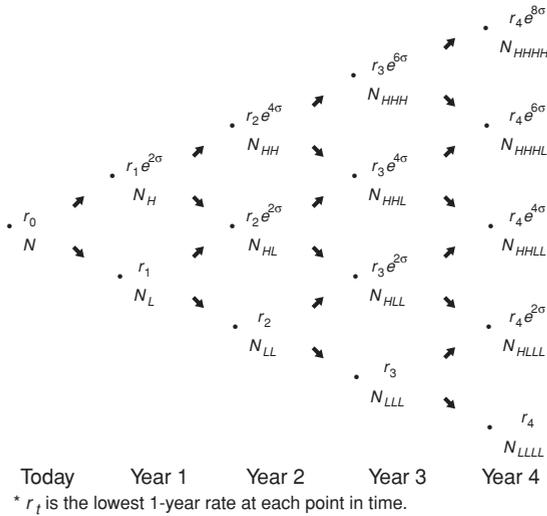


Figure 2 Four-Year Binomial Interest Rate Tree with 1-Year Rates*

So, for example, if $r_{2,LL}$ is 4.6958%, and assuming once again that σ is 10%, then

$$r_{2,HH} = 4.6958\%(e^{4 \times 0.10}) = 7.0053\%$$

and

$$r_{2,HL} = 4.6958\%(e^{2 \times 0.10}) = 5.7354\%$$

This relationship between rates holds for each point in time. Figure 2 shows the interest rate tree using this notation.

Determining the Value at a Node

In general, to get a security's value at a node we follow the fundamental rule for valuation: The value is the present value of the expected cash flows. The appropriate discount rate to use for cash flows one year forward is the 1-year rate at the node where we are computing the value. Now there are two present values in this case: the present value of the cash flows in the state where the 1-year rate is the higher rate, and one where it is the lower rate state. We have assumed that the probability of both outcomes is equal. Figure 3 provides an illustration for a node assuming that the 1-year rate is r^* at the node where the valuation is sought and letting:

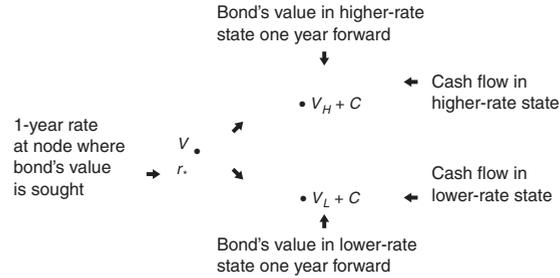


Figure 3 Calculating a Value at a Node

V_H = the bond's value for the higher 1-year rate state

V_L = the bond's value for the lower 1-year rate state

C = coupon payment

From where do the future values come? Effectively, the value at any node depends on the future cash flows. The future cash flows include (1) the coupon payment one year from now and (2) the bond's value one year from now, both of which may be uncertain. Starting the process from the last year in the tree and working backwards to get the final valuation resolves the uncertainty. At maturity, the instrument's value is known with certainty—par. The final coupon payment can be determined from the coupon rate, or from prevailing rates to which it is indexed. Working back through the tree, we realize that the value at each node is quickly calculated. This process of working backward is often referred to as *recursive valuation*.

Using our notation, the cash flow at a node is either:

- $V_H + C$ for the higher 1-year rate
- $V_L + C$ for the lower 1-year rate

The present value of these two cash flows using the 1-year rate at the node, r^* , is:

$$\frac{V_H + C}{(1 + r^*)} = \text{present value for higher 1-year rate}$$

$$\frac{V_L + C}{(1 + r^*)} = \text{present value for lower 1-year rate}$$

Then, the value of the bond at the node is found as follows:

$$\text{Value at a node} = \frac{1}{2} \left[\frac{V_H + C}{(1 + r^*)} + \frac{V_L + C}{(1 + r^*)} \right]$$

CALIBRATING THE LATTICE

We noted above the importance of the no-arbitrage condition that governs the construction of the lattice. To assure this condition holds, the lattice must be calibrated to the current par yield curve, a process we demonstrate here. Ultimately, the lattice must price optionless par bonds at par.

Assume the on-the-run par yield curve for a hypothetical issuer as it appears in Table 1. The current 1-year rate is known, 3.50%. Hence, the next step is to find the appropriate 1-year rates one year forward. As before, we assume that volatility, σ , is 10% and construct a 2-year tree using the 2-year bond with a coupon rate of 4.2%, the par rate for a 2-year security.

Figure 4 shows a more detailed binomial tree with the cash flow shown at each node. The root rate for the tree, r_0 , is simply the current 1-year rate, 3.5%. At the beginning of Year 2 there are two possible 1-year rates, the higher rate and the lower rate. We already know the relationship between the two. A rate of 4.75% at N_L has been arbitrarily chosen as a starting point. An iterative process determines the proper rate (that is, trial and error). The steps are described and illustrated below. Again, the goal is a rate that, when applied in the tree, provides a value of par for the 2-year, 4.2% bond.

Step 1: Select a value for r_1 . Recall that r_1 is the lower 1-year rate. In this first trial, we arbitrarily selected a value of 4.75%.

Table 1 Issuer Par Yield Curve

Maturity	Par Rate	Market Price
1 year	3.50%	100
2 years	4.20%	100
3 years	4.70%	100
4 years	5.20%	100

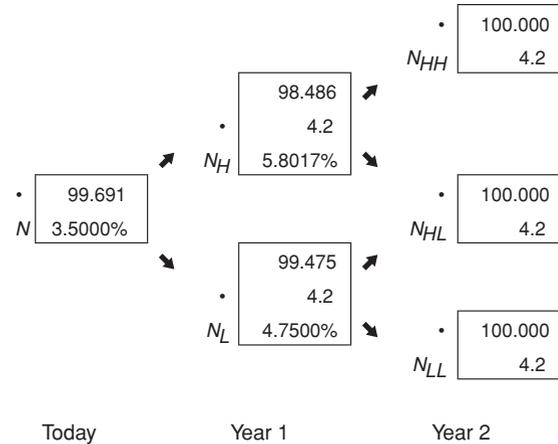


Figure 4 The 1-Year Rates for Year 1 Using the 2-Year 4.2% On-the-Run Issue: First Trial

Step 2: Determine the corresponding value for the higher 1-year rate. As explained earlier, this rate is related to the lower 1-year rate as follows: $r_1 e^{2\sigma}$. Since r_1 is 4.75%, the higher 1-year rate is 5.8017% ($= 4.75\% e^{2 \times 0.10}$). This value is reported in Figure 4 at node N_H .

Step 3: Compute the bond's value one year from now as follows:

- Determine the bond's value two years from now. In our example, this is simple. Since we are using a 2-year bond, the bond's value is its maturity value (\$100) plus its final coupon payment (\$4.2). Thus, it is \$104.2.
- Calculate V_H . Cash flows are known. The appropriate discount rate is the higher 1-year rate, 5.8017% in our example. The present value is \$98.486 ($= \$104.2 / 1.058017$).
- Calculate V_L . Again, cash flows are known—the same as those in Step 3b. The discount rate assumed for the lower 1-year rate is 4.75%. The present value is \$99.475 ($= \$104.2 / 1.0475$).

Step 4: Calculate V .

- Add the coupon to both V_H and V_L to obtain the values at N_H and N_L , respectively. In our example we have \$102.686 for the higher rate and \$103.675 for the lower rate.

b. Calculate V . The 1-year rate is 3.50%.
 (Note: At this point in the valuation, r^* is the root rate, 3.50%. Therefore, $\$99.691 = 1/2(\$99.214 + \$100.169)$.)

Step 5: Compare the value in Step 4 to the bond's market value. If the two values are the same, then the r_1 used in this trial is the one we seek. If, instead, the value found in Step 4 is not equal to the market value of the bond, then r_1 in this trial is not the 1-year rate that is consistent with the current yield curve. In this case, the five steps are repeated with a different value for r_1 .

When r_1 is 4.75%, a value of \$99.691 results in Step 4, which is less than the observed market price of \$100. Therefore, 4.75% is too large and the five steps must be repeated trying a lower rate for r_1 .

Let's jump right to the correct rate for r_1 in this example and rework Steps 1 through 5. This occurs when r_1 is 4.4448%. The corresponding binomial tree is shown in Figure 5. The value at the root is equal to the market value of the 2-year issue (par).

We can "grow" this tree for one more year by determining r_2 . Now we will use the 3-year on-the-run issue, the 4.7% coupon bond, to get r_2 . The same five steps are used in an iterative

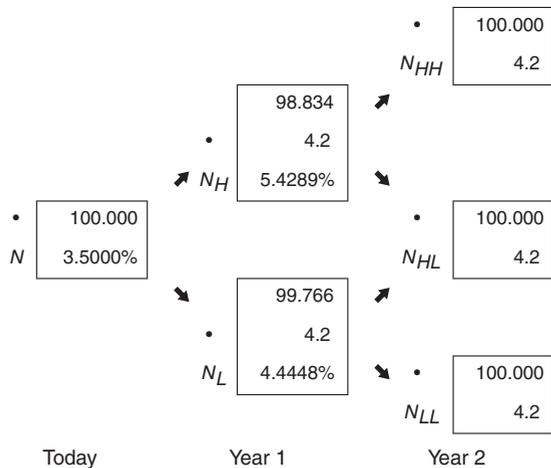


Figure 5 The 1-Year Rates for Year 1 Using the 2-Year 4.2% On-the-Run Issue

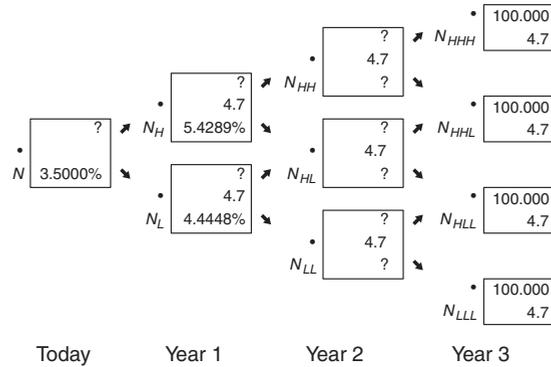


Figure 6 Information for Deriving the 1-Year Rates for Year 2 Using the 3-Year 4.7% On-the-Run Issue

process to find the 1-year rates in the tree two years from now. Our objective is now to find the value of r_2 that will produce a bond value of \$100. Note that the two rates one year from now of 4.4448% (the lower rate) and 5.4289% (the higher rate) do not change. These are the fair rates for the tree one year forward.

The problem is illustrated in Figure 6. The cash flows from the 3-year, 4.7% bond are in place. All we need to perform a valuation are the rates at the start of Year 3. In effect, we need to find r_2 such that the bond prices at par. Again, an arbitrary starting point is selected, and an iterative process produces the correct rate.

The completed version of Figure 6 is found in Figure 7. The value of r_2 , or equivalently

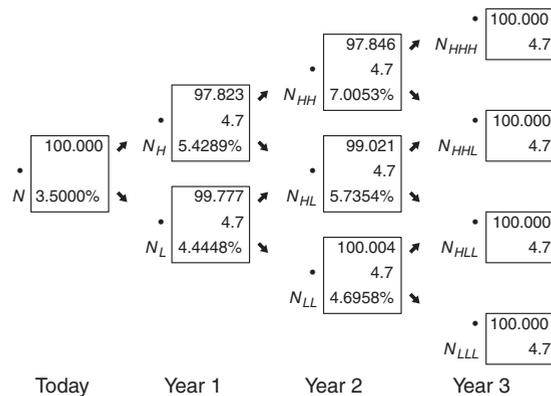


Figure 7 The 1-Year Rates for Year 2 Using the 3-Year 4.7% On-the-Run Issue

$r_{2,LL}$, which will produce the desired result is 4.6958%. The corresponding rates $r_{2,HL}$ and $r_{2,HH}$ would be 5.7354% and 7.0053%, respectively. To verify that these are the correct 1-year rates two years from now, work backwards from the four nodes at the right of the tree in Figure 7. For example, the value in the box at N_{HH} is found by taking the value of \$104.7 at the two nodes to its right and discounting at 7.0053%. The value is \$97.846. Similarly, the value in the box at N_{HL} is found by discounting \$104.70 by 5.7354% and at N_{LL} by discounting at 4.6958%.

USING THE LATTICE FOR VALUATION

To illustrate how to use the lattice for valuation purposes, consider a 6.5% option-free bond with four years remaining to maturity. Since this bond is option-free, it is not necessary to use the lattice model to value it. All that is necessary to obtain an arbitrage-free value for this bond is to discount the cash flows using the spot rates obtained from bootstrapping the yield curve shown in Table 1. (All calculations are highly sensitive to the number of decimal places chosen.) The spot rates are as follows:

1-year	3.5000%
2-year	4.2147%
3-year	4.7345%
4-year	5.2707%

Discounting the 6.5% 4-year option-free bond with a par value of \$100 at the above spot rates would give a bond value of \$104.643.

Figure 8 contains the fair tree for a four-year valuation. Figure 9 shows the various values in the discounting process using the lattice in Figure 8. The root of the tree shows the bond value of \$104.643, the same value found by discounting at the spot rate. This demonstrates that the lattice model is consistent with the valuation of an option-free bond when using spot rates.

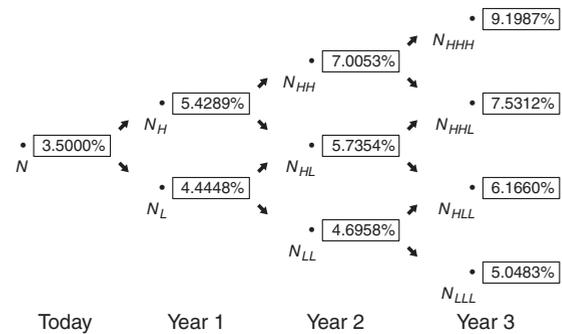


Figure 8 Binomial Interest Rate Tree for Valuing up to a 4-Year Bond for Issuer (10% Volatility Assumed)

The lesson here can be applied to more complex instruments, those with option features that require the lattice-based process for proper valuation and derivatives such as swaptions. Regardless of the security or derivative to be valued, the generation of the lattice follows the same no-arbitrage principles outlined here. Subsequently, cash flows are determined at each node, the recursive valuation process undertaken to arrive at fair values. Hence, a single lattice and a valuation process prove to be robust means for obtaining fair values for a wide variety of fixed income instruments.

KEY POINTS

- The complication in valuing bonds with embedded options and option-type derivatives is that cash flows depend on interest rates in the future.
- In practice, several interest rate models have been employed to construct an interest rate lattice. In each case, interest rates can realize one of several possible levels when we move from one period to the next. There are binomial lattices (two possible rates in the next period), trinomial lattices (three possible rates in the next period), and even more complex models that allow more than three possible rates in the next period.

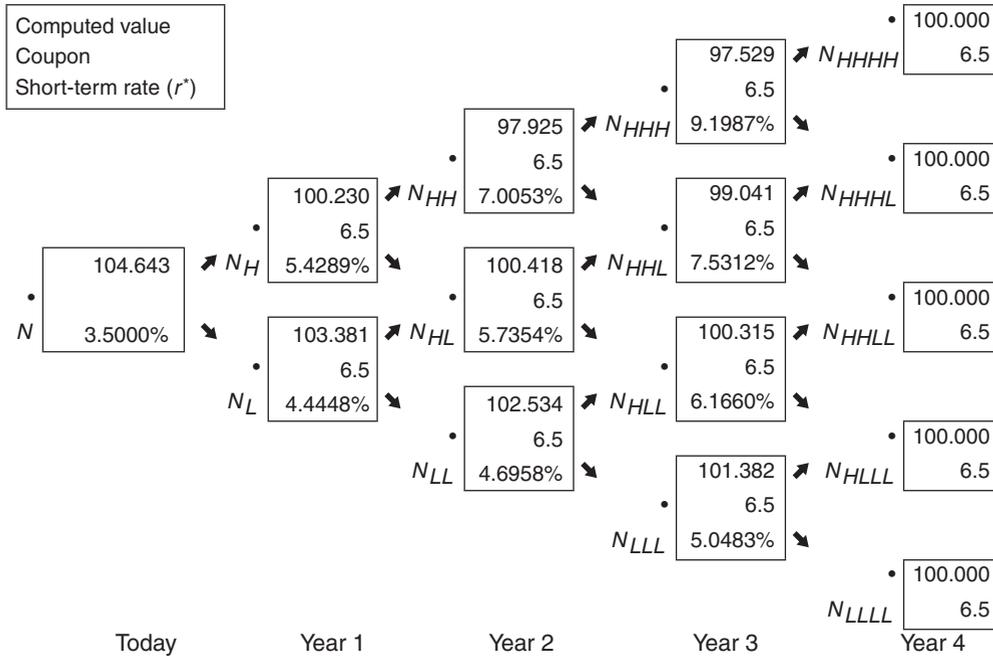


Figure 9 Valuing an Option-Free Bond with Four Years to Maturity and a Coupon Rate of 6.5% (10% Volatility Assumed)

- Several models have been developed to value bonds with embedded options and option-type interest rate derivatives, the most common model being a one-factor model.
- The lattice framework uses an arbitrage-free interest rate lattice or tree to generate the cash flows over the life of the financial instrument and then to determine the present value of the cash flow. The present value of the cash flow is then the fair value of the financial instrument.
- The lattice must be constructed so as to be consistent with (that is, calibrated to) the observed market value of an on-the-run option-free issue.

NOTE

1. For an extensive discussion of the application to the valuation of embedded options in bonds see Kalotay, Williams, and Fabozzi (1993), and for the application to interest rate swaptions see Fabozzi and Buetow (2000).

REFERENCES

Fabozzi, F. J., and Buetow, G. W. (2000). *Valuation of Interest Rate Options and Swaptions*. Hoboken, NJ: John Wiley & Sons.

Kalotay, A. J., Williams, G. O., and Fabozzi, F. J. (1993). A model for the valuation of bonds and embedded options. *Financial Analysts Journal* 49, 3: 35–46.

Using the Lattice Model to Value Bonds with Embedded Options, Floaters, Options, and Caps/Floors

FRANK J. FABOZZI, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

ANDREW KALOTAY, PhD
President, Andrew Kalotay Associates

MICHAEL DORIGAN, PhD
Senior Quantitative Analyst, PNC Capital Advisors

Abstract: In principle, the valuation of a financial instrument is straightforward: It is the present value of the expected cash flow. For fixed income securities, the expected cash flow, ignoring the possibility of default, is the periodic interest payments and the maturity value. The interest rates used to discount the expected cash flows are obtained from an appropriate benchmark spot rate curve. When a fixed-rate or floating-rate bond has an interest-sensitive embedded option such as a call option, put option, or a cap in the case of a floater, the expected cash flow will be dependent on future interest rates. To value fixed income securities with embedded options, the lattice framework is the standard tool in practice. The same lattice-based framework is also used to value interest-sensitive derivatives such as options, caps, and floors.

We will demonstrate in this entry how the *lattice* framework provides a robust means for valuing fixed-rate and floating-rate bonds and interest rate derivatives. In addition, we extend the application of the interest rate tree to the calculation of the option-adjusted spread, as well as the effective duration and convexity of a fixed income instrument. The model described below was first introduced by Kalotay, Williams, and Fabozzi (1993).

FIXED-COUPON BONDS WITH EMBEDDED OPTIONS

The valuation of bonds with embedded options proceeds in the same fashion as in the case of an *option-free bond*. However, the added complexity of an embedded option requires an adjustment to the cash flows on the tree depending on the structure of the option. A decision on whether to call or put must be made at nodes

on the tree where the option is eligible for exercise. Examples for both *callable* and *puttable* bonds follow. The analysis can be extended to cases where there are several embedded options such as a callable bond with an accelerated sinking fund provision.

Valuing a Callable Bond

In the case of a call option, the call will be made when the present value (PV) of the future cash flows is greater than the call price at the node where the decision to exercise is being made. Effectively, the following calculation is made:

$$V_t = \text{Min}[\text{Call Price}, \text{PV}(\text{Future Cash Flows})]$$

where V_t represents the PV of future cash flows at the node. This operation is performed at each node where the bond is eligible for call.

For example, consider a 6.5% bond with four years remaining to maturity that is callable in one year at \$100. We will value this bond, as well as the other instruments in this entry, using a binomial tree. The on-the-run yield curve for the issuer used to construct the tree is given in Table 1. The methodology for constructing the *binomial interest rate tree* from the yield curve is not discussed here but is explained in Entry 16. Application of the methodology results in the binomial interest rate tree in Figure 1. In constructing the binomial tree in Figure 1, it is assumed that interest rate volatility is 10% and that cash flows occur at the end of the year. This binomial tree will be used throughout this entry.

Figure 2 shows two values are now present at each node of the binomial tree. The discounting process is used to calculate the first of the two values at each node. The second value is the value based on whether the issue will be called. To simplify the analysis, it is assumed that the

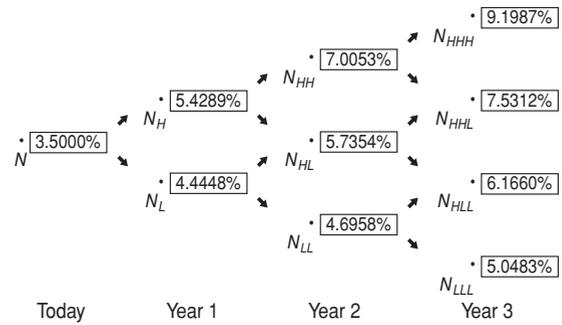


Figure 1 Binomial Interest Rate Tree for Valuing up to a Four-Year Bond for Issuer (10% Volatility Assumed)

issuer calls the issue if the PV of future cash flows exceeds the call price. This second value is incorporated into the subsequent calculations.

In Figure 3 certain nodes from Figure 2 are highlighted. Panel (a) of the figure shows nodes where the issue is not called (based on the simple call rule used in the illustration) in year 2 and year 3. The values reported in this case are the same as in the valuation of an option-free bond. Panel (b) of the figure shows some nodes where the issue is called in year 2 and year 3. Notice how the methodology changes the cash flows. In year 3, for example, at node N_{HLL} the *recursive valuation process* produces a PV of 100.315.¹ However, given the call rule, this issue would be called. Therefore, 100 is shown as the second value at the node and it is this value that is then used as the valuation process continues. Taking the process to its end, the value for this callable bond is 102.899.

The value of the call option is computed as the difference between the value of an optionless bond and the value of a callable bond. In our illustration, the value of the option-free bond can be shown to be 104.643. The value of the callable bond is 102.899. Hence, the value of the call option is 1.744 (=104.634 – 102.899).

Table 1 Issuer Par Yield Curve

Maturity	Par Rate	Market Price
1 year	3.50%	100
2 years	4.20%	100
3 years	4.70%	100
4 years	5.20%	100

Valuing a Puttable Bond

A puttable bond is one in which the bondholder has the right to force the issuer to pay off the

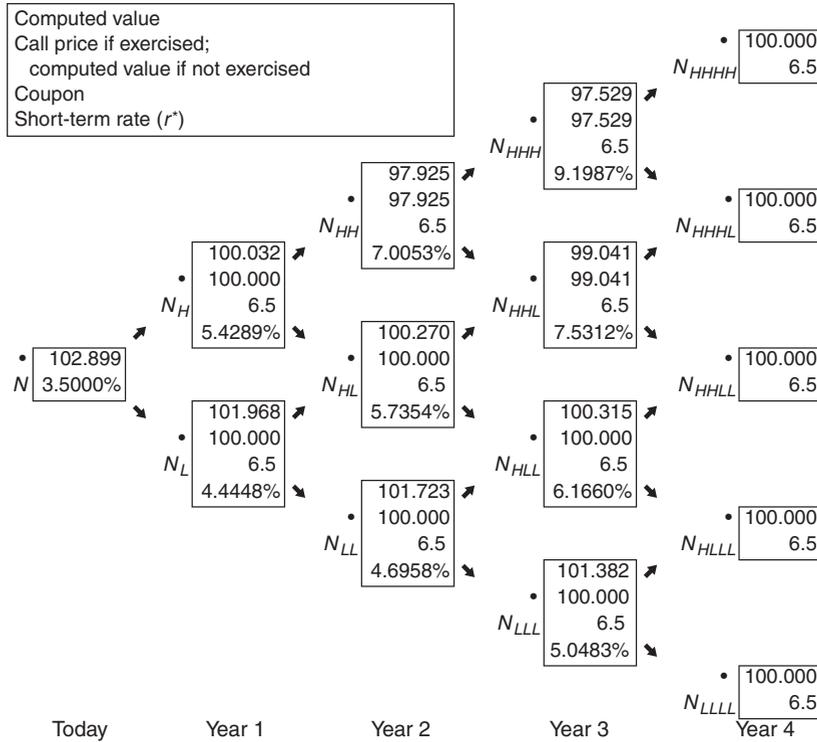


Figure 2 Valuing a Callable Bond with Four Years to Maturity, a Coupon Rate of 6.5%, and Callable after the First Year at 100 (10% Volatility Assumed)

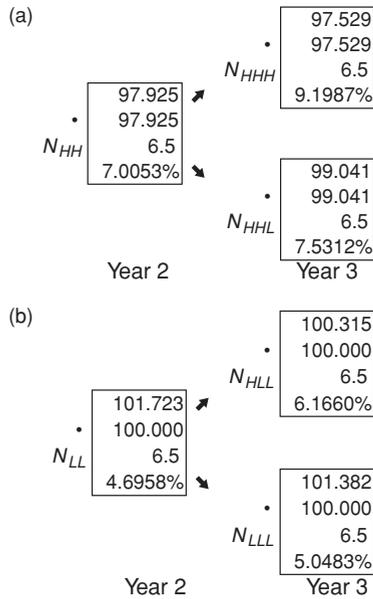


Figure 3 Highlighting Nodes in Years 2 and 3 for a Callable Bond: (a) Nodes Where the Call Option Is Not Exercised and (b) Selected Nodes Where the Call Option Is Exercised

bond prior to the maturity date. The analysis of the puttable bond follows closely that of the callable bond. In the case of the puttable, we must establish the rule by which the decision to put is made. The reasoning is similar to that for the callable bond. If the PV of the future cash flows is less than the put price (that is, par), then the bond will be put. In equation form,

$$V_t = \text{Max}[\text{Put Price}, \text{PV}(\text{Future Cash Flows})]$$

Figure 4 is analogous to Figure 3. It shows the binomial tree with the values based on whether or not the investor exercises the put option at each node. The bond is puttable any time after the first year at par. The value of the bond is 105.327. Note that the value is greater than the value of the corresponding option-free bond.

With the two values in hand, we can calculate the value of the put option. Since the value of the puttable bond is 105.327 and the value of the

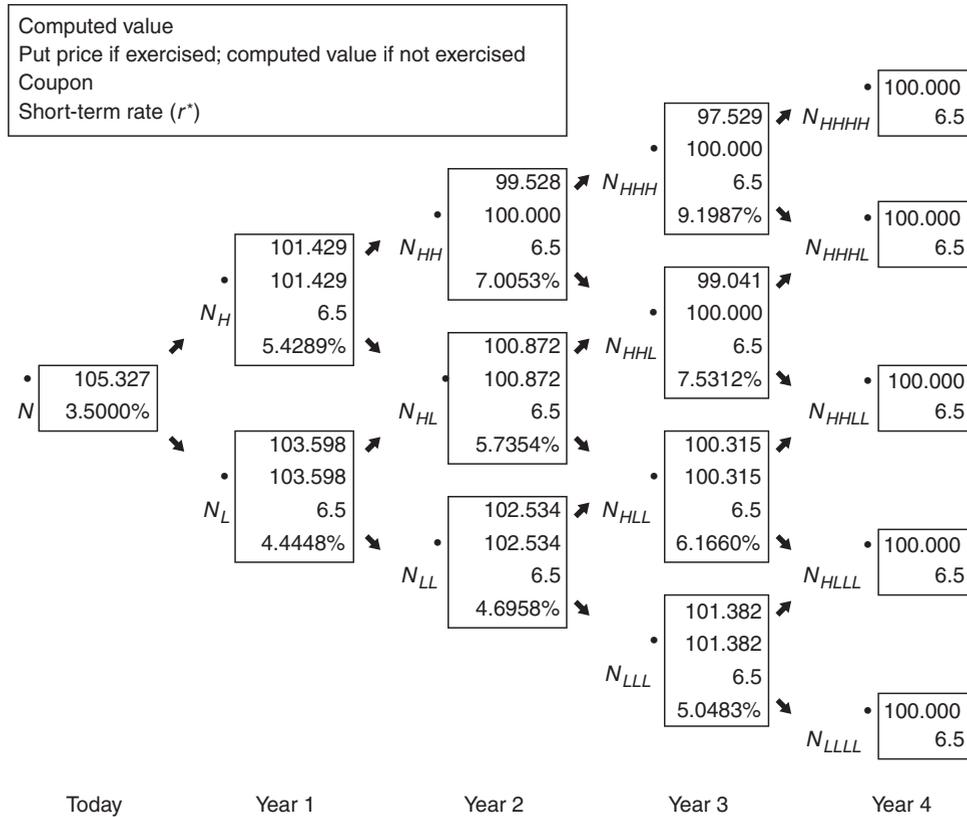


Figure 4 Valuing a Puttable Bond with Four Years to Maturity, a Coupon Rate of 6.5%, and Puttable after the First Year at 100 (10% Volatility Assumed)

corresponding option-free bond is 104.643, the value of the embedded put option purchased by the investor is effectively 0.684.

Suppose that a bond is both puttable and callable. The procedure for valuing such a structure is to adjust the value at each node to reflect whether the issue would be put or called. Specifically, at each node there are two decisions about the exercising of an option that must be made. If it is called, the value at the node is replaced by the call price. The valuation procedure then continues using the call price at that node. If the call option is not exercised at a node, it must be determined whether or not the put option will be exercised. If it is exercised, then the put price is substituted at that node and is used in subsequent calculations.

FLOATING-COUPON BONDS WITH EMBEDDED OPTIONS

Simple discounted cash flow methods of analysis fail to handle floaters with embedded or option-like features. In this section we demonstrate how to use the lattice model to value (1) a *capped floater*, and (2) a *callable capped floater*. We will streamline the notation used in the binomial tree in the figures shown in this section.

Valuing Capped Floating-Rate Bonds

Consider a floating-rate bond with a coupon indexed to the 1-year rate (the reference rate) plus a spread. For our purposes, assume a 25

basis point (bp) spread to the reference rate. The coupon adjusts at each node to reflect the level of the reference rate plus the spread.

Using the same valuation method as before, we can find the value at each node. Recall the value of the bond is 100 (par) at the end of year 4. Consider N_{HLL} .

$$N_{HLL} = \frac{1}{2} \left[\frac{100 + 6.416}{1.06166} + \frac{100 + 6.416}{1.06166} \right] = 100.235$$

Stepping back one period

$$N_{LL} = \frac{1}{2} \left[\frac{100.235 + 4.9458}{1.046958} + \frac{100.238 + 4.9458}{1.046958} \right] = 100.465$$

Following this same procedure, we arrive at the price of 100.893. How would this

change if the interest rate on the bond were capped?

Assume that the cap is 7.25%. In Figure 5 we've taken the tree from Figure 1 and, as was the case with the optionless fixed-coupon bond, at each node we've entered the cash flow expected at the end of each period based on the reset formula. As rates move higher there is a possibility that the current reference rate exceeds the cap. Such is the case at N_{HHH} and N_{HLL} . The coupon is subject to the following constraint:

$$C_t = \text{Min}[r_t, 7.25\%]$$

As a result of the cap, the value of the bond in the upper nodes at $t = 3$ falls below par. For example,

$$N_{HHH} = \frac{1}{2} \left[\frac{100 + 7.25}{1.091987} + \frac{100 + 7.25}{1.09198} \right] = 98.215$$

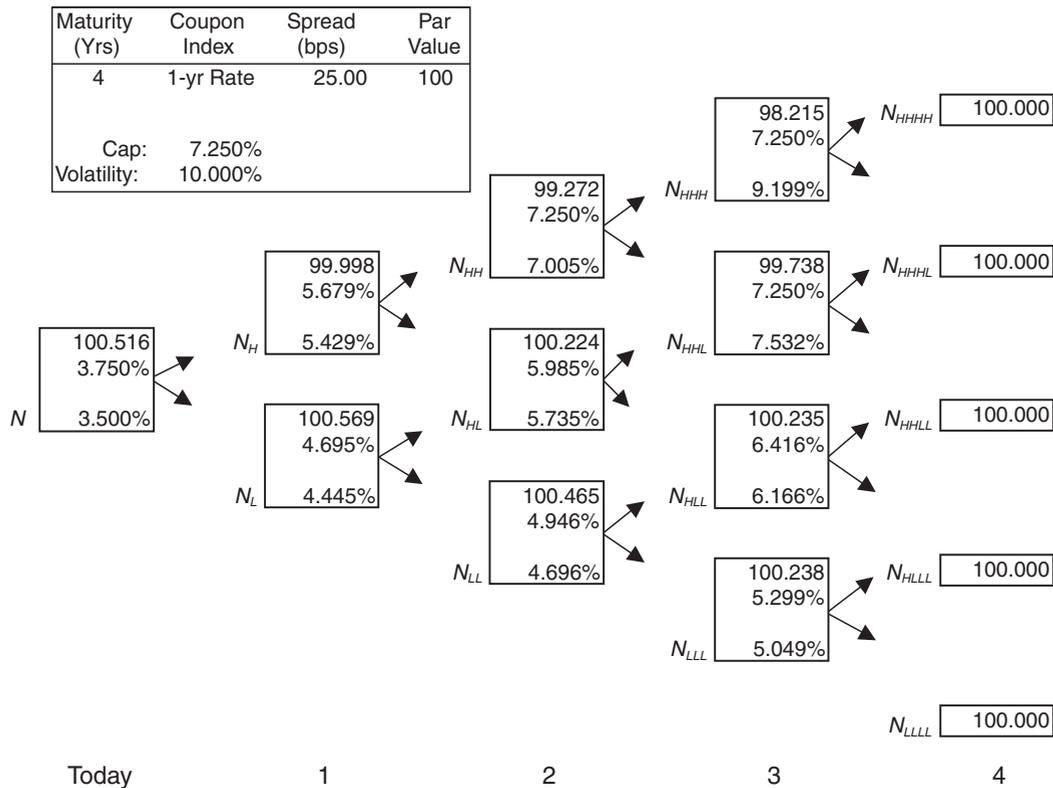


Figure 5 Valuation of a Capped Floating-Rate Bond

Valuing recursively through the tree, we arrive at the current value of the capped floater, 100.516, a value lower than the plain vanilla floater. This last calculation gives us a means for pricing the embedded option. Without a cap, the bond is priced at 100.893. The difference between these two prices is the value of the cap, 0.377. It is important to note that the price of the cap is volatility dependent. Any change in the volatility would result in a different valuation for the cap. The greater the volatility, the higher the price of the option, and vice versa.

We can extend the application of the lattice to the initial pricing of securities. What if an issuer wanted to offer this bond at par? In such a case, an adjustment has to be made to the coupon. To lower the price from 100.516 to par, a lower spread over the reference rate is offered to investors. Figure 6 shows the relationship between the spread over the 1-year reference rate and the bond price. At a spread of 8.70 bps over the 1-year reference rate, the capped floater in Figure 5 will be priced at par. Again, the spread of 8.7 bps is volatility dependent.

Callable Capped Floating-Rate Bonds

Now consider a call option on the capped floater. As was the case for a fixed-coupon bond, we must be careful to specify the appropriate rules for calling the bond on the valuation tree. It turns out that the rule is the same for floaters and fixed-coupon bonds. Any time the bond

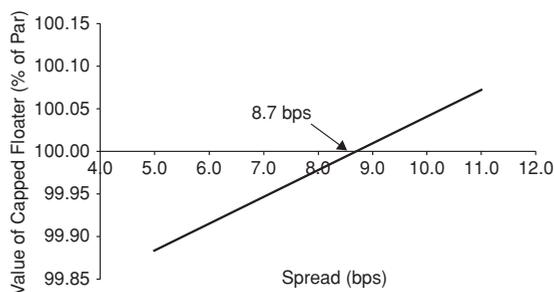


Figure 6 Spread to Index to Price Cap at Par

has a PV above par at a node where the bond is callable, the bond will be called. (Here we assume a par call to simplify the illustration.)

Before we get into the details, it is important to motivate the need for a call on a floating-rate bond. The value of a cap to the issuer increases as market rates near the cap and there is the potential for rates to exceed the cap prior to maturity. As rates decline, so does the value of the cap. The problem for the issuer in the event of low rates is the additional basis-point spread it is paying for a cap that now has little or no value. Thus, when rates decline, a call has value to the issuer because it can call and reissue at a different spread.

Suppose that the capped floater is callable at par anytime after the first year. Figure 7 provides details on the effect of the call option on valuation of the capped floater. Again, for a callable bond, when the present value exceeds par in a recursive valuation model, the bond is called. In the case of our 4-year bond, in Figure 7 the value of the bond at several lower nodes is now 100, the call price. The full effect of the call option on price is evident with today's price for the bond moving to 99.9140.

The by-product of this analysis is the value of the call option on a capped floater. We now have the fair value of the capped floater versus the callable capped floater. So the call option has a value of $100.516 - 100.189 = 0.327$.

How would one structure the issue so that it is priced at par? We have to offer a lower spread over the floating rate than the holder is already receiving for accepting the cap. In this case, we need to move the total spread over the 1-year floating rate to 13.37 bps. Figure 8 shows the relationship between spread and value.

VALUING CAPS AND FLOORS

An *interest rate cap* is nothing more than a package or strip of options. More specifically, a cap is a strip of European options on interest rates.

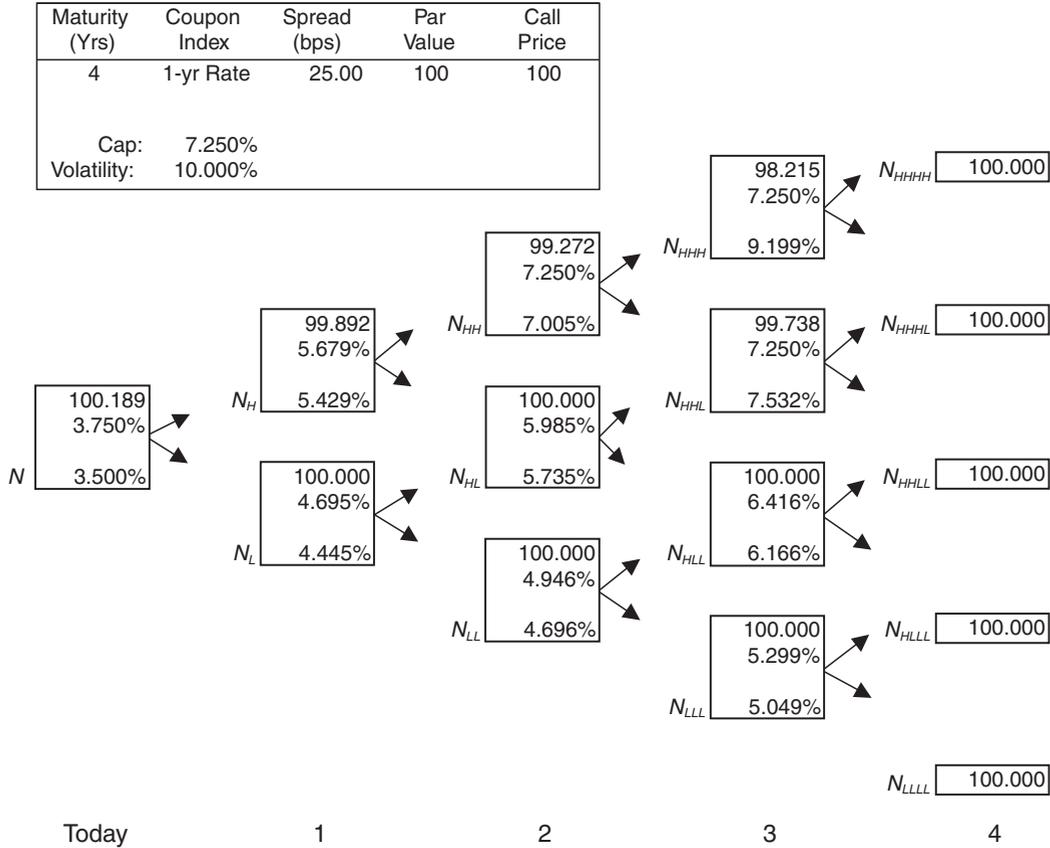


Figure 7 Valuation of a Capped Floating-Rate Bond

Thus, to value a cap, the value of each period's cap, called a caplet, is found and all the caplets are then summed.

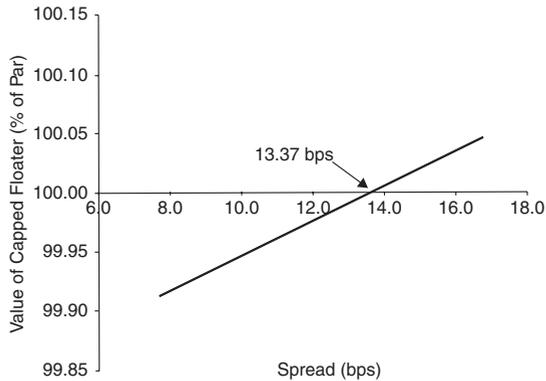


Figure 8 Spread to Index to Price Callable Cap at Par

In order to value caps and floors, a modification of the lattice framework is required. The modification is necessary because of the timing of the payments for a cap and floor: Settlement for the typical cap and floor is paid in arrears. Payment in arrears means that the interest rate paid is determined at the beginning of the period, but the actual payment is made at the end of the period (that is, beginning of the next period). This modification complicates the notation and will not be made here but is explained in Fabozzi (2006).

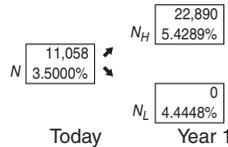
To illustrate, we once again use the binomial tree given in Figure 1 to value a cap. Consider a 5.2% 3-year cap with a notional amount of \$10 million. The reference rate is the 1-year rate. The payoff for the cap is annual.

The three panels in Figure 9 show how this cap is valued by valuing the three caplets. The value for the caplet for any year, say year X, is found as follows. First, calculate the payoff in year X at each node as either:

1. Zero if the one-year rate at the node is less than or equal to 5.2%, or
2. The notional amount of \$10 million times the difference between the 1-year rate at the node and 5.2% if the 1-year rate at the node is greater than 5.2%.

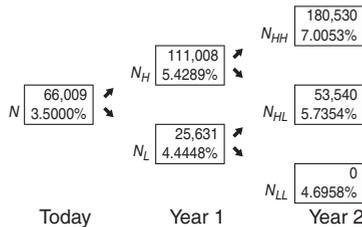
Assumptions:
 Cap rate: 5.2%
 Notional amount: \$10,000,000
 Payment frequency: Annual

Panel A: The Value of the Year 1 Caplet



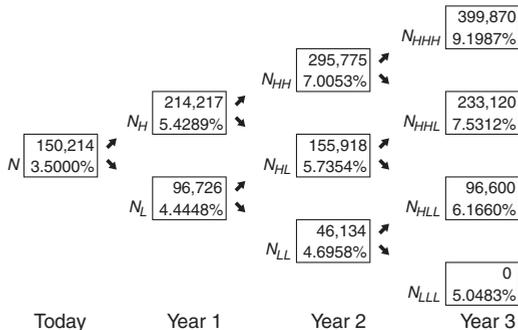
Value of Year 1 caplet = \$11,058

Panel B: The Value of the Year 2 Caplet



Value of Year 2 caplet = \$66,009

Panel C: The Value of the Year 3 Caplet



Value of Year 3 caplet = \$150,214

Summary: Value of 3-Year Cap = \$11,058 + \$66,009 + \$150,214 = \$227,281

Note on calculations: Payoff in last box of each figure is \$10,000,000 Maximum [(Rate at node - 5.2%), 0]

Then, the recursive valuation process is used to determine the value of the year X caplet.

For example, consider the year 3 caplet. At the top node in year 3 of Panel C of Figure 9, the 1-year rate is 9.1987%. Since the 1-year rate at this node exceeds 5.2%, the payoff in year 3 is:

$$\$10,000,000 \times (0.091987 - 0.052) = \$399,870$$

For node N_{HH} we look at the value for the cap at the two nodes to its right, N_{HHH} and N_{HHL} . Discounting the values at these nodes, \$399,870 and \$233,120, by the interest rate from the binomial tree at node N_{HH} , 7.0053%, we arrive at a value of \$295,755. That is,

$$\begin{aligned} \text{Value at } N_{HH} &= [\$399,870/(1.070053) \\ &\quad + \$233,120(1.070053)]/2 \\ &= \$295,775 \end{aligned}$$

The values at nodes N_{HH} and N_{HL} are discounted at the interest rate from the binomial tree at node N_H , 5.4289%, and then the value is computed. That is,

$$\begin{aligned} \text{Value at } N_H &= [\$295,775/(1.054289) \\ &\quad + \$155,918/(1.054289)]/2 \\ &= \$214,217 \end{aligned}$$

Finally, we get the value at the root, node N , which is the value of the year 3 caplet found by discounting the value at N_H and N_L by 3.5% (the interest rate at node N). Doing so gives:

$$\begin{aligned} \text{Value at } N &= [\$214,217/(1.035) \\ &\quad + \$96,726/(1.035)]/2 \\ &= \$150,214 \end{aligned}$$

Following the same procedure, the value of the year 2 caplet is \$66,009 and the value of the year 1 caplet is \$11,058. The value of the cap is then the sum of the three caplets.

Thus, the value of the cap is \$227,281, found by adding \$11,058, \$66,009, and \$150,214. The

Figure 9 Valuation of a Three-Year 5.2% Cap (10% Volatility Assumed)

valuation of an *interest rate floor* is done in the same way.

VALUATION OF TWO MORE EXOTIC STRUCTURES

The lattice-based recursive valuation methodology is robust. To further support this claim, we address the valuation of two more exotic structures—the step-up callable note and the range floater.

Valuing a Step-Up Callable Note

Step-up callable notes are callable instruments whose coupon rate is increased (that is, “stepped up”) at designated times. When the coupon rate is increased only once over the security’s life, it is said to be a single step-up callable note. A multiple step-up callable note is a step-up callable note whose coupon is increased more than one time over the life of the security. Valuation using the lattice model is similar to that for valuing a callable bond

described above except that the cash flows are altered at each node to reflect the coupon characteristics of a step-up note.

Suppose that a four-year step-up callable note pays 4.25% for two years and then 7.5% for two more years. Assume that this note is callable at par at the end of year 2 and year 3. We will use the binomial tree given in Figure 1 to value this note.

Figure 10 shows the value of the note if it were not callable. The valuation procedure is the recursive valuation from Figure 2. The coupon in the box at each node reflects the step-up terms. The value is 102.082. Figure 11 shows that the value of the single step-up callable note is 100.031. The value of the embedded call option is equal to the difference in the optionless step-up note value and the step-up callable note value, 2.051.

Now we move to another structure where the coupon floats with a reference rate, but is restricted. In this next case, a range is set in which the bond pays the reference rate when the rate falls within a specified range, but outside the range no coupon is paid.

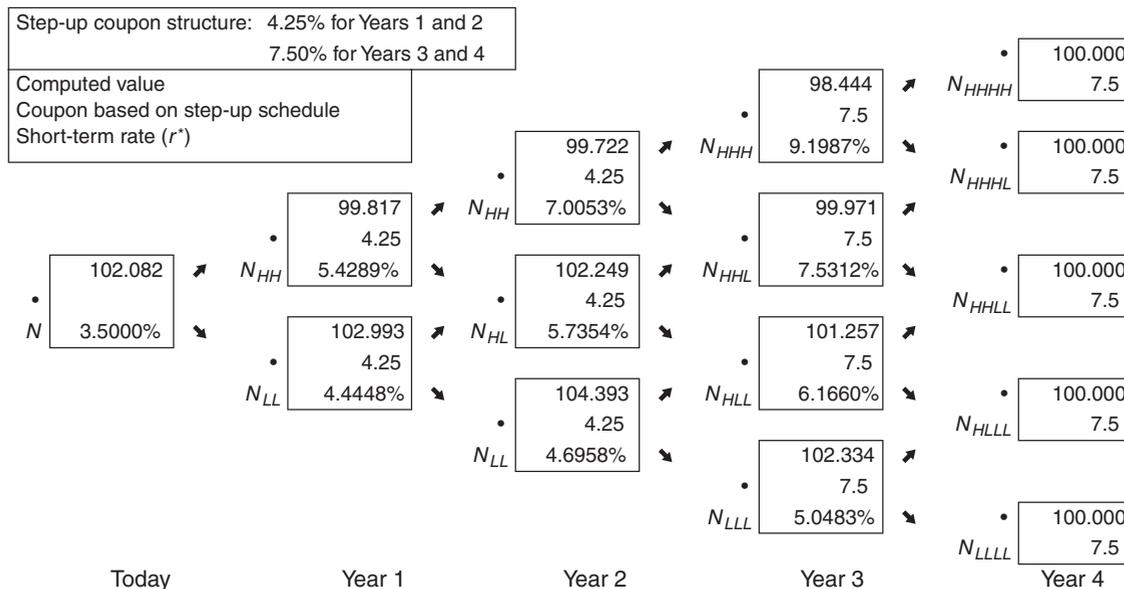


Figure 10 Valuing a Single Step-Up Noncallable Note with Four Years to Maturity (10% Volatility Assumed)

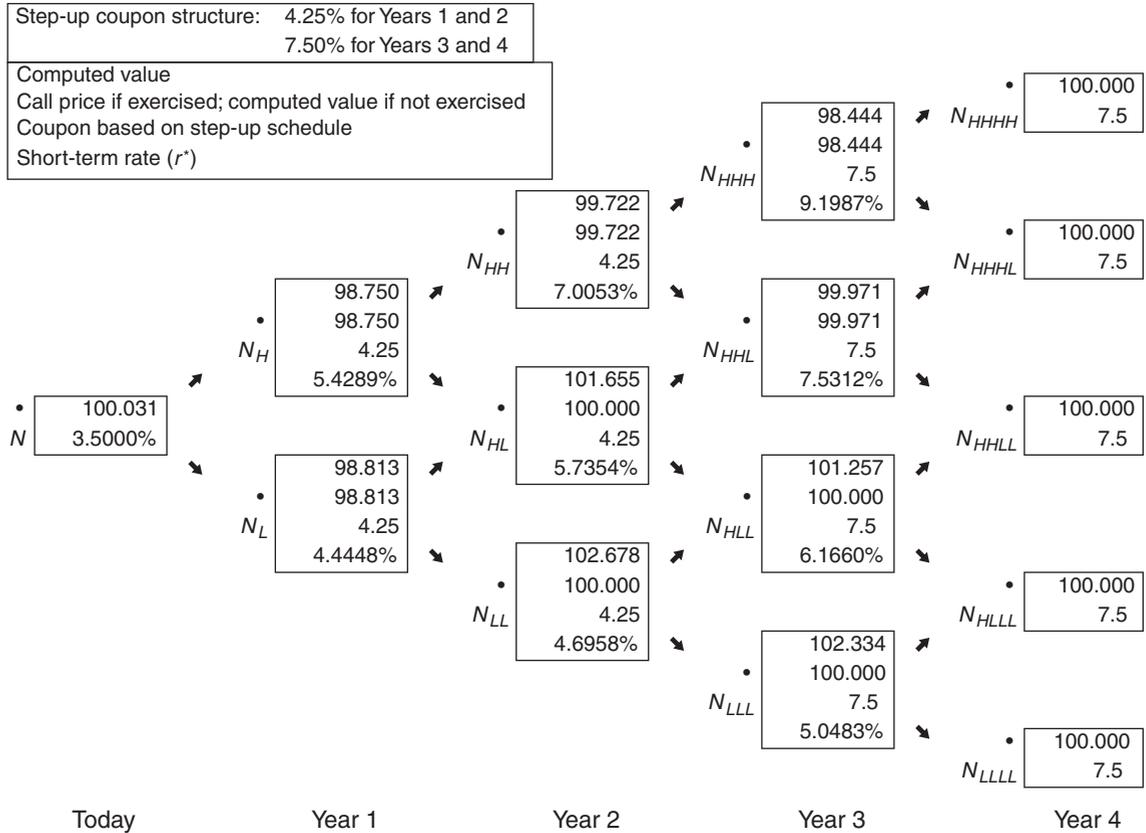


Figure 11 Valuing a Single Step-Up Callable Note with Four Years to Maturity, Callable in Two Years at 100 (10% Volatility Assumed)

Valuing a Range Note

A *range note* is a security that pays the reference rate only if the rate falls within a band. If the reference rate falls outside the band, whether the lower or upper boundary, no coupon is paid. Typically, the band increases over time.

To illustrate, suppose that the reference rate is, again, the 1-year rate and the note has three years to maturity. Suppose further that the band (or coupon schedule) is defined as in Table 2. Figure 12 shows the interest rate tree and the cash flows expected at the end of each year. Ei-

Table 2 Coupon Schedule (Bands) for a Range Note

	Year 1	Year 2	Year 3
Lower Limit	3.00%	4.00%	5.00%
Upper Limit	5.00%	6.25%	8.00%

ther the 1-year reference rate is paid, or nothing. In the case of this 3-year note, there is only one state in which no coupon is paid. Using our recursive valuation method, we can work back through the tree to the current value, 98.963.

VALUING AN OPTION ON A BOND

Thus far we have seen how the lattice can be used to value bonds with embedded options. The same tree can be used to value a stand-alone *option on a bond*.

To illustrate how this is done, consider a 2-year American call option on a 6.5% 2-year Treasury bond with a strike price of 100.25 which will be issued two years from now. We will assume that the on-the-run Treasury yields are

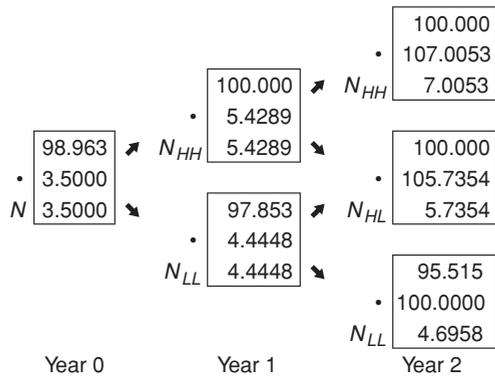


Figure 12 Valuation of a Three-Year Range Floater

those represented in Figure 13. Within the binomial tree we find the value of the Treasury bond at each node. Figure 14 shows the value of our hypothetical Treasury bond (excluding coupon interest) at each node at the end of year 2.

The decision rule at a node for determining the value of an option on a bond depends on whether or not the call or put option being valued is in the money. Moreover, the exercise decision is only applied at the expiration date. That is, a call option will be exercised at the option's expiration date if the bond's value at a node is greater than the strike price. In the case of a put option, the option will be exercised if the strike price at a node is greater than the bond's value (that is, if the put option is in the money).

Three values for the underlying 2-year bond are shown in Figure 14: 97.925, 100.418, and

102.534. Given these three values, the value of a call option with a strike price of 100.25 can be determined at each node. For example, if in year 2 the price of this Treasury bond is 97.925, then the value of the call option would be zero. In the other two cases, since the value at the end of year 2 is greater than the strike price, the value of the call option is the difference between the price of the bond at the node and 100.25.

Given these values, the binomial tree is used to find the present value of the call option using recursive valuation. The discount rates are the now familiar 1-year forward rates from the binomial tree. The expected value at each node for year 1 is found by discounting the call option value from year 2 using the rate at the node. Move back one more year to "Today." The value of the option is 0.6056.

The same procedure is used to value a put option on a bond.

EXTENSIONS

We next demonstrate how to compute the option-adjusted spread, effective duration, and the convexity for a fixed income instrument with an embedded option.

Option-Adjusted Spread

We have concerned ourselves with valuation to this point. However, financial market transactions determine the actual price for a fixed

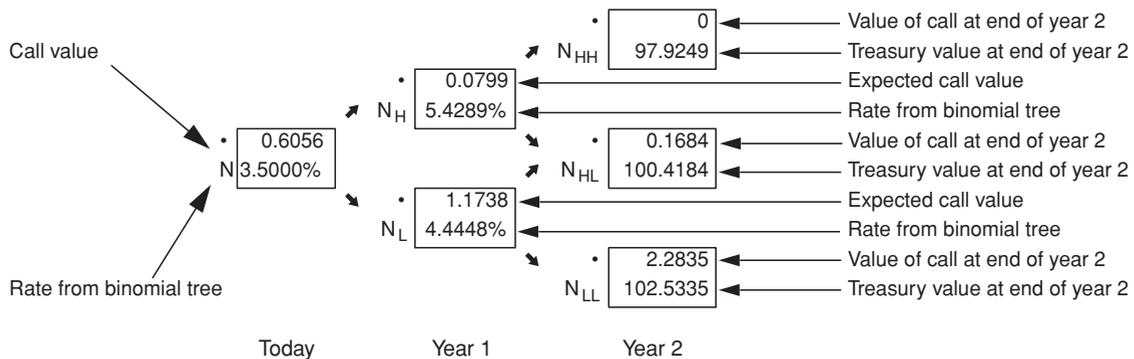
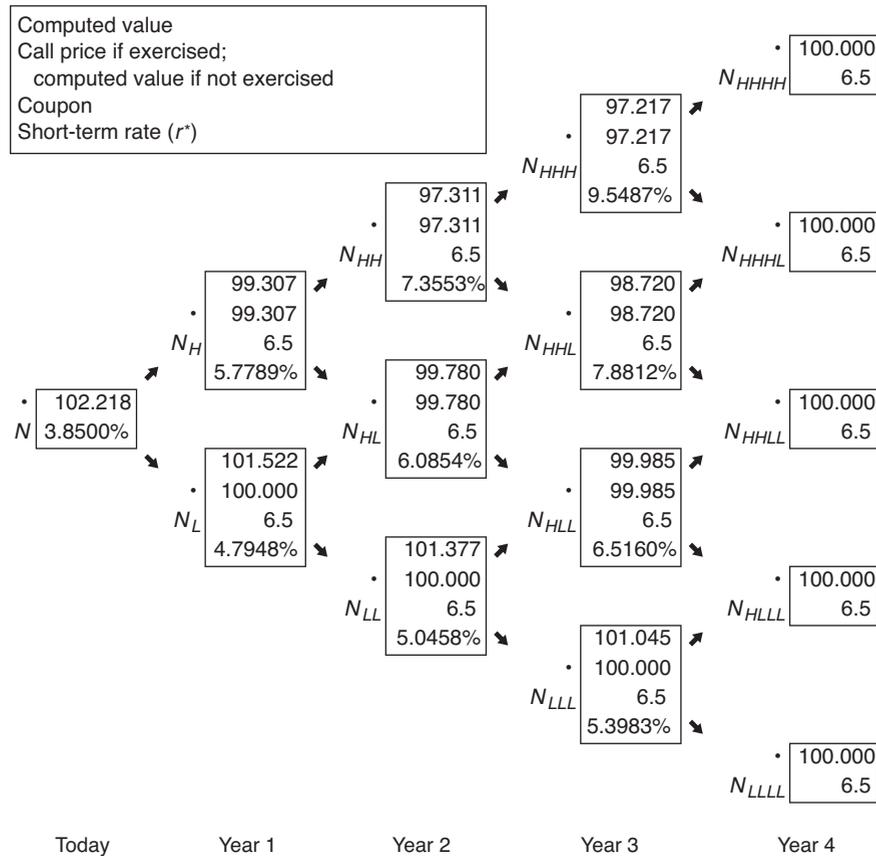


Figure 13 Using the Arbitrage-Free Binomial Method
 Expiration: 2 years; Strike Price: 100.25; Current Price: 104.643; Volatility Assumption: 10%



* Each 1-year rate is 35 basis points greater than in Figure 1.

Figure 14 Demonstration That the Option-Adjusted Spread is 35 Basis Points for a 6.5% Callable Bond Selling at 102.218 (Assuming 10% Volatility)

income instrument, not a series of calculations on an interest rate lattice. If markets are able to provide a meaningful price (usually a function of the liquidity of the market in which the instrument trades), this price can be translated into an alternative measure of relative value, the *option-adjusted spread* (OAS).

The OAS for a security is the fixed spread (usually measured in basis points) over the benchmark rates that equates the output from the valuation process with the actual market price of the security.² For an optionless security, the calculation of OAS is a relatively simple iterative process. The process is much more analytically challenging with the added complexity of optionality. And, just as the value of the op-

tion is volatility dependent, the OAS for a fixed income security with embedded options or an option-like interest rate product is volatility dependent.

Recall our illustration in Figure 2 where the value of a callable bond was calculated as 102.899. Suppose that we had information from the market that the price is actually 102.218. We need the OAS that equates the value from the lattice with the market price. Since the market price is lower than the valuation, the OAS is a positive spread to the rates in the figure, rates which we assume to be benchmark rates.

The solution in this case is 35 basis points, which is incorporated into Figure 14 that shows the value of the callable bond after adding

35 basis points to each rate. The simple, binomial tree provides evidence of the complex calculation required to determine the OAS for a callable bond. In Figure 2, the bond is called at N_{HLL} . However, once the tree is shifted 35 bps in Figure 14, the PV of future cash flows at N_{HLL} falls below the call price to 99.985, so the bond is not called at this node. Hence, as the lattice structure grows in size and complexity, the need for computer analytics becomes obvious.

Effective Duration and Effective Convexity

Duration and convexity provide a measure of the interest rate risk inherent in a fixed income security.³ We rely on the lattice model to calculate the *effective duration* and *effective convexity* of a bond with an embedded option and other option-like securities. The formulas for these two risk measures are given below:

$$\text{Effective duration} = \frac{V_- - V_+}{2 V_0(\Delta r)}$$

$$\text{Effective convexity} = \frac{V_+ - V_- - 2V_0}{2 V_0(\Delta r)^2}$$

where V_- and V_+ are the values derived following a parallel shift in the yield curve down and up, respectively, by a fixed spread. The model adjusts for the changes in the value of the embedded call option that result from the shift in the curve in the calculation of V_- and V_+ .

Note that the calculations must account for the OAS of the security. Below we provide the steps for the proper calculation of V_+ . The calculation for V_- is analogous.

Step 1: Given the market price of the issue, calculate its OAS.

Step 2: Shift the on-the-run yield curve up by a small number of basis points (Δr).

Step 3: Construct a binomial interest rate tree based on the new yield curve from Step 2.

Step 4: Shift the binomial interest rate tree by the OAS to obtain an “adjusted tree.” That

is, the calculation of the effective duration and convexity assumes a constant OAS.

Step 5: Use the adjusted tree in Step 4 to determine the value of the bond, V_+ .

We can perform this calculation for our 4-year callable bond with a coupon rate of 6.5%, callable at par selling at 102.218. We computed the OAS for this issue as 35 basis points. Figure 15 shows the adjusted tree following a shift in the yield curve up by 25 basis points, and then adding 35 basis points (the OAS) across the tree. The adjusted tree is then used to value the bond. The resulting value, V_+ is 101.621.

To determine the value of V_- , the same five steps are followed except that in Step 2, the on-the-run yield curve is shifted down by the same number of basis points (Δr). It can be demonstrated that for our callable bond, the value for V_- is 102.765.

The results are summarized below:

$$\Delta r = 0.0025$$

$$V_+ = 101.621$$

$$V_- = 102.765$$

$$V_0 = 102.218$$

Therefore,

$$\text{effective duration} = \frac{102.765 - 101.621}{2(102.218)(0.0025)} = 2.24$$

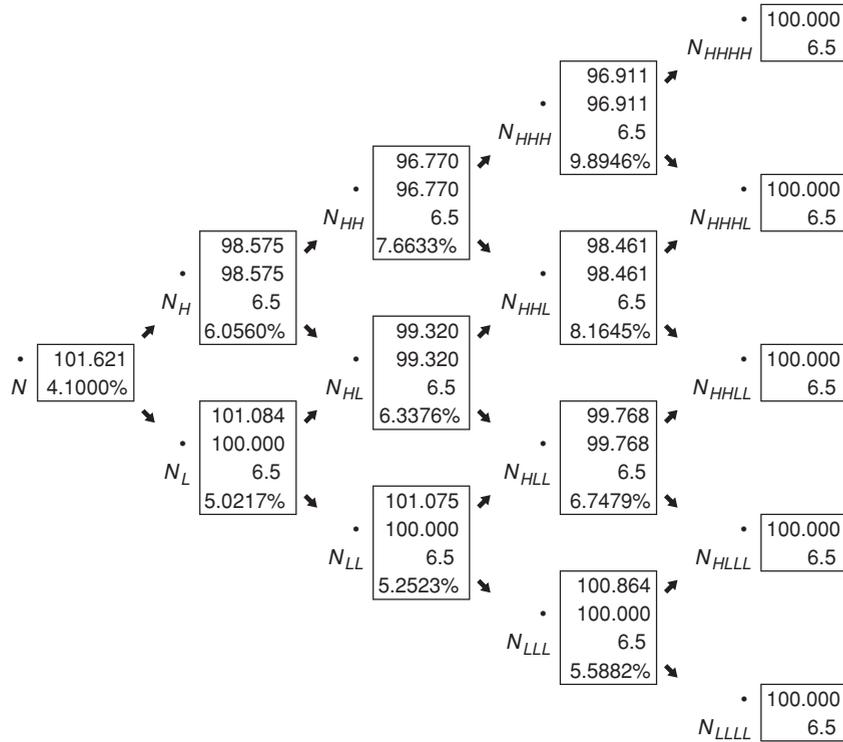
$$\text{Effective convexity} = \frac{101.621 + 102.765 - 2(102.218)}{2(102.218)(0.0025)^2}$$

$$= -39.1321$$

Notice that this callable bond exhibits negative convexity.

KEY POINTS

- The valuation of an option-free bond is straightforward. However, once there is a provision in the bond structure that grants the issuer and/or the investor an option, valuation becomes more difficult.
- The standard technology employed to value bonds with embedded options that depend



* +25 basis point shift in on-the-run yield curve.

Figure 15 Determination of V_+ for Calculating Effective Duration and Convexity*

on future interest rates, such as callable and puttable bonds, is the lattice framework.

- The initial step in the lattice approach is to generate an arbitrage-free lattice or interest rate tree from an appropriate on-the-run yield curve.
- Based on rules specified by the modeler for when an option will be exercised, a lattice of future cash flows is obtained and then valued using the interest rates in the lattice.
- The same model is used to value interest rate-sensitive derivatives such as options on bonds, interest rate caps, and interest rate caps and floors.
- Other useful analytical measures can be obtained using the lattice model. These measures include the option-adjusted spread—a measure of relative value—and effective duration and effective convexity—measures of price sensitivity to changes in interest rates.

NOTES

1. See Kalotay, Williams, and Fabozzi (1993).
2. For a discussion of OAS, see Fabozzi (1990, 2012).
3. See Fabozzi (1999, 2012) for a discussion of effective duration and convexity.

REFERENCES

Fabozzi, F. J. (1999). *Duration, Convexity, and Other Bond Risk Measures*. Hoboken, NJ: John Wiley & Sons.

Fabozzi, F. J. (2012). *Bond Markets, Analysis, and Strategies*, 8th ed. Upper Saddle River, NJ: Pearson.

Kalotay, A. J., Williams, G. O., and Fabozzi, F. J. (1993). A model for the valuation of bonds and embedded options. *Financial Analysts Journal* 49, 3: 35–46.

Understanding the Building Blocks for OAS Models

PHILIP O. OBAZEE

Senior Vice President and Head of Derivatives,
Delaware Investments

Abstract: Ubiquity of option-adjusted spread (OAS) in finance practice is remarkable, in light of the fact that there is no general consensus on its implementation. Investors in mortgage-backed (MBS) and asset-backed (ABS) securities hold a long position in noncallable bonds and short positions in prepayment (call) options. The noncallable bond is a bundle of zero coupon bonds, and the call option gives the borrower the right to prepay the loan at any time prior to the scheduled principal repayment dates. The call option component of the valuation consists of intrinsic and time values. To the extent that the option embedded in ABS/MBS is a delayed American exercise style, the time value component associated with prepayment volatility needs to be evaluated. To evaluate this option, OAS analysis uses an option-based technique to price ABS/MBS under different interest rate scenarios. Hence, OAS is the spread differential between the zero volatility spread and option value components of an ABS/MBS.

Investors and analysts continue to wrestle with the differences in option-adjusted-spread (OAS) values for securities they see from competing dealers and vendors. And portfolio managers continue to pose fundamental questions about OAS with which we all struggle in the financial industry. Some of the frequently asked questions are

- How can we interpret the difference in dealers' OAS values for a specific security?
- What is responsible for the differences?
- Is there really a correct OAS value for a given security?

In this entry, we examine some of the questions about OAS analysis, particularly the basic building block issues about OAS implementa-

tion. Because some of these issues determine "good or bad" OAS results, we believe there is a need to discuss them. To get at these fundamental issues, we hope to avoid sounding pedantic by relegating most of the notations and expressions to the endnotes.

Clearly, it could be argued that portfolio managers do not need to understand the OAS engine to use it but that they need to know how to apply it in relative value decisions. This argument would be correct if there were market standards for representing and generating interest rates and prepayments. In the absence of a market standard, investors need to be familiar with the economic intuitions and basic assumptions made by the underlying models. More important, investors need to understand what

works for their situation and possibly identify those situations in which one model incorrectly values a bond. Although pass-throughs are commoditized securities, OAS results still vary considerably from dealer to dealer and vendor to vendor. This variance is attributable to differences in the implementation of the respective OAS models.

Unlike other market measures, for example, yield to maturity and the weighted average life of a bond, which have market standards for calculating their values, OAS calculations suffer from the lack of a standard and a black-box mentality. The lack of a standard stems from the required inputs in the form of interest rate and prepayment models that go into an OAS calculation. Although there are many different interest rate models available, there is little agreement on which one to use. Moreover, there is no agreement on how to model prepayments. The black-box mentality comes from the fact that heavy mathematical machinery and computational algorithms are involved in the development and implementation of an OAS model. This machinery is often so cryptic that only a few initiated members of the intellectual tribe can decipher it. In addition, dealers invest large sums in the development of their term structures and prepayment models and, consequently, they are reluctant to share it.

In this entry, we review some of the proposed term structures and prepayments. Many of the term structure models describe “what is” and only suggest that the models could be used. Which model to use perhaps depends on the problem at hand and the resources available. In this entry, we review some of the popular term structure models and provide some general suggestions on which ones should not be used.

Investors in asset-backed securities (ABS) and mortgage-backed securities (MBS) hold long positions in noncallable bonds and short positions in call (prepayment) options. The noncallable bond is a bundle of zero-coupon bonds (e.g., Treasury strips), and the call option gives

the borrower the right to prepay the mortgage at any time prior to the maturity of the loan. In this framework, the value of MBS is the difference between the value of the noncallable bond and the value of the call (prepayment) option. Suppose a theoretical model is developed to value the components of ABS/MBS. The model would value the noncallable component, which we loosely label the zero volatility component, and the call option component. If interest rate and prepayment risks are well accounted for, and if those are the only risks for which investors demand compensation, one would expect the theoretical value of the bond to be equal to its market value. If these values are not equal, then market participants demand compensation for the unmodeled risks. One of these unmodeled risks is the forecast error associated with the prepayments. By this, we mean the actual prepayment may be faster or slower than projected by the model. Other unmodeled risks are attributable to the structure and liquidity of the bond. In this case, OAS is the market price for the unmodeled risks.

To many market participants, however, OAS indicates whether a bond is mispriced. All else being equal, given that interest rate and prepayment risks have been accounted for, one would expect the theoretical price of a bond to be equal to its market price. If these two values are not equal, a profitable opportunity may exist in a given security or a sector. Moreover, OAS is viewed as a tool that helps identify which securities are cheap or rich when the securities are relatively priced.

The zero volatility component of ABS/MBS valuation is attributable to the pure interest rate risk of a known cash flow—a noncallable bond. The forward interest rate is the main value driver of a noncallable bond. Indeed, the value driver of a noncallable bond is the sum of the rolling yield and the value of the convexity. The rolling yield is the return earned if the yield curve and the expected volatility are unchanged. Convexity refers to the curvature of the price-yield curve. A noncallable bond

exhibits varying degrees of positive convexity. Positive convexity means a bond's price rises more for a given yield decline than it falls for the same yield. By unbundling the noncallable bond components in ABS/MBS to their zero-coupon bond components, the rolling yield becomes dominant. Hence, it is called the zero volatility component—that is, the component of the yield spread that is attributable to no change in the expected volatility.

The call option component in ABS/MBS valuation consists of intrinsic and time values. To the extent the option embedded in ABS/MBS is the delayed American exercise style—in other words, the option is not exercised immediately but becomes exercisable any time afterward—the time value component dominates. Thus, in valuing ABS/MBS, the time value of the option associated with the prepayment volatility needs to be evaluated. To evaluate this option, OAS analysis uses an option-based technique to evaluate ABS/MBS prices under different interest rate scenarios. OAS is the spread differential between the zero volatility and option value components of MBS. These values are expressed as spreads measured in basis points.

The option component is the premium paid (earned) from going long (shorting) a prepayment option embedded in the bond. The bondholders are short the option, and they earn the premium in the form of an enhanced coupon. Mortgage holders are long the prepayment option, and they pay the premium in spread above the comparable Treasury. The option component is the cost associated with the variability in cash flow that results from prepayments over time.

The two main inputs into the determination of an OAS of a bond are as follows:

- Generate the cash flow as a function of the principal (scheduled and unscheduled) and coupon payments.
- Generate interest rate paths under an assumed term structure model.

At each cash flow date, a spot rate determines the discount factor for each cash flow. The present value of the cash flow is equal to the sum of the product of the cash flow and the discount factors.¹ When dealing with a case in which uncertainty about future prospects is important, the cash flow and the spot rate need to be specified to account for the uncertainty.² The cash flow and spot rate become a function of time and the state of the economy. The time consideration is that a dollar received now is worth more than one received tomorrow. The state of the economy consideration accounts for the fact that a dollar received in a good economy may be perceived as worth less than a dollar earned in a bad economy. For OAS analysis, the cash flow is run through different economic environments represented by interest rates and prepayment scenarios. The spot rate, which is used to discount the cash flow, is run through time steps and interest rate scenarios. The spot rate represents the instantaneous rate of risk-free return at any time, so that \$1 invested now will have grown by a later time to \$1 multiplied by a continuously compounded rollover rate during the time period.³ Arbitrage pricing theory stipulates the price one should pay now to receive \$1 at later time is the expected discount of the payoff.⁴ So by appealing to the arbitrage pricing theory, we are prompted to introduce an integral representation for the value equation; in other words, the arbitrage pricing theory allows us to use the value additivity principle across all interest rate scenarios.

IS IT EQUILIBRIUM OR AN ARBITRAGE MODEL?

Market participants are guided in their investment decision making by received economic philosophy or intuition. Investors, in general, look at value from either an absolute or relative value basis. Absolute value basis proceeds from the economic notion that the market clears at an exogenously determined price

that equates supply-and-demand forces. Absolute valuation models are usually supported by general or partial *equilibrium* arguments. In implementing market measure models that depend on equilibrium analysis, the role of an investor's preference for risky prospects is directly introduced. The formidable task encountered with respect to preference modeling and the related aggregation problem has rendered these types of models useless for most practical considerations. One main exception is the present value rule that explicitly assumes investors have a time preference for today's dollar. Where the present value function is a monotonically decreasing function of time, today's dollar is worth more than a dollar earned tomorrow. Earlier term structure models were supported by equilibrium arguments, for example, the Cox-Ingersoll-Ross (CIR) model.⁵ In particular, CIR provides an equilibrium foundation for a class of yield curves by specifying the endowments and preferences of traders, which, through the clearing of competitive markets, generates the proposed term structure model.

Relative valuation models rely on *arbitrage* and dominance principles and characterize asset prices in terms of other asset prices. A well-known example of this class is the Black-Scholes⁶ and Merton⁷ option pricing model. Modern term structure models, for example, Hull-White,⁸ Black-Derman-Toy (BDT),⁹ and Heath-Jarrow-Morton (HJM),¹⁰ are based on arbitrage arguments. Although relative valuation models based on arbitrage principles do not directly make assumptions about investors' preferences, there remains a vestige of the continuity of preference, for example, the notion that investors prefer more wealth to less. Thus, whereas modelers are quick in attributing "arbitrage-freeness" to their models, assuming there are no arbitrage opportunities implies a continuity of preference that can be supported in equilibrium. So, if there are no arbitrage opportunities, the model is in equilibrium for some specification of endowments and preferences. The upshot is that the distinc-

tion between equilibrium models and arbitrage models is a stylized fetish among analysts to demarcate models that explicitly specify endowment and preference sets (equilibrium) and those models that are outwardly silent about the preference set (arbitrage). Moreover, analysts usually distinguish equilibrium models as those that use today's term structure as an output and no-arbitrage models as those that use today's term structure as an input.

Arbitrage opportunity exists in a market model if there is a strategy that guarantees a positive payoff in some state of the world with no possibility of negative payoff and no initial net investment. The presence of arbitrage opportunity is inconsistent with economic equilibrium populated by market participants that have increasing and continuous preferences. Moreover, the presence of arbitrage opportunity is inconsistent with the existence of an optimal portfolio strategy for market participants with nonsatiated preferences (prefer more to less) because there would be no limit to the scale at which they want to hold an arbitrage position. The economic hypothesis that maintains two perfect substitutes (two bonds with the same credit quality and structural characteristics issued by the same firm) must trade at the same price is an implication of no arbitrage. This idea is commonly referred to as the law of one price. Technically speaking, the fundamental theorem of asset pricing is a collection of canonical equivalent statements that implies the absence of arbitrage in a market model. The theorem provides for weak equivalence between the absence of arbitrage, the existence of a linear pricing rule, and the existence of optimal demand from some market participants who prefer more to less. The direct consequence of these canonical statements is the pricing rule: the existence of a positive linear pricing rule, the existence of positive risk-neutral probabilities, and associated riskless rate or the existence of a positive state price density.

In essence, the pricing rule representation provides a way of correctly valuing a

security when the arbitrage opportunity is eliminated. A fair price for a security is the arbitrage-free price. The arbitrage-free price is used as a benchmark in relative value analysis to the extent that it is compared with the price observed in actual trading. A significant difference between the observed and arbitrage-free values may indicate the following profit opportunities:

- If the arbitrage price is above the observed price, all else being equal, the security is cheap and a long position may be called for.
- If the arbitrage price is below the observed price, all else being equal, the security is rich and a short position may be called for.

In practice, the basic steps in determining the arbitrage-free value of the security are as follows:

- Specify a model for the evolution of the underlying security price.
- Obtain a risk-neutral probability.
- Calculate the expected value at expiration using the risk-neutral probability.
- Discount this expectation using the risk-free rates.

In studying the solution to the security valuation problem in the arbitrage pricing framework, analysts usually use one of the following:

- Partial differential equation (PDE) framework
- Equivalent martingale measure framework

The PDE framework is a direct approach and involves constructing a risk-free portfolio, then deriving a PDE implied by the lack of arbitrage opportunity. The PDE is solved analytically or evaluated numerically.¹¹

Although there are few analytical solutions for pricing PDEs, most of them are evaluated using numerical methods such as lattice, finite difference, and Monte Carlo. The equivalent martingale measure framework uses the notion of arbitrage to determine a probability measure under which security prices are martingales once discounted. The new probability

measure is used to calculate the expected value of the security at expiration and discounting with the risk-free rate.

WHICH IS THE RIGHT MODEL OF THE INTEREST RATE PROCESS?

The bare essential of the bond market is a collection of zero-coupon bonds for each date, for example, now, that mature later. A zero-coupon bond with a given maturity date is a contract that guarantees the investor \$1 to be paid at maturity. The price of a zero-coupon bond at time t with a maturity date of T is denoted by $P(t, T)$. In general, analysts make the following simplifying assumptions about the bond market:

- There exists a frictionless and competitive market for a zero-coupon bond for every maturity date. By a frictionless market, we mean there is no transaction cost in buying and selling securities and there is no restriction on trades such as a short sale.
- For every fixed date, the price of a zero-coupon bond, $\{P(t, T); 0 \leq t \leq T\}$, is a stochastic process with $P(t, t) = 1$ for all t . By stochastic process, we mean the price of a zero-coupon bond moves in an unpredictable fashion from the date it was bought until it matures. The present value of a zero-coupon bond when it was bought is known for certain and it is normalized to equal one.
- For every fixed date, the price for a zero-coupon bond is continuous in that at every trading date the market is well bid for the zero-coupon bond.

In addition to zero-coupon bonds, the bond market has a money market (bank account) initialized with a unit of money.¹² The bank account serves as an accumulator factor for rolling over the bond.

A term structure model establishes a mathematical relationship that determines the price of a zero-coupon bond, $\{P(t, T); 0 \leq t \leq T\}$, for all dates t between the time the bond is bought (time 0) and when it matures (time T). Alternatively, the term structure shows the relationship between the yield to maturity and the time to maturity of the bond. To compute the value of a security dependent on the term structure, one needs to specify the dynamic of the interest rate process and apply an arbitrage restriction. A term structure model satisfies the arbitrage restriction if there is no opportunity to invest risk-free and be guaranteed a positive return.¹³

To specify the dynamic of the interest rate process, analysts have always considered a dynamic that is mathematically tractable and anchored in sound economic reasoning. The basic tenet is that the dynamic of interest rates is governed by time and the uncertain state of the world. Modeling time and uncertainty are the hallmarks of modern financial theory. The uncertainty problem has been modeled with the aid of the probabilistic theory of the stochastic process. The stochastic process models the occurrence of random phenomena; in other words, the process is used to describe unpredictable movements. The stochastic process is a collection of random variables that take values in the state space. The basic elements distinguishing a stochastic process are state space¹⁴ and index parameter,¹⁵ and the dependent relationship among the random variables (e.g., X_t).¹⁶ The Poisson process and Brownian motion are two fundamental examples of continuous time stochastic processes.

In everyday financial market experiences, one may observe, at a given instant, three possible states of the world: Prices may go up a tick, decrease a tick, or do not change. The ordinary market condition characterizes most trading days; however, security prices may from time to time exhibit extreme behavior. In financial modeling, there is the need to distinguish between rare and normal events. Rare events

usually bring about discontinuity in prices. The Poisson process is used to model jumps caused by rare events and is a discontinuous process. Brownian motion is used to model ordinary market events for which extremes occur only infrequently according to the probabilities in the tail areas of normal distribution.¹⁷

Brownian motion is a continuous martingale. Martingale theory describes the trend of an observed time series. A stochastic process behaves like a martingale if its trajectories display no discernible trends.

- A stochastic process that, on average, increases is called a submartingale.
- A stochastic process that, on average, declines is called a supermartingale.

Suppose one has an interest in generating a forecast of a process (e.g., R_t – interest rate) by expressing the forecast based on what has been observed about R based on the information available (e.g., F_t) at time t .¹⁸ This type of forecast, which is based on conditioning on information observed up to a time, has a role in financial modeling. This role is encapsulated in a martingale property.¹⁹ A martingale is a process, the expectation for which future values conditional on current information are equal to the value of the process at present. A martingale embodies the notion of a fair gamble: The expected gain from participating in a family of fair gambles is always zero and, thus, the accumulated wealth does not change in expectation over time. Note the actual price of a zero-coupon bond does not move like a martingale. Asset prices move more like sub-martingales or supermartingales. The usefulness of martingales in financial modeling stems from the fact that one can find a probability measure that is absolutely continuous with objective probability such that bond prices discounted by a risk-free rate become martingales. The probability measures that convert discounted asset prices into martingales are called equivalent martingale measures. The basic idea is that, in the absence of an arbitrage opportunity, one can

find a synthetic probability measure Q absolutely continuous with respect to the original measure P so that all properly discounted asset prices behave as martingales. A fundamental theorem that allows one to transform R_t into a martingale by switching the probability measure from P to Q is called the Girsanov theorem.

The powerful assertion of the Girsanov theorem provides the ammunition for solving a stochastic differential equation driven by Brownian motion in the following sense: By changing the underlying probability measure, the process that was driving the Brownian motion becomes, under the equivalent measure, the solution to the differential equation. In financial modeling, the analog to this technical result says that in a risk-neutral economy assets should earn a risk-free rate. In particular, in the option valuation, assuming the existence of a risk-neutral probability measure allows one to dispense with the drift term, which makes the diffusion term (volatility) the dominant value driver.

To model the dynamic of interest rates, it is generally assumed the change in rates over instantaneous time is the sum of the drift and diffusion terms (see Figure 1).²⁰ The drift term could be seen as the average movement of the process over the next instants of time, and the diffusion is the amplitude (width) of the movement. If the first two moments are sufficient to describe the distribution of the asset

return, the drift term accounts for the mean rate of return and the diffusion accounts for the standard deviation (volatility). Empirical evidence has suggested that interest rates tend to move back to some long-term average, a phenomenon known as mean reverting that corresponds to the Ornstein-Uhlenbeck process (see Figure 2).²¹ When rates are high, mean reversion tends to cause interest rates to have a negative drift; when rates are low, mean reversion tends to cause interest rates to have a positive drift.

The highlights of the preceding discussion are as follows:

- The modeler begins by decomposing bonds to their bare essentials, which are zero-coupon bonds.
- To model a bond market that consists of zero-coupon bonds, the modeler makes some simplifying assumptions about the structure of the market and the price behaviors.
- A *term structure model* establishes a mathematical relationship that determines the price of a zero-coupon bond and, to compute the value of a security dependent on the term structure, the modeler needs to specify the dynamic of the interest rate process and apply arbitrage restriction.
- The *stochastic process* is used to describe the time and uncertainty components of the price of zero-coupon bonds.

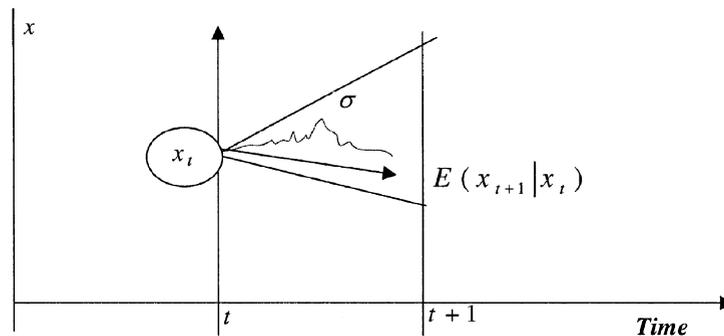


Figure 1 Drift and Diffusion

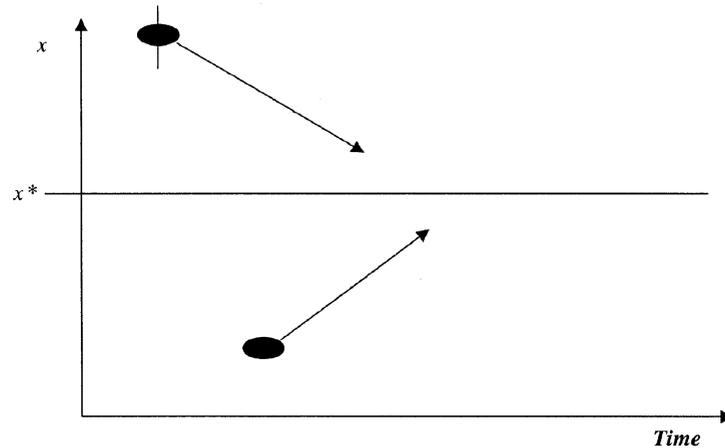


Figure 2 Process with Mean Reversion (Ornstein-Uhlenbeck Process)

- There are two basic types of stochastic processes used in financial modeling: The Poisson process is used to model jumps caused by rare events, and Brownian motion is used to model ordinary market events for which extremes occur only infrequently.
- We assume the market for zero-coupon bonds is well bid, that is, the zero-coupon price is continuous. Brownian motion is the suitable stochastic process to describe the evolution of interest rates over time. In particular, Brownian motion is a continuous martingale. Martingale theory describes the trend of the observed time series.
- Once we specify the evolution of interest rate movements, we need an arbitrage pricing theory that tells us the price one should pay now to receive \$1 later is an expected discounted payoff. The issue to be resolved is, What are the correct expected discount factors to use? The discount must be determined by the market and based on risk-adjusted probabilities. In particular, when all bonds are properly risk-adjusted, they should earn risk-free rates; if not, arbitrage opportunity exists to earn riskless profit.
- The risk-adjusted probability consistent with the no-arbitrage condition is the equivalent martingale measure; it is the probability mea-

sure that converts the discounted bond price to a martingale (fair price). The elegance of the martingale theory is the “roughs and tumbles” one finds in the world of partial differentiation are to some extent avoided and the integral representation it allows fits nicely with Monte Carlo simulations.

Several term structure models have been proposed with subtle differences. However, the basic differences amount to how the dynamic of the interest rate is specified, the number of factors that generate the rate process, and whether the model is closed by equilibrium or arbitrage arguments.

Which of these models to use in OAS analysis depends on the available resources. Where resource availability is not an issue, we favor models that account for the path-dependent nature of mortgage cash flows. Good rules-of-thumb in deciding which model to use are as follows:

- *Flexibility*: How flexible is the model?
- *Simplicity*: Is the model easy to understand?
- *Specification*: Is the specification of the interest rate process reasonable?
- *Realism*: How real is the model?
- *Good fit*: How well does the result fit the market data?

- *Internal consistency rule*: A necessary condition for the existence of market equilibrium is the absence of arbitrage, and the external consistency rule requires models to be calibrated to market data.

TERM STRUCTURE MODELS: WHICH IS THE RIGHT APPROACH FOR OAS?

Numerical schemes are constructive or algorithmic methods for obtaining practical solutions to mathematical problems. They provide methods for effectively finding practical solutions to asset pricing PDEs.

The first issue in a numerical approach is discretization. The main objective for discretizing a problem is to reduce it from continuous parameters formulation to an equivalent discrete parameterization in a way that makes it amenable to practical solution. In financial valuation, one generally speaks of a continuous time process in an attempt to find an analytical solution to a problem; however, nearly all the practical solutions are garnered by discretizing space and time. Discretization involves finding numerical approximations to the solution at some given points rather than on a continuous domain.

Numerical approximation may involve the use of a pattern, lattice, network, or mesh of discrete points in place of the (continuous) whole domain, so that only approximate solutions are obtained for the domain in the isolated points, and other values such as integrals and derivatives can be obtained from the discrete solution by the means of interpolation and extrapolation.

With the discretization of the continuous domain come the issues of adequacy, accuracy, convergence, and stability. Perhaps how these issues are faithfully addressed in the implementation of OAS models speaks directly to the type of results achieved. Although these numerical

techniques—lattice methods, finite difference methods, and Monte Carlo methods—have been used to solve asset pricing PDEs, the lattice and Monte Carlo methods are more in vogue in OAS implementations.

Lattice Method

The most popular numerical scheme used by financial modelers is the *lattice* (or *tree*) *method*. A lattice is a nonempty collection of vertices and edges that represent some prescribed mathematical structures or properties. The node (vertex) of the lattice carries particular information about the evolution of a process that generates the lattice up to that point. An edge connects the vertices of a lattice. A lattice is initialized at its root, and the root is the primal node that records the beginning history of the process.

The lattice model works in a discrete framework and calculates expected values on a discrete space of paths. A node in a given path of a nonrecombining lattice distinguishes not only the value of the underlying claim there but also the history of the path up to the node. A bushy tree represents every path in the state space and can numerically value path-dependent claims. A node in a given path of a bushy tree distinguishes not only the value of the underlying claim there but also the history of the path to the node. There is a great cost in constructing a bushy tree model. For example, modeling a 10-year Treasury rate in a binary bushy tree with each time period equal to one coupon payment would require a tree with 2^{20} (1,048,576) paths. Figure 3 shows a schematic of a bushy tree.

In a lattice construction, it is usually assumed the time to maturity of the security, T , can be divided into discrete (finite and equal) time-steps M , $\Delta t = T/M$. The price of the underlying security is assumed to have a finite number of “jumps” (or up-and-down movements) N between the time-steps Δt . In a recombining lattice, the price or yield of the underlying security is assumed to be affected by N and not

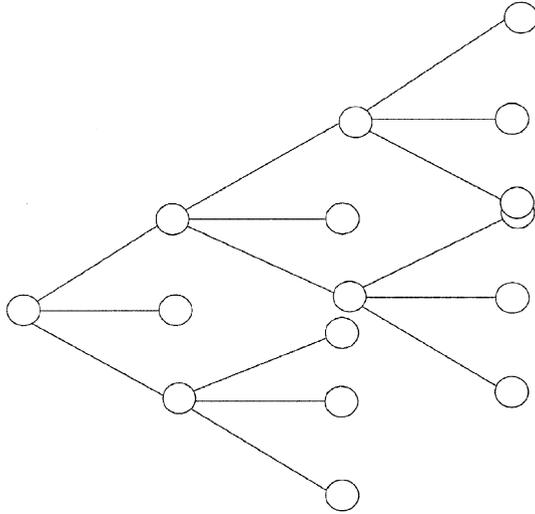


Figure 3 Bushy or Nonrecombining Tree

the sequences of the jumps. For computational ease, N is usually set to be two or three; the case where $N = 2$ is called binomial lattice (or tree), and $N = 3$ is the trinomial lattice. Figures 4 and 5 show the binomial and trinomial lattices, respectively, for the price of a zero-coupon bond.

Monte Carlo Method

The *Monte Carlo method* is a numerical scheme for solving mathematical models that involve random sampling. This scheme has been used to solve problems that are either deterministic or probabilistic in nature. In the most common

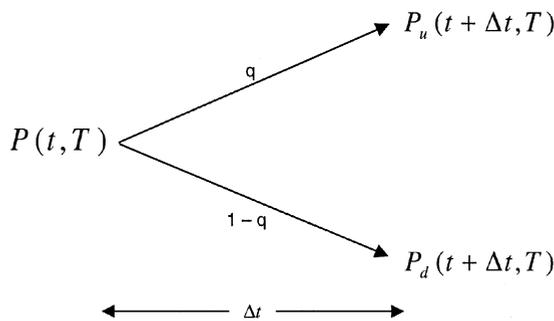


Figure 4 Binomial Lattice for the Price of a Zero-Coupon Bond

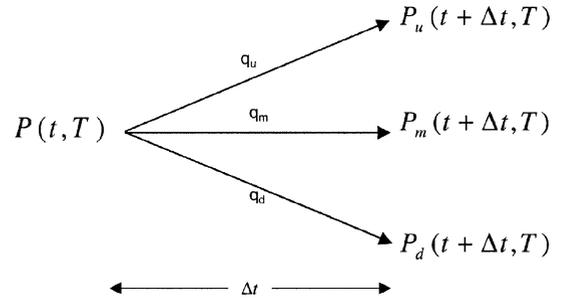


Figure 5 Trinomial Lattice for the Price of a Zero-Coupon Bond

application, the Monte Carlo method uses random or pseudo-random numbers to simulate random variables. Although the Monte Carlo method provides flexibilities in dealing with a probabilistic problem, it is not precise especially when one desires the highest level of accuracy at a reasonable cost and time.

Aside from this drawback, the Monte Carlo method has been shown to offer the following advantages:

- It is useful in dealing with multidimensional problems and boundary value problems with complicated boundaries.
- Problems with random coefficients, random boundary values, and stochastic parameters can be solved.
- Solving problems with discontinuous boundary functions, nonsmooth boundaries, and complicated right-hand sides of equations can be achieved.

The application of the Monte Carlo method in computational finance is predicated on the integral representation of security prices. The approach taken consists of the following:

- Simulating in a manner consistent with a risk-neutral probability (equivalent martingale) measure the sample path of the underlying state variables
- Evaluating the discounted payoff of the security on each sample path
- Taking the expected value of the discounted payoff over the entire sample paths

The Monte Carlo method computes a multidimensional integral—the expected value of discounted cash flows over the space of sample paths. For example, let $f(x)$ be an integral function over d -dimensional unit hypercube, then a simple (or crude) estimate of the integral is equal to the average value of the function f over n points selected at random (more appropriately, pseudorandom) from the unit hypercube. By the law of large numbers,²² the Monte Carlo estimate converges to the value as n tends to infinity. Moreover, we know from the central limit theorem that the standard error of estimate tends toward zero as $1/(\sqrt{n})$. To improve on the computational efficiency of the crude Monte Carlo method, there are several variance-reduction techniques available.

IS THERE A RIGHT WAY TO MODEL PREPAYMENTS?

Because cash flows are one of the most important inputs in determining the value of a security, there has to be a model for cash flow. The cash flow model consists of a model for distributing the coupon and scheduled principal payments to the bondholders, as contained in the deal prospectus, and a *prepayment model* that projects unscheduled principal payments. The basic types of prepayment models are as follows:

- *Rational prepayment models.* These models apply an option-theoretic approach and link prepayment and valuation in a single unified framework.
- *Econometric prepayment models.* This class of models is based on econometric and statistical analysis.
- *Reduced-form prepayment models.* This type of model uses past prepayment rates and other endogenous variables to explain current prepayment. It fits the observed prepayment data, unrestricted by theoretical consideration.

The reduced-form prepayment model is the most widely used approach among dealers and prepayment vendors because of its flexibility and unrestricted calibration techniques. The basic determinants of the voluntary and involuntary components of total prepayments are collateral and market factors. Collateral factors are the origination date, weighted average coupon (WAC), and weighted average maturity, and the market-related factors are benchmark rates and spreads.

KEY POINTS

- There are foundational issues that explain (1) why there is a difference in dealers' OAS values for a specific bond, (2) what may be responsible for the differences, and (3) why one OAS value may be more correct than another.
- As a general guideline, portfolio managers should become familiar with the economic intuitions and basic assumptions made by the models.
- The reasonableness of the OAS values produced by different models should be considered. Moreover, because prepayment options are not traded in the market, calibrating OAS values using the prices of these options is not possible.
- Interest rate models, which are closed by precluding arbitrage opportunities, are more tractable and realistic.
- Interest rate models that account for the path-dependent natures of ABS and MBS cash flows are more robust.
- With the path-dependent natures of ABS and MBS cash flows come the difficulties of implementation, in particular, the speed of calculation; the toss-up here is between the lattice and Monte Carlo schemes.
- There is a tendency for market participants to believe that because we are talking about interest rate scenarios, the ideal candidate for the job would be Monte Carlo techniques, but this should not necessarily be the case. Although lattice implementation could do a

good job, the success of this scheme depends highly on ad hoc techniques that have not been time-tested. Hence, whereas the OAS implementation scheme is at the crux of what distinguishes good or bad results, the preferred scheme is an open question that critically depends on available resources.

- Reduced-form prepayment models should be favored because of their flexibility and unrestricted calibration techniques. In particular, a model that explicitly identifies its control parameters and is amenable to the perturbation of these parameters is more robust and transparent.
- With respect to how to interpret the differences in dealers' OAS value for a specific security, decisions by dealers, vendors, and portfolio managers to choose one interest rate and prepayment model over others and the different approaches they take in implementing these models largely account for the wide variance in OAS results. Moreover, to complicate the issue, the lack of a market for tradable prepayment options makes calibrating the resulting OAS values dicey at best.
- As for whether there is a correct OAS value for a given security, examining the change in OAS value over time, the sensitivity of OAS parameters, and their implications to relative value analysis are some of the important indicators of the reasonableness of the OAS value.

NOTES

1. In the world of certainty, the present value is

$$PV = \sum_{i=1}^n \frac{cf_i}{(1+r_i)^i}$$

where r_i is the spot rate applicable to cash flow cf_i . In terms of forward rates, the equation becomes

$$PV = \sum_{i=1}^n \frac{cf_i}{(1+f_1)(1+f_2)\dots(1+f_n)}$$

where f_i is the forward rate applicable to cash flow cf_i .

2. The present value formula becomes more complicated and could be represented as

$$PV_{\Omega} = \sum_{\omega_i} \sum_{t_i}^T \frac{cf(t_i, \omega_i)}{(1+r(t_i, \omega_i))}$$

$$\forall i = 1, 2, \dots, N$$

where

PV_{Ω} = the present value of uncertain cash flow

$cf(t_i, \omega_i)$ = the cash flow received at time t_i and state ω_i

$r(t_i, \omega_i)$ = the spot rate applicable at time t_i and state ω_i

For OAS analysis, a stylized version of the previous equation is given by

$$PV_{\Omega} = \lim_{n \rightarrow \infty} \frac{1}{N} \frac{cf(t_i, \omega_i)}{(1+r(t_i, \omega_i))}$$

$$\forall i = 1, 2, \dots, N$$

3. $\$1 \left[\exp \left(\int_t^T r(u) du \right) \right]$
4. $p(t, T) = E \left[\exp \left(- \int_t^T r(u, du) \middle| F_t \right) \right]$
5. Cox, Ingersoll, and Ross (1985).
6. Black and Scholes (1973).
7. Merton (1974).
8. Hull and White (1990).
9. Black, Derman, and Toy (1990).
10. Heath, Jarrow, and Morton (1992).
11. For example, the PDE for a zero-coupon bond price is

$$\frac{\partial p}{\partial t} + \frac{1}{2} \sigma^2 \frac{\partial^2 p}{\partial r^2} + (\mu - \lambda \sigma) \frac{\partial p}{\partial r} - rp = 0$$

where

p = zero-coupon price

r = instantaneous risk-free rate

μ = the drift rate

σ = volatility

λ = market price of risk

To solve the zero-coupon price PDE, we must state the final and boundary

conditions. The final condition that corresponds to payoff at maturity is $p(r, T) = k$.

12. The bank account is denoted by

$$B(t) = \exp \left[\int_0^t r(u) du \right]$$

and $B(0) = 1$.

13. Technically, the term structure model is said to be arbitrage-free if and only if there is a probability measure \mathbf{Q} on Ω ($\mathbf{Q} \sim \mathbf{P}$) with the same null

$$Z(t, T) = \frac{P(t, T)}{B(t)}, 0 \leq t \leq T$$

set as \mathbf{P} , such that for each t , the process is a martingale under \mathbf{Q} .

14. State space is the space in which the possible values of X_t lie. Let S be the state space. If $S = (0, 1, 2, \dots)$, the process is called the discrete state process. If $S = \mathfrak{R}(-\infty, \infty)$ that is the real line, and the process is called the real-valued stochastic process. If S is Euclidean d -space, then the process is called the d -dimensional process.
15. Index parameter: If $T = (0, 1, \dots)$, then X_t is called the discrete-time stochastic process. If $T = \mathfrak{R}_+[0, \infty)$, then X_t is called a continuous time stochastic process.
16. Formally, a stochastic process is a family of random variables $X = \{x_t; t \in T\}$, where T is an ordered subset of the positive real line \mathfrak{R}_+ . A stochastic process X with a time set $[0, T]$ can be viewed as a mapping from $\Omega \times [0, T]$ to \mathfrak{R} with $x(\omega, t)$ denoting the value of the process at time t and state ω . For each $\omega \in \Omega$, $\{x(\omega, t); t \in [0, T]\}$ is a sample path of X sometimes denoted as $x(\omega, \bullet)$. A stochastic process $X = \{x_t; t \in [0, T]\}$ is said to be adapted to filtration F if x_t is measurable with respect to F_t for all $t \in [0, T]$. The adaptedness of a process is an informational constraint: The value of the process at any time t cannot depend on the information yet to be revealed strictly after t .
17. A process X is said to have an independent increment if the random variables $x(t_1) -$

$x(t_0), x(t_2) - x(t_1) \dots$ and $x(t_n) - x(t_{n-1})$ are independent for any $n \geq 1$ and $0 \leq t_0 < t_1 < \dots < t_n \leq T$. A process X is said to have a stationary independent increment if, moreover, the distribution of $x(t) - x(s)$ depends only on $t - s$. We write $z \sim N(\mu, \sigma^2)$ to mean the random variable z has normal distribution with mean μ and variance σ^2 . A standard Brownian motion W is a process having continuous sample paths, stationary independent increments, and $W(t) \sim N(\mu, t)$ (under probability measure P). Note that if X is a continuous process with stationary and independent increments, then X is a Brownian motion. A strong Markov property is a memoryless property of a Brownian motion. Given X as a Markov process, the past and future are statistically independent when the present is known.

18. We write

$$E_t[R_t] = E[R_T | F_t], t < T$$

19. More concretely, given a probability space, a process $\{R_t; t \in (0, \infty)\}$ is a martingale with respect to information sets F_t , if for all $t > 0$,
1. R_t is known, given F_t , that is, R_t is F_t adapted
 2. Unconditional forecast is finite; $E|R_t| < \infty$
 3. And if

$$E_t[R_t] = R_T, \quad \forall t < T$$

with a probability of 1. The best forecast of unobserved future value is the last observation on R_t .

20. In particular, assume

$$dX(t) = \alpha(t, X(t))dt + \beta(t, X(t))dW(t)$$

for which the solution $X(t)$ is the factor. Depending on the application, one can have n -factors, in which case we let X be an n -dimensional process and W an n -dimensional Brownian motion. Assume the stochastic differential equation for $X(t)$ describes the interest process $r(t)$, (i.e., $r(t)$

is a function of $X(t)$. A one-factor model of interest rate is

$$dr(t) = \alpha(t)dt + \beta(t)dW(t)$$

21. This process is represented as

$$dr = a(b - r)dt + \sigma r^\beta dW$$

where a and b are called the reversion speed and level, respectively.

22. *Strong Law of Large Numbers*. Let $X = X_1, X_2 \dots$ be an independent identically distributed random variable with $E(X^2) < \infty$ then the mean of the sequence up to the n th term, though itself a random variable, tends as n get larger and larger, to the expectation of X with probability 1. That is

$$P\left(\lim_{n \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n X_i\right) = E(X)\right) = 1$$

REFERENCES

- Black, F., Derman, E., and Toy, W. (1990). A one-factor model of interest rates and its application to Treasury bond options. *Financial Analysts Journal* 46: 33–39.
- Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.
- Cox, J., Ingersoll, J., and Ross, S. (1985). A theory of the term structure of interest rates. *Econometrica* 53: 385–408.
- Heath, D., Jarrow, R., and Morton, A. (1992). Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation. *Econometrica* 60: 77–105.
- Hull, J., and White, A. (1990). Pricing interest rate derivatives securities. *Review of Financial Studies* 3: 573–592.
- Merton, R. (1974). The theory of rational option pricing. *Bell Journal of Economics and Management Science* 4: 141–183.

Quantitative Models to Value Convertible Bonds

FILIPPO STEFANINI

Head of Hedge Funds and Manager Selection, Eurizon Capital SGR

Abstract: Convertible bonds are bonds that give their holders the right to periodic coupon payments and, as of a fixed date, the right to convert the bonds into a fixed number of shares. If the bondholder decides to exercise his conversion right, instead of being paid back the par value of the bonds, he will receive a fixed number of shares in exchange. There are several options embedded in a convertible bond. There is obviously a call option on the underlying stock. All convertible bonds are callable. A convertible bond may be puttable. The presence of all of these options complicates the valuation of convertible bonds. There are models that practitioners use for valuation purposes. These models are classified as analytical models and numerical models.

Convertibles are ideal securities for arbitrage, because the convertible itself, namely the underlying stock and the associated derivatives, are traded along predictable ratios, and any discrepancy or misprice would give rise to arbitrage opportunities for fund managers. Traders use quantitative models to identify convertible bonds whose market value differs from their theoretical price. However, unlike callable bonds or puttable bonds that have interest rate-embedded options, a convertible bond also has an embedded equity option. This complicates the quantitative modeling of these securities.

Quantitative models, or valuation models, for convertible bonds are divided into two categories: *analytical models* and *numerical models*. In this entry, we describe the more commonly used model in both of these categories.

ANALYTICAL MODELS

Ingersoll (1977) proposed a valuation model for convertible bonds based on the option theory and on the Black-Scholes option pricing model. The model's main assumptions are:

- Markets operate continuously.
- There are no transaction costs.
- Share prices follow an Ito diffusion process.
- Securities prices have a lognormal distribution.
- The underlying stock volatility is constant.

Ingersoll's model assumes that prices vary continuously, that is, there is always liquidity in the market and there are no limits to securities lending and short selling. It also assumes that the company's market value follows an Ito diffusion process, that is, a continuous Brownian

motion. Under this assumption, it is possible to set up a closed analytical formula to calculate the value of a convertible bond.

The model can be applied only to European convertibles, namely, convertibles that can be exercised only upon expiration. Moreover, the model makes it clear how complex the valuation of convertible bonds is, and it provides a highly interesting theoretical reference, in that it can reach an analytical solution to the valuation of convertibles. Yet, we know all too well that interest rates, credit spreads, currencies, and dividends are not constant, and the clauses and provisions written in the prospectus of a convertible are often highly varied and complicated, making it fairly difficult to apply analytical valuation models. This is why it is necessary to turn to numerical approximation models.

The Ingersoll Model

As just noted, the *Ingersoll model* provides an analytic solution for the pricing of a convertible bond, given some general market assumptions. The strongest assumptions are:

- Capital markets are perfect with no transaction costs, no taxes, and equal access to information for all investors.
- Trading takes place continuously in time and there are no restrictions against borrowing or short sales.
- The market value of the company follows an Ito diffusion process.

The Black-Scholes option pricing model is used to value the convertible bond as a contingent claim on the firm as a whole.

Consider a convertible bond that is convertible only at maturity, therefore with a European call option embedded. Let

$$\gamma = \frac{n}{n + N}$$

equal the dilution factor, indicating the fraction of the common equity that would be held by

the convertible bond issue's owners if the entire issue were converted:

- V = market value of the company
- τ = maturity of the convertible bond
- B = balloon payment (nominal value of the convertible bond)
- r = interest rate

In light of the continuous-time analysis, the functional form to assume for the call price of a convertible bond is the exponential:

$$K(\tau) = B \cdot e^{-\rho\tau}$$

where

- ρ = rate of change in the call price
- σ^2 = the instantaneous variance of returns of the stock underlying the convertible bond

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$

is the cumulative normal distribution

$$F(V, \tau; B, 0) = B \cdot e^{-r\tau} \cdot \left[\Phi \left(\frac{-\log(B \cdot e^{-r\tau}/V) + \frac{1}{2}\sigma^2\tau}{\sigma\sqrt{\tau}} \right) + V \cdot \Phi \left(\frac{-\frac{\log(B \cdot e^{-r\tau}/V) + \frac{1}{2}\sigma^2\tau}{\sigma\sqrt{\tau}}}{\frac{B \cdot e^{-r\tau}}{V}} \right) \right]$$

$$W(\gamma V, \tau; B) = \gamma \cdot V \cdot \Phi \left(\frac{\log(\frac{\gamma V}{B}) + (r + \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}} \right) - B \cdot e^{-r\tau} \cdot \Phi \left(\frac{\log(\frac{\gamma V}{B}) + (r + \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}} - \sigma\sqrt{\tau} \right)$$

The value of the convertible bond is

$$H(V, \tau) = F(V, \tau; B, 0) + W(\gamma V, \tau; B) + \left(\frac{K(\tau)}{\gamma V} \right)^{2(r-\rho)/\sigma^2} \cdot \left(F(\gamma V \cdot e^{(\rho-r)\tau}, \tau; B \cdot e^{(r-\rho)\tau}, 0) - F(\gamma V \cdot e^{(\rho-r)\tau}, \tau; \frac{B}{\gamma} \cdot e^{(r-\rho)\tau}, 0) \right)$$

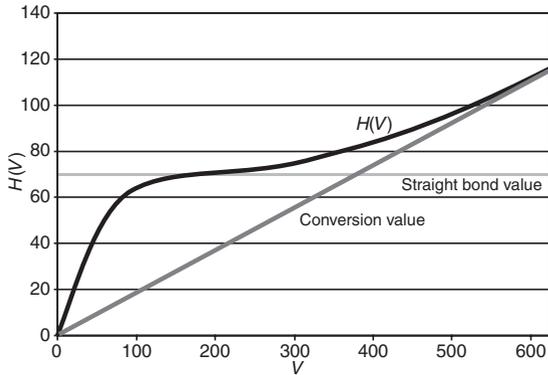


Figure 1 Plot of a Convertible Bond Function for Different Firm Values

To illustrate the model, let's plot the function H with the following parameters:

- $B = 100$
- $\rho = 0,02$
- $\gamma = 0,2$
- $\sigma^2 = 5\%$
- $r = 7\%$

V ranges from 0 to 625.

The plots are shown in Figure 1. The straight lines cross at

$$V(\tau) = \frac{K(\tau)}{\gamma}$$

NUMERICAL MODELS

The most widely used mathematical models among hedge fund managers for the valuation of convertible bonds are numerical methods, among which are the binomial and trinomial trees, the three-dimensional binomial model, implied trees, and the Monte Carlo simulation model.

The *binomial tree model* was introduced by Cox, Ross, and Rubinstein (1979) and by Sharpe in his textbook (Sharpe, 1978). This model allows one to build a tree of possible share prices between now and the convertible's maturity date. This tree is then used to find the convertible's current value by calculating its value along all the tree's nodes. In the binomial tree model,

the tree has two branches that develop from every node, while in the trinomial tree model there are three branches diverging from each node. The higher the number of nodes, the more accurate the model is. The binomial model makes it possible to also value an American option that would otherwise find no solution in a closed form. If the number of time steps grows bigger, the binomial tree tends toward the Black-Scholes continuous formula for European options.

All these models are helpful when making a decision, but many of the options embedded in a convertible do not fit the models and therefore the fund manager's skill and a rigorous risk management discipline become more precious. The manager's art lies in finding innovative ways to evaluate convertible bonds without being swamped with too many details.

The trinomial tree model was introduced by Boyle (1986). The share price can move in three directions from every single node and therefore the number of time steps can be reduced to reach the same precision obtained with the binomial tree.

The *Monte Carlo method*, named after the casino of the Principality of Monaco, is a statistical simulation method, according to which data obtained through the generation of random numbers coming from a given statistical distribution is considered empirical and is used to estimate the parameters under consideration. Thousands of random samples are generated, derived from the assumed statistical distribution, which takes as parameters the maximum likelihood estimators using real data, and then these data are used to estimate the parameters under examination.

The Binomial Tree Model

Here, we will describe a version of the Cox-Ross-Rubinstein model as modified by Goldman Sachs. The binomial tree model can be used to evaluate convertible bonds with either an

embedded European call option or an embedded American call option.

To determine the value of the convertible bond, it's necessary to build four different trees in the following order:

1. Stock price tree.
2. Conversion probability tree.
3. Credit-adjusted spread tree.
4. Convertible bond value tree that is calculated backward from the previous trees.

In the first step we build the stock price tree. The binomial tree model allows us to build up a picture of how a stock is likely to perform between now and the maturity of the convertible bond (T). The number of nodes (N) is calculated from the maturity of the convertible bond according to the formula $T \cdot (T + 1)/2$. The more nodes, the more accurate will be the model.

Between a node and the following node, the stock price can move upward or downward. The jump of the stock price depends on the length of the time interval $\Delta t = T/N$ and on the stock price volatility σ . Therefore

$$\begin{aligned} u &= e^{\sigma\sqrt{\Delta t}} \text{ (upward move)} \\ d &= e^{-\sigma\sqrt{\Delta t}} \text{ (downward move)} \end{aligned}$$

The stock price, S , at each node is set equal to

$$S \cdot u^i \cdot d^{j-i}$$

where $i = 0, 1, \dots, j$

N is the time step and i is the number of upward moves.

The probability of a downward move in stock price at the next time step Δt is

$$p = \frac{e^{b\sqrt{\Delta t}-d}}{u-d}$$

while the probability of a downward move must be $(1 - p)$, since the probability of going either up or down equals unity.

In the second step we build the conversion probability tree. We calculate the conversion probabilities backward, starting from the leaves of the stock price tree. If it's optimal to convert the bond, the conversion probability is 1, other-

wise it is 0. For the steps before the end of the tree, the conversion probability is 1 if it's optimal to convert the bond; otherwise, it is equal to

$$q_{n,i} = p \cdot q_{n+1,i+1} + (1 + p) \cdot q_{n+1,i}$$

In the third step we build the credit-adjusted spread tree. If the convertible bond is out-of-the-money, futures cash flows should be discounted to a rate equal to the risk-free rate, r , plus a credit spread, k , of that particular bond. In fact, if the stock price is much lower than the conversion price, the convertible bond behaves like a plain vanilla bond. If the convertible bond is in-the-money, future cash flows must be discounted at the risk-free rate. In this case, the convertible bond behaves like a stock. Therefore, instead of using a fixed discount rate r , in each node is calculated a discount rate $r_{n,i}$ and a conversion probability $q_{n,i}$ is used. The discount rate is equal to

$$r_{n,i} = q_{n,i} \cdot r + (1 - q_{n,i}) \cdot (r + k)$$

In the fourth step, we build the convertible bond value tree. At each node of the tree, the price of the convertible bond is equal to the maximum between the conversion value of the bond and the face value plus the final coupon. The tree is built backward: from the leaves back to the root of the tree. The root of the tree is the price of the convertible bond.

If it's optimal to convert the bond at a node, then that node is assigned the conversion value; otherwise, the price of the convertible bond is

$$\begin{aligned} P_{n,i} &= \max[mS, p \cdot P_{n+1,i+1} \cdot e^{r_{n+1,i+1} \cdot \Delta t} \\ &\quad + (1 - p) \cdot P_{n+1,i} \cdot e^{-r_{n+1,i} \cdot \Delta t}] \end{aligned}$$

where m is the conversion ratio.

For example, let's determine the price of a convertible bond with the binomial tree method, starting with the following data:

- $T = 5$ years (maturity)
- $\Delta t = 1$ year (step)
- $N = 5$ (number of nodes)
- $r = 4\%$ (risk-free rate)
- $k = 2\%$ (credit spread)

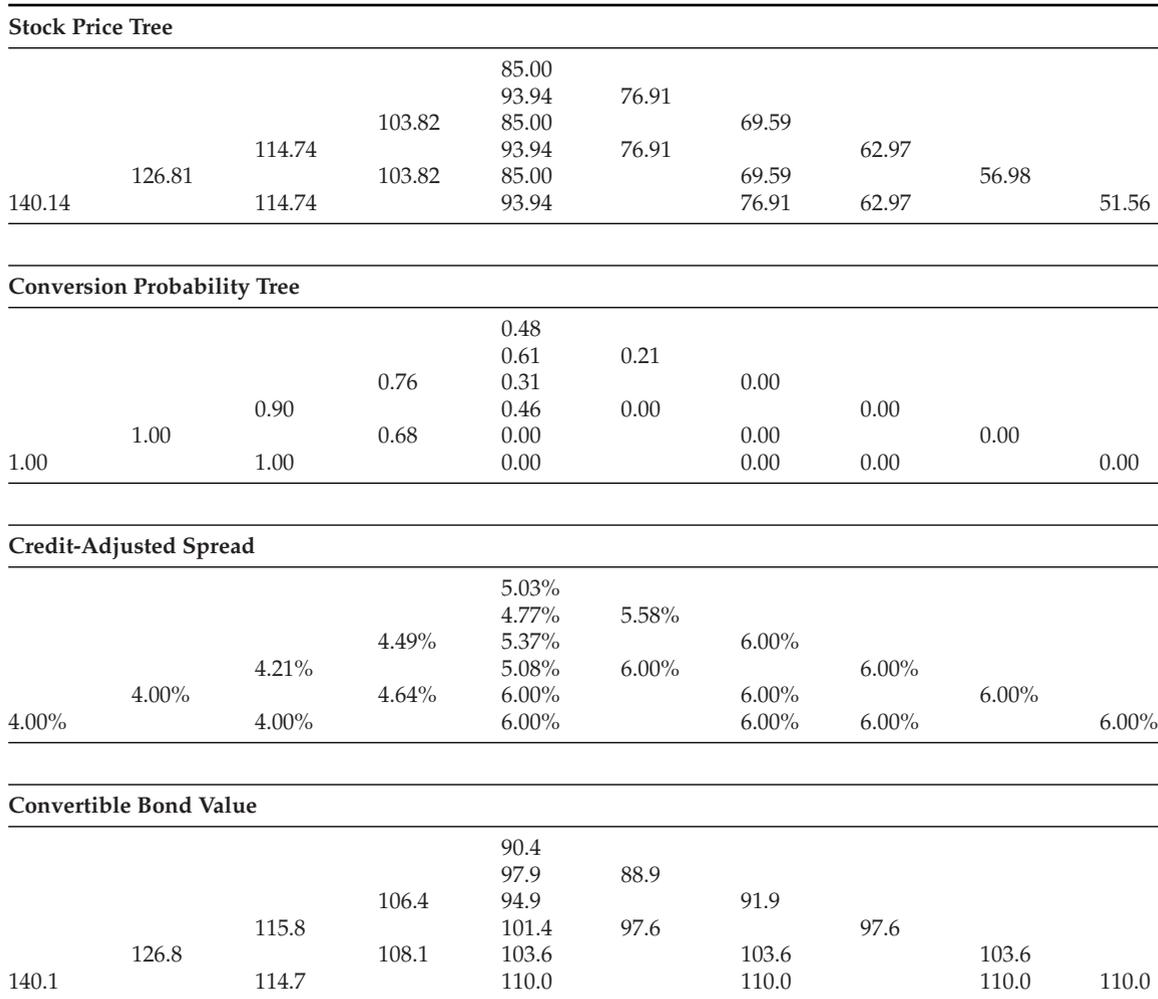


Figure 2 Binomial Trees Necessary to Determine the Value of a Convertible Bond

The convertible bond has nominal value 100 and coupon 10%.

$m = 100\%$ (conversion ratio)

$S = 85$ (stock price)

$\sigma = 10\%$ (stock volatility)

With the formulas discussed above we calculate

$u = 1.1052$ (upwards move)

$d = 0.9048$ (downwards move)

$p = 0.6787$ (probability of an upward move of the stock price in the next time interval Δt)

As shown in Figure 2 we built first the stock price tree, then the conversion probability tree,

then the credit-adjusted spread tree, and finally the convertible bond value tree. The value in the root of the tree is 90.4, which is the price of the convertible bond.

KEY POINTS

- To implement strategies involving convertible bonds, traders and fund managers require a valuation model.
- Analytical models provide a closed-form solution for the value of a convertible bond, and the most commonly used model in practice is the Ingersoll model.

- While there are several models that fall into the realm of numerical models, the one commonly used is the binomial tree model, which requires the construction of a stock price tree, conversion probability tree, credit-adjusted spread tree, and convertible bond value tree that is calculated backward from the previous trees.

REFERENCES

- Boyle, P. P. (1986). Option valuation using a three jump process. *International Options Journal* 3: 7–12.
- Boyle, P. P. (1988). A lattice framework for option pricing with two state variables. *Journal of Financial and Quantitative Analysis* 23, 1: 1–12.
- Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 3: 637–654.
- Cox, J. C., Ross, S. A., and Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of Financial Economics* 7, 3: 229–263.
- Cox, J. C. and Rubinstein, M. (1985). *Option Markets*. Englewood Cliffs, NJ: Prentice Hall.
- Ingersoll, J. (1977). A contingent-claims valuation of convertible securities. *Journal of Financial Economics* 4, 2: 289–322.
- Sharpe, W. F. (1978). *Investments*. Englewood Cliffs, NJ: Prentice Hall.

Quantitative Approaches to Inflation-Indexed Bonds

WESLEY PHOA, PhD

Senior Vice President, Capital International Research, Inc.

Abstract: Inflation-indexed bonds, as a way of financing government debt, were proposed in the 1920s by economists such as Alfred Marshall and John Maynard Keynes. In Israel, they have been issued since the 1950s and have often dominated that country's bond market. Inflation-indexed sovereign bonds now exist in a broad range of developed countries, as well as in a number of emerging markets. A wide variety of bond structures and tax regimes exist. Issuance volumes and the breadth of the investor base vary widely from country to country; liquidity varies from reasonably good to very poor. When inflation-indexed bonds were introduced in the United States in 1997, there was some disagreement about the degree to which inflation-indexed bonds—called Treasury inflation-protected securities or TIPS—are “risk-free” and the role they should play in a portfolio. In particular, it had not been universally appreciated that these bonds can have volatile mark-to-market returns.

Since their introduction in 1997, *Treasury inflation-protected securities (TIPS)* have become an established part of the U.S. bond market. This entry reviews the structure of TIPS and the factors that drive TIPS returns; examines the role that TIPS play in a broader bond portfolio, and the nature of TIPS interest rate risk; and discusses some methods employed by TIPS investors to assess value and risk.

BOND STRUCTURES AND THE CONCEPT OF REAL YIELD

The key features of the TIPS bond structure are summarized here:

- TIPS pay interest semiannually. Interest payments are based on a fixed coupon rate. However, the underlying principal amount of the bonds is indexed to inflation; this inflation-adjusted principal amount is used to calculate the coupon payments, which therefore also rise with inflation. At maturity, the redemption value of the bonds is equal to their inflation-adjusted principal amount, rather than their original par amount.
- The inflation-adjusted principal amount is equal to the original par amount multiplied by an index ratio, which is based on changes in the *Consumer Price Index (CPI)* and which is recalculated every day. The index ratio is simply the reference CPI on the relevant date divided by the reference CPI on the

issue date. Negative inflation adjustments are not made.

- The reference CPI for the first day of any month is defined to be the non-seasonally adjusted CPI-U for the third preceding calendar month, while the reference CPI for any subsequent day in that month is determined by linearly interpolating the reference CPI for the first of the month and the reference CPI for the first day of the next month.
- Price-yield calculations are as follows. Compute the “real price” of the bond from the quoted *real yield* via the standard bond pricing formula, using an actual/actual day count basis, round to 3 decimal places (in \$100); then multiply the real price by the index ratio to obtain the inflation-adjusted price. Accrued interest is computed in exactly the same way, except that no rounding is carried out.

An attractive feature of the TIPS structure is that inflation indexation occurs with no substantial lag. In the U.K., there is an eight-month lag in the inflation adjustment of index-linked gilts; in Australia and New Zealand, there is a three- to six-month lag. The lag means that real returns from these *inflation-indexed bonds* are subject to short-term inflation risk and considerably complicates the analysis of the bonds.

The obvious question, of course, is: Where does the real yield come from, and how much can it change? To investors used to thinking of bond yields as being driven by inflation expectations, it is not obvious that real yields should be volatile at all—except perhaps because of temporary imbalances in supply and demand, or changes in liquidity. After all, there are respectable economic theories that suggest that real interest rates should be constant. But in practice, there are various economic reasons why real yields do in fact fluctuate.¹

Causes of Real Yield Volatility

The real yield may be defined as the long-term cost of risk-free capital (net of inflation). That

is, since TIPS are competing with other investments, real yields on TIPS will move with the cost of capital in the economy as a whole. Of course, other factors affect real yields: For example, index-linked gilts in the U.K. have had artificially low real yields because of their favorable tax treatment and because of a regulatory requirement (since loosened) making it virtually obligatory for pension funds to own them. However, in this entry we will focus on economic and market factors.

Long-term real yields are influenced by expectations about future long-term real interest rates. The two main macroeconomic factors that affect these expectations are:

1. The *domestic* factor: long-term expected growth in real gross domestic product (GDP). Strong growth generally drives up real interest rates, since the demand for capital tends to rise, and borrowers—expecting higher real returns—are prepared to shoulder higher real borrowing costs.
2. The *international* factor: long-term expected changes in the current account deficit. Demand for capital is by definition higher in countries with a large current account deficit, driving up domestic interest rates in order to attract required international investment.

Note that short-term trends in real GDP and the current account deficit can have a strong influence on real yields, because they tend to influence the long-term expectations of investors.² (Roll [1996] has also argued, based on an analysis of tax effects, that real yields should also rise when expected inflation rises; this argument is outlined later in this entry. For the moment we ignore tax effects.)

Real yields on inflation-linked bonds are also influenced by relative demand for these bonds when compared with competing investments that may offer investors some protection—albeit imperfect—against inflation. The balance between competing investments constantly shifts, depending on subjective

factors such as investor aversion to different kinds of risk. Relevant investments include:

1. *Money market investments*: If investors are confident that short-term interest rates will move broadly in line with inflation—which was the case for US monetary policy during the “Great Moderation” period from the early 1980s up to the recent financial crisis, but not before or since—then real returns on money market instruments will be relatively stable over the long term.
2. *Equities*: When profit margins are stable, corporate profits, and hence dividends and dividend growth rates, tend to rise with the price level; thus, it is reasonable to regard equities as an inflation hedge in the long term (remembering that equity investors are exposed to additional risks in comparison to holders of inflation-indexed bonds).
3. *Corporate bonds*: As with equities, corporate bond performance is partly linked to inflation: Rising price levels drive up corporate revenues and reduce the real value of existing fixed-rate debt, and both these factors can cause yield spreads to tighten. However, this relationship is often weak and dominated by other factors.
4. *Commodities*: A basket of commodities also provides a partial hedge against inflation; in practice, this investment alternative was not historically as important as the previous three, though its importance has increased considerably since 2005 as financial innovation has expanded the investor base.

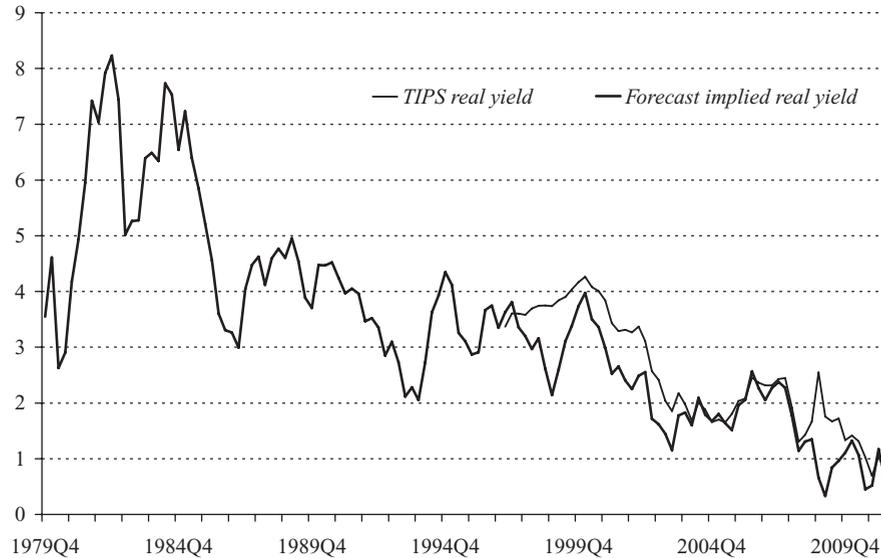
To summarize: Real yields are far from stable, and the behavior of real yields is just as complex as the behavior of nominal yields. Real yields are influenced by both economic fundamentals and market supply/demand factors across asset classes. It is not at all obvious that inflation-linked bonds should be “among the least risky of all assets.” Indeed, in the Australian market these securities were long regarded as highly risky in comparison to nominal bonds—though this is partly because of their poorer liquidity.

In all countries where inflation-linked bonds are actively traded, real yields have, historically, been quite volatile. Like nominal yields, market real yields trade in ranges of hundreds of basis points (see Figure 1). Historical examples from other countries include:

- In the U.K., real yields on long index-linked gilts fluctuated between 2% and 4.5% in the period 1981–1993.³ In the period 1984–1994, real yields on short index-linked gilts fluctuated between 1.5% and 5.75%, partly reflecting instability in monetary policy.⁴
- In Israel from 1984–1993, long-dated real yields fluctuated between –1.5% and 3.3%; however, they more typically traded in the range $\pm 1\%$.⁵
- In Australia, real yields have varied from a high of 5.75% in 1986 and 1994 to a low of 3.25% in 1993.⁶

Real yields are often estimated by subtracting current (i.e., recent historical) inflation from current nominal bond yields; but this procedure is obviously illogical, as it assumes that expected inflation is equal to current inflation. One can get a better idea of what market real yields would have been by taking nominal yields and subtracting a consensus inflation forecast. Figure 1 shows the 10-year nominal Treasury yield minus the 10-year consensus CPI forecast, as reported in the Philadelphia Fed’s Survey of Professional Forecasters; this measures investors’ expectations of real returns on 10-year Treasury bonds and is therefore a reasonable estimate of the 10-year real yield going back several decades. Figure 1 also shows the market real yield of the 10-year TIPS (dating back only to 1997); it is correlated with the survey-based real yield estimate, but not perfectly. We discuss this divergence at the end of the entry.

Even though using consensus data has a number of drawbacks, this rough analysis yields some useful results. The figure shows clearly how long-dated real yields soared in the early



Source: Bloomberg; Federal Reserve Board

Figure 1 U.S. 10-Year Real Yield Estimated from Consensus Long-Term CPI Forecasts and TIPS Real Yield

1980s, due to the extreme instability in monetary policy. They stabilized after 1985, once the Fed stopped targeting monetary aggregates and adopted interest rate targeting instead. Since then they have fluctuated between 5% (in the overheated economy of the late 1980s) and less than 0.5% (in the crisis and postcrisis periods). Note the apparent link between long-term real yields and current GDP growth in recent years.

Figure 2 shows a more detailed history of TIPS real yields since issuance. It also shows the yield spread between the 10-year TIPS and the 10-year CMT nominal yield. This may be regarded as a rough measure of the market's inflation expectations over the next 10 years.

It's interesting that 10-year TIPS real yields have never been stable, whereas 10-year TIPS break-even inflation was remarkably stable from about 2004–2007, a period of relative macroeconomic stability and strong Fed credibility. Also note the extraordinary period of volatility during the crisis period of late 2008 and early 2009, during which TIPS were highly correlated with risky asset classes such as equities and credit (as predicted above, but per-

haps not for the fundamental economic reasons cited).

A derivative market for inflation swaps has developed alongside the cash market for inflation-linked bonds. While inflation swaps will not be discussed explicitly in this entry, much of the material is also applicable to them.

Existence of an Inflation Risk Premium

It is often asserted that real yields on inflation-linked bonds should reflect an *inflation risk premium*, since investors are not exposed to inflation risk as they are with nominal bonds. Note that if future inflation were known—not necessarily zero—there would be no inflation risk premium; it is uncertainty about inflation that creates a risk premium. The more volatile inflation is expected to be, the higher the inflation risk premium on nominal bonds should be, and the lower real yields should be in relation to nominal yields.

It is important to note that it is uncertainty about future inflation that should determine the

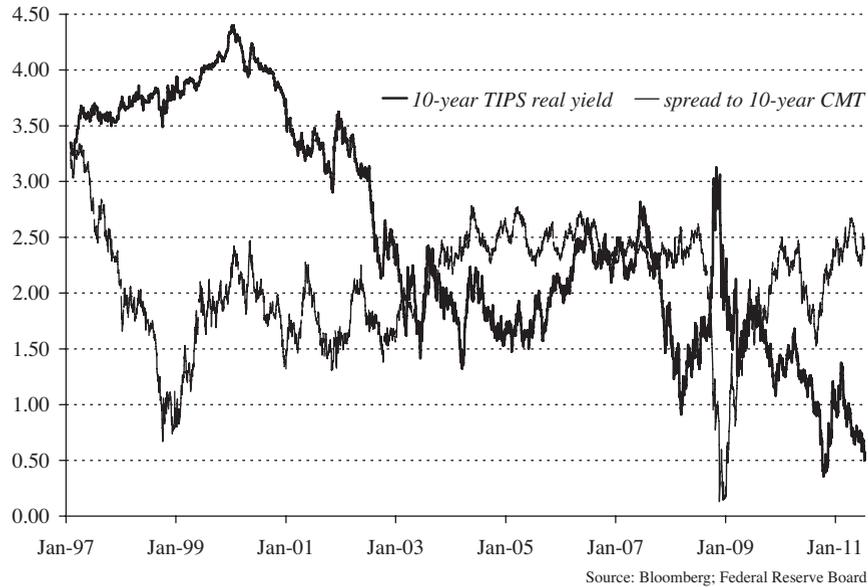


Figure 2 TIPS Real Yield History and Spread to Nominal Yield Curve (“Break-Even Inflation”)

risk premium, not the historical volatility of inflation. For example, the inflationary episode of the 1970s is not relevant unless investors think it may be repeated. Investors’ expectations about the future volatility of inflation are not directly observable, but it may be helpful to look at economists’ estimates. It is also useful to compare expected inflation volatility with expected volatility in real interest rates, since both factors are relevant to the risk/return opportunities offered by inflation-indexed bonds.

Note that if the inflation risk premium exists, one would not expect it to be unvarying. Since it is related to market expectations about potential uncertainty in inflation, it is comparable to option-implied volatility. One would thus expect the inflation risk premium to depend on bond maturity, and also to vary over time; for example, if the market lost confidence in the Fed’s ability or willingness to control inflation, the inflation risk premium would rise, causing nominal yields to rise relative to real yields. However, since the inflation risk premium is determined by inflation uncertainty over a long period (10 years for the

10-year TIPS), sudden changes should be unusual. Absent unusual shocks to Fed credibility, the inflation risk premium should experience moderate fluctuations, like long-dated swaption implied volatilities, and not sharp ones, like short-dated exchange-traded option implied volatilities.

In the absence of a complete inflation-linked derivatives market, the inflation risk premium is not directly observable. Furthermore, naive attempts to measure it can lead to grossly overstated estimates, and a number of proposed methods for measuring it turn out to be spurious. For example, it has been asserted that the differential between money market and bond yields arises because of an inflation risk premium, which can thus be estimated by looking at the long-term average spread between the Fed Funds rate and the two-year bond yield (about 70 bp in the period since deregulation). This argument has a grain of truth, but the conclusion is incorrect as it stands. The slope of the yield curve reflects a term premium that is not solely attributable to inflation risk. In addition, there are other reasons why money

market yields are usually lower than bond yields: Liquidity preference and the impact of capital charges both have important effects. Furthermore, if the spread between money market and bond yields reflects a risk premium, this is not just an inflation risk premium but a real rate risk premium as well.

Also, the argument that the difference between the Fed Funds rate and the two-year bond yield equals the inflation risk premium implies that the yields of money market securities reflect no inflation risk premium, while this risk premium is fully priced into two-year bond yields. This would only be plausible if money market securities were not (perceived to be) subject to inflation risk, and this is far from obvious, particularly since real money market returns were frequently negative during the 1970s.

Thus we must look for more valid ways of estimating what the inflation risk premium should be. There is no strong consensus in the literature, and a surprisingly wide range of estimates appears in the literature, from around 100 bp to modestly negative.⁷ However, the most credible estimates tend to fall in the zero to 50 bp range.⁸

One approach is to try to observe inflation uncertainty directly and then derive a “fair” inflation risk premium by applying a market price of risk. Figure 3 shows the probabilities attached by economists to various GDP growth and inflation scenarios; it is taken from the Survey of Professional Forecasters.

Economists’ forecasts recognize that both inflation and real yields are volatile, and that they have comparable volatilities. It is tempting to conclude that nominal bond yields should indeed reflect an inflation risk premium, since returns on nominal bonds are affected by both inflation volatility and real yield volatility, while returns on inflation-linked bonds are only affected by real yield volatility. And the reported uncertainty in inflation naively leads to an (again, very rough) estimate of the inflation risk premium at the upper end of the range

mentioned above. But this conclusion is not necessarily correct.

Based on an analysis of 30 years’ worth of cross-country panel data, Judson and Orphanides (1999) have shown that—as one might expect—there is a strong negative correlation between inflation and growth. Thus, as inflation rises, real yields should fall, and vice versa; in other words, the risks arising from fluctuations in inflation and fluctuations in real yields at least partly offset each other, at least over the medium to long term. It is therefore conceivable that, over the medium to long term, a portfolio of nominal bonds may be less risky, not more risky, than a portfolio of inflation-linked bonds, in which case an inflation risk premium need not exist at all. Certainly the situation is more complex than it seems at first.

We can actually use the earlier “economists’ estimates” of volatility in real GDP growth and CPI inflation, together with the implied volatility of short-term rates, to compute a rough estimate of the correlation between inflation and growth. Assuming that nominal rates are solely determined by growth and inflation, we have:

$$\sigma_{nom}^2 = \sigma_{GDP}^2 + \sigma_{CPI}^2 + 2\rho\sigma_{GDP}\sigma_{CPI}$$

where σ_{GDP} , σ_{CPI} denote the volatility of growth and inflation respectively, and ρ is the correlation between growth and inflation.

If all three volatilities are around 1% per annum, then solving this formula for ρ gives an estimate of around 0.5. However, it would not be meaningful to try to compute a more precise estimate this way.

Incidentally, Judson and Orphanides (1999) also found a strong negative correlation between inflation volatility and growth. In other words, if inflation is expected to become more volatile, real yields should fall, that is, inflation-indexed bond prices should rally. However, in this scenario inflation-linked bonds should outperform nominal bonds, since the inflation risk premium should rise.

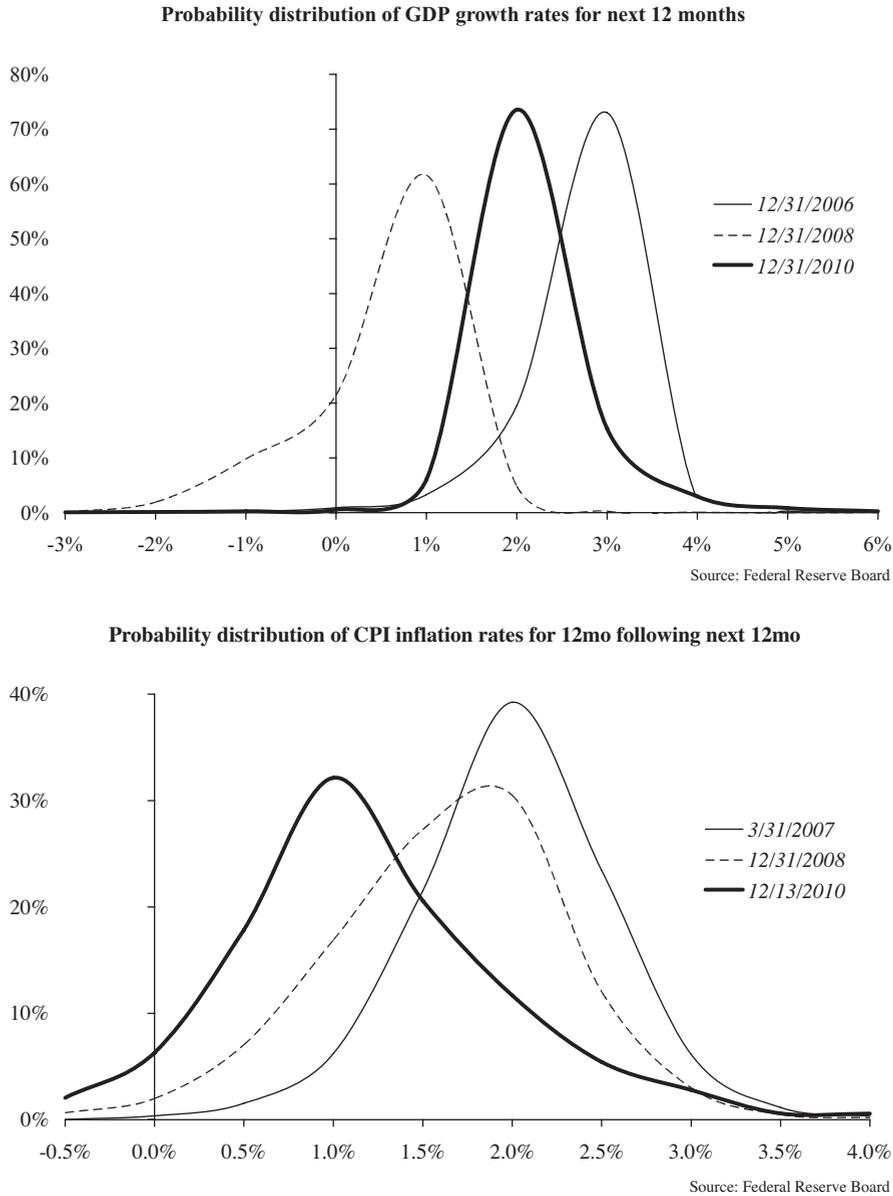


Figure 3 Economists' Uncertainty about Future GDP Growth and Future Inflation

INFLATION-INDEXED BONDS IN A NOMINAL PORTFOLIO

TIPS behave in unique ways and resemble neither nominal Treasuries nor spread products. It's therefore worth going back to basics in order to understand the nature of the interest rate

risk inherent in TIPS and the role they can play in broader portfolios.

What Is the Duration of an Inflation-Indexed Bond?

Inflation-indexed bonds are often used for specialized purposes (e.g., asset/liability

management for insurance companies offering inflation-linked life annuities, or for defined benefit plans where benefits are subject to cost of living adjustments), and may thus be segregated from other fixed-income holdings. However, if they are held in the same portfolio as nominal bonds, an interesting problem arises when attempting to define their price sensitivity to rate changes as measured by “duration.” We first examine the simplest possible definition of duration and its consequences; then we look at some alternative definitions.

It is easy to compute the duration of an inflation-indexed bond using exactly the same method as one would use for a nominal bond. Because of its low real coupon and low real yield, an inflation-indexed bond tends to have a much longer duration than a nominal bond of comparable maturity.

But what does this duration mean? The duration of a nominal Treasury bond measures its sensitivity to changes in nominal yields, that is, to changes in inflation and real interest rate expectations. By contrast, the duration of an inflation-linked bond measures its sensitivity to changes in real yields, that is, to changes in real interest rate expectations alone. In other words, the two durations are not comparable: They are measuring different things. So, for example, it does not make sense to look for a “reference” nominal yield for the TIPS real yield: While the TIPS yield may appear to trade off the 10-year Treasury during some periods, or off the 5-year Treasury during other periods, there is no fundamental reason why any such relationship should persist.

This creates a problem at the portfolio level. If we try to compute a portfolio duration by adding up the durations of nominal and inflation-indexed bond holdings, what does the resulting figure mean? Two portfolios could have the same duration but, depending on the relative weighting of index-linked bonds, might have a very different response to a change in investors’ economic expectations. A simple duration target is no longer an

effective way of controlling portfolio interest rate risk.

Thus, when a portfolio contains both nominal and inflation-linked bonds, it is critical to monitor and report the relative weights and durations of the “nominal” and “real” components of the portfolio separately. One approach is to report two durations for the portfolio, which distinguish two sources of risk:

1. A “portfolio real yield duration” equal to the sum of the durations of both nominal and inflation-indexed bond holdings. This shows how the portfolio value will respond to a change in market real yields (which also affect nominal yields).
2. A “portfolio inflation duration” equal to the duration of the nominal bond holdings alone. This shows how the portfolio value will respond to a change in market inflation expectations (which affect nominal yields but not real yields).

Similarly, care must be taken when carrying out portfolio simulations. For example, when carrying out parallel interest rate simulations, it is standard practice to apply an identical yield shift to all securities in the portfolio. For a portfolio containing both nominal and inflation-indexed bonds, this actually corresponds to a “real yield simulation.” One should also carry out “expected inflation simulations,” where the yield shift is applied to nominal but not inflation-indexed bond yields.

There is one practical situation in which it makes sense to compare the durations of a nominal and inflation-indexed bond directly: when designing trading strategies based on expected inflation. Suppose the central banking authority is targeting a long-term core CPI inflation rate of no more than 2%; and suppose that the 10-year nominal yield is 3.5% while the 10-year real yield is 0.5%. This means that the market is predicting an average headline CPI inflation rate, over the next 10 years, of 3%. If one had faith in the central bank’s ability to meet its inflation target and one did not believe headline

CPI would consistently outpace core CPI over the next decade, nominal bonds would look undervalued relative to inflation-indexed bonds.

How should one exploit this perceived opportunity without changing exposure to other sources of risk? The correct way is to execute a duration-matched swap, selling 10-year inflation-indexed bonds and buying 10-year nominal bonds. If inflation expectations fall, the strategy would realize a profit. If real interest rate expectations change (i.e., if real yields change), there would be no effect—which is the intention. (This kind of strategy can be implemented more precisely using a full term structure of market inflation forecasts, and incorporating short horizon economist forecasts.)

The above duration calculation is based on the (known) real cash flows and discounts at the real yield. There are other potential ways to compute the “duration” of an inflation-linked bond, which involve forecasting the (unknown) nominal cash flows and discounting using nominal yields on a zero coupon curve basis. The three most obvious alternatives are:

1. Using a fixed inflation forecast, generate projected bond cash flows (one should use a forecast that ensures that the net present value of

the forecast cash flows, discounted using the current nominal zero coupon curve, is equal to the current bond price). Compute the duration of this fixed cash flow stream using ± 100 bp shifts in the nominal zero coupon curve.

2. The same, except that when shifting the zero coupon curve by ± 100 bp, one recalculates the bond cash flows based on a new inflation forecast, adjusted by $\pm 1\%$. That is, the cash flow stream is assumed to depend on the level of nominal yields.
3. The same, except that one adjusts the inflation forecast by an amount different from $\pm 1\%$. For example, Figure 4 shows that historically, a 10 bp rise in U.S. nominal yields corresponded, on average, to a 9 bp rise in market long-term inflation expectations (though with much variation around that average). Thus one might adjust the inflation forecast by (say) $\pm 0.9\%$. The precise number depends on the reference Treasury yield, and the historical period used to estimate the relationship.

In each case, some minor variations are possible; for example, either constant or time-varying inflation forecasts could be used. These

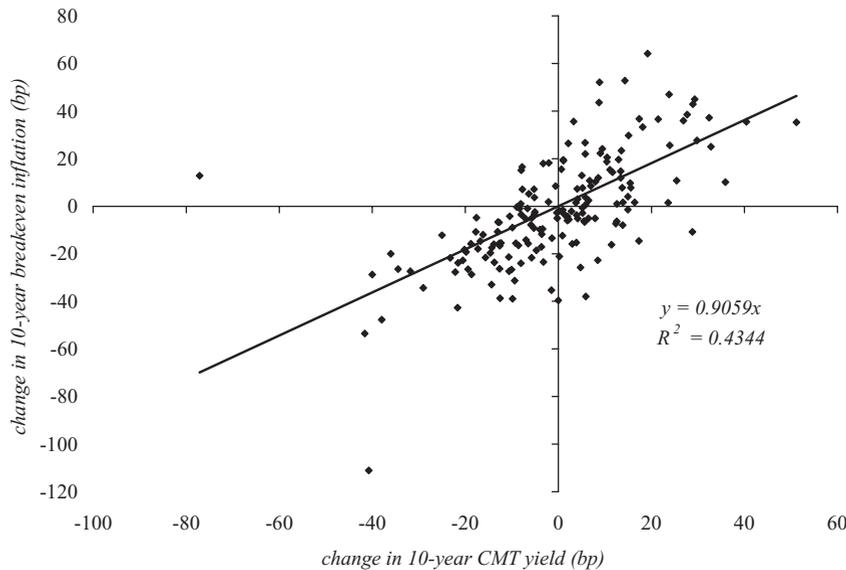


Figure 4 Shift in Nominal Yield versus Shift in Break-Even Inflation

calculations can be related to the above concepts of “real yield duration” and “inflation duration” in the following way:

1. Assuming a fixed cash flow stream (i.e., a fixed inflation scenario) amounts to assuming that the ± 100 bp shift in nominal yields is due to a change in real yields, not a change in inflation expectations. Thus, this calculation determines the sensitivity to a change in real yields, that is, it is essentially computing a real yield duration and produces an answer very close to the duration calculation described above.
2. Assuming an inflation scenario that varies by $\pm 1\%$ amounts to assuming that the ± 100 bp shift in nominal yields is due to a change in inflation expectations. Thus, this calculation measures an inflation duration, that is, a sensitivity to a shift in market inflation expectations, which is conceptually different from the real yield duration. The inflation duration of a TIPS will be approximately zero, but it may depend on the precise way the calculation is carried out.
3. Assuming an inflation scenario that varies by some amount based on the empirical relationship between nominal yields and market inflation expectations amounts to calculating a nominal yield duration, which attempts to measure the sensitivity of an inflation-linked bond to a shift in nominal yields.

Real yield duration is the most important of these risk measures—and, as we have seen, it can be calculated without using an inflation forecast. The inflation duration is not a useful risk measure for TIPS; however, in the U.K. and Australian markets, where inflation-indexed Treasuries have some residual inflation sensitivity due to the lag in inflation indexation, inflation duration is perhaps worth monitoring. The definition of nominal yield duration makes essential use of an estimate about an empirical relationship that is probably unstable, severely limiting the usefulness of this risk measure.

Note that if inflation-indexed Treasury bonds did have stable nominal durations—that is, if they did respond in an absolutely predictable way to a change in nominal yields—then they would not be a very useful risk management tool, since their mark-to-market behavior could be perfectly replicated by nominal bonds, which, moreover, are more liquid. In fact, experience shows that inflation-indexed bonds cannot be hedged perfectly with nominal bonds.

One can also attempt to compute a “tax-adjusted duration” for an inflation-linked bond, which takes its tax treatment into account; this may be of importance in the U.K., where inflation accruals are not taxed. In the U.S. market inflation-linked and nominal bonds are taxed on a broadly consistent basis; in particular, by analogy with Treasury STRIPS, the inflation adjustment to the bond principal is taxable as it occurs, and not simply at bond maturity. Thus, just as one continues to use pretax durations for Treasury STRIPS despite their tax treatment, it seems reasonable to use pretax durations for TIPS as well. The trading behavior of inflation-linked bonds in a range of markets suggests that pretax duration measures suffice for most day-to-day interest rate risk management. However, it is worth discussing tax briefly.

The Impact of Taxation: An Outline

Inflation-indexed bonds attempt to eliminate inflation risk, but it reappears on an after-tax basis. We begin with the fact that tax affects returns on both nominal bonds and inflation-indexed bonds in an unfortunate way: High inflation results in lower after-tax real returns. For inflation-indexed bonds, an investor would reason as follows:⁹

$$\begin{aligned}
 & \text{forecast after-tax real yield} \\
 &= \text{forecast after-tax nominal yield} \\
 &\quad - \text{forecast inflation} \\
 &= \text{tax rate} \times \text{forecast pretax nominal yield} \\
 &\quad - \text{forecast inflation} \\
 &= \text{tax rate} \times (\text{pretax real yield} \\
 &\quad + \text{forecast inflation}) - \text{forecast inflation}
 \end{aligned}$$

$$= \text{tax rate} \times \text{pretax real yield} - \\ (1 - \text{tax rate}) \cdot \text{forecast inflation}$$

For nominal bonds, the reasoning is similar:

$$\begin{aligned} &\text{forecast after-tax real yield} \\ &= \text{forecast after-tax nominal yield} \\ &\quad - \text{forecast inflation} \\ &= \text{tax rate} \times \text{forecast pretax nominal yield} \\ &\quad - \text{forecast inflation} \\ &= \text{tax rate} \times (\text{pretax real yield} \\ &\quad + \text{market inflation}) - \text{forecast inflation} \\ &= \text{tax rate} \times \text{pretax real yield} \\ &\quad - (\text{forecast inflation} \\ &\quad - \text{tax rate} \cdot \text{market inflation}) \end{aligned}$$

where “forecast inflation” refers to the investor’s inflation forecast and “market inflation” refers to the market’s inflation forecast as reflected in the spread between market nominal yields and market real yields.

Thus an investor who agrees with the market’s inflation forecast and who is thus indifferent between inflation-linked bonds and nominal bonds on a pretax basis will also be indifferent on an after-tax basis. The arguments show that projected after-tax real returns on both inflation-indexed and nominal bonds depend on forecast inflation.

An important consequence is that since U.S. inflation-indexed bonds and nominal bonds are affected equally, inflation-linked bonds do not protect investors against the negative after-tax impact of high inflation. Thus, TIPS real yields reflect only a premium for “pretax inflation risk.” By contrast, since U.K. index-linked gilts receive preferential tax treatment, their yields also reflect a premium for “after-tax inflation risk.” The price paid by U.K. investors, as observed by Roll (1996) and by Brown and Schaefer (1996), is lower liquidity: The market for index-linked gilts is confined to investors with high marginal tax rates and to investors who have other incentives, such as regulatory incentives, to own inflation-linked securities.¹⁰ Roll (1996) points out a further consequence: If the demand for inflation-indexed or nomi-

nal bonds is a function of expected after-tax returns, pretax real yields should rise as expected inflation rises, to maintain a constant after-tax real yield. It is not clear whether real yields on inflation-indexed bonds actually behave in this way, although the Australian experience in 1994 suggests that they do. In any case, this introduces a further source of uncertainty about the future behavior of real yields.

Inflation-Indexed Bonds and Portfolio Efficiency

Inflation-indexed bonds have a risk profile quite different from that of nominal bonds. In fact, it could be argued that for asset allocation purposes, they should not be grouped with nominal bonds but should be treated as an entirely separate asset class. We will use portfolio theory to explore the consequences of adopting this point of view. More specifically, we will try to determine what weight TIPS should have in efficient portfolios with varying degrees of risk, and what impact their inclusion has on expected returns.

For simplicity, we focus on maximizing nominal returns in the U.S. market, and we work in a total return framework. Other kinds of analysis are possible; for example, Eichholtz, Naber, and Petri (1993) discuss the problem of matching inflation-indexed liabilities in the U.K. and Israeli markets.

The results of any Markowitz-style analysis are always highly dependent on the expected returns, volatilities, and correlations used. The assumptions we use are set out in Table 1 and are broadly based on market data and presumed market expectations. They were derived as follows:

1. Expected nominal returns for cash and nominal bonds (aggregate bond index) are based on current market yields—this is more meaningful than using historical returns. For simplicity, we assume that nominal bonds and TIPS have the same expected return (in practice, one would derive expected returns more

Table 1 Assumptions Used in Efficient Portfolio Analysis

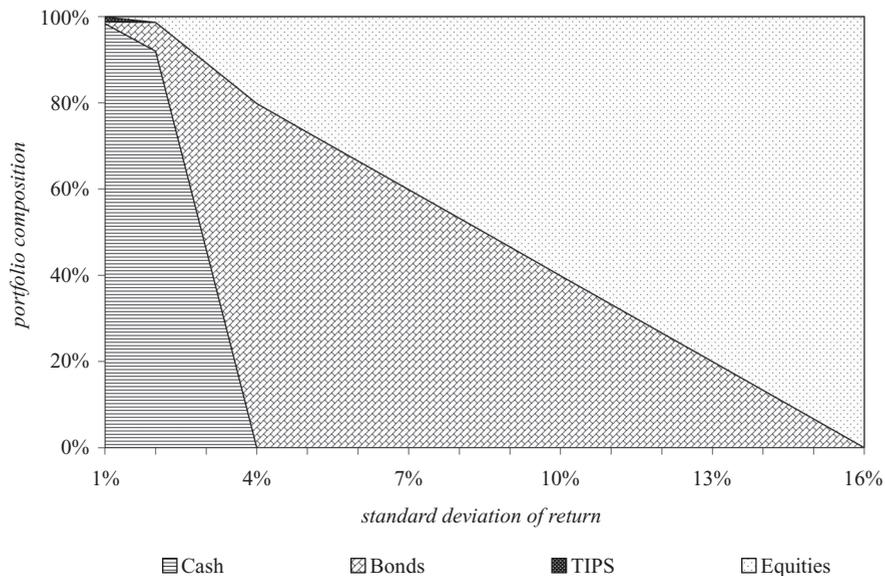
	Cash	Bonds	TIPS	Equities
Expected return	2.0%	2.8%	2.8%	5.8%
Return volatility	0.6%	3.6%	5.9%/3.6%	16.5%
Correlations				
Cash	1.00	0.06	-0.03	0.01
Bonds		1.00	0.73	-0.01
TIPS			1.00	0.03
Equities				1.00

carefully). The expected return for equities is obtained by adding a risk premium of 3% to that for bonds.

- Return volatilities for cash, nominal bonds, TIPS, and equities are historical, calculated using monthly Barclays index return data and S&P 500 return data over the period 1997–2011. In addition to the historical volatility for TIPS—which is quite high, largely due to the experience during the crisis period—we also carry out an alternative analysis that assumes that they have the same volatility as nominal bonds.
- We use historical correlations estimated using the same time period.

At first glance, the results seem highly dependent on the volatility assumption used for TIPS. Figure 5 shows the composition of theoretically efficient portfolios with varying degrees of risk, using the realistic volatility assumption; Figure 6 shows the same, using the low volatility assumption. Using the higher volatility, TIPS play almost no role in any efficient portfolio; for example, at moderate risk levels, nominal bonds are preferred because of their lower correlation with equities. However, using the lower volatility, TIPS have a much more important role to play. They partly displace cash at low risk levels, and more importantly they partly displace nominal bonds at moderate risk levels. Only the equity weightings remain more or less unchanged.

But how much value do TIPS actually add? Figure 7 shows the efficient frontier; that is, expected returns from efficient portfolios, calculated using both the realistic and low TIPS volatility assumptions. Above the 2% risk level, they are very close: Expected returns differ only marginally. That is, even assuming that TIPS will have a very low return volatility does not significantly increase their expected value

**Figure 5** Composition of Efficient Portfolios, 5.9% Volatility Assumption

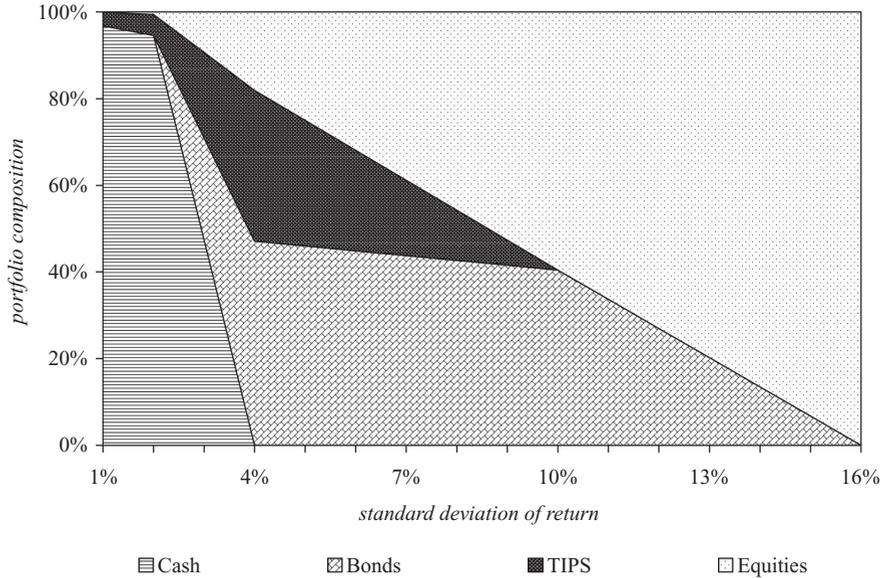


Figure 6 Composition of Efficient Portfolios, 3.6% Volatility Assumption

added to portfolio returns unless different expected return assumptions are used as well (or unless we move beyond the pure mean-variance framework).

Figure 8 is even more telling. It shows expected returns from efficient portfolios under the low TIPS volatility assumption for both

unconstrained portfolios and portfolios from which TIPS have been excluded. Even at moderate risk levels, where TIPS are most important, the difference in expected returns is extremely modest. Moreover, an investor who currently held a TIPS-free portfolio, and who wanted to capture these additional few basis points by

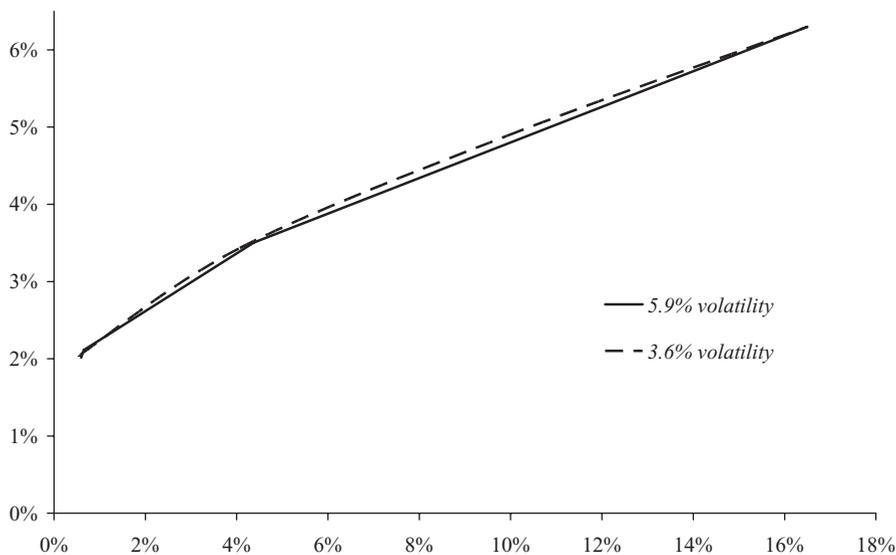


Figure 7 Efficient Frontier for the Two Different Volatility Assumptions

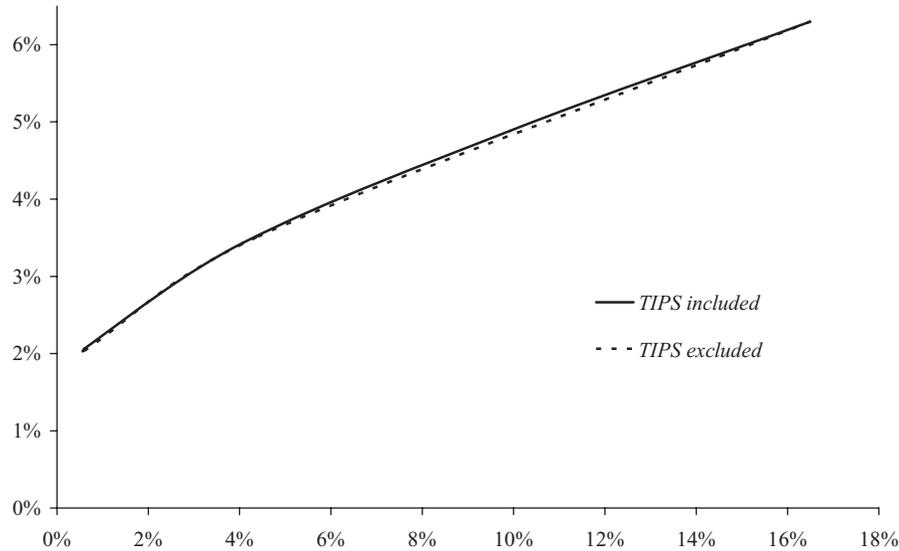


Figure 8 Efficient Portfolios Including and Excluding TIPS

purchasing TIPS, would have to trade over a quarter of the portfolio to achieve the optimal asset class weightings.

The overall conclusions are that (1) a realistic TIPS return volatility assumption, consistent with historical experience, implies that TIPS do not add much value to asset allocation; and (2) even under a very optimistic TIPS return volatility assumption, the value added by TIPS is modest. The main reasons are that TIPS do not have a higher expected return than nominal bonds, but have a slightly higher assumed correlation with equities.

These results should be compared with the findings of Eichholtz, Naber, and Petri (1993), who used data from 1983–1991 and discovered a significant difference between relatively low-inflation countries such as the U.K. in that period and countries such as Israel where inflation had been extremely high and volatile.

- Results for the U.K.: If the goal is to maximize total return, inflation-linked bonds do not appear in any efficient portfolio. If inflation-linked liabilities are included in the problem (but setting regulatory considerations aside), they appear in very low-risk efficient portfolios, but with negligible weight: less than 1%.

- Results for Israel: If the goal is to maximize total return, inflation-linked bonds play a minor role in low-risk portfolios but a major role in risky portfolios, sometimes having a weight of over 50%. If inflation-linked liabilities are included in the problem, inflation-linked bonds play a major role at all levels of risk, with weights between 44% and 88%.

TIPS provide insurance against inflation, and each investor's subjective assessment of future inflation risk and the need for inflation protection must strongly influence any conclusions about the role of TIPS. U.S. investors will have to decide which set of results provides more useful guidance.

ADVANCED ANALYTICAL APPROACHES TO INFLATION-INDEXED BONDS

As the U.S. TIPS market has matured, with a full term structure of maturities and a trading history spanning several business cycles and inflation environments, investors have developed many analytical approaches in the search for

investment opportunities. Rather than attempting a comprehensive survey, the remainder of this entry gives two brief examples, both of them focusing on economic factors rather than supply/demand relationships or “market technicals.” Standard econometric techniques turn out to be useful.

Link between TIPS Performance and Short-Term Inflation

The relative performance of TIPS versus nominal Treasuries is determined both by daily mark-to-market movements and by inflation accrual, which influences both inflation-adjusted principal and interest payments. Inflation accrual is clearly determined by realized headline CPI inflation (relative to nominal yields). Since headline CPI is quite volatile, this “carry” component of TIPS returns can often be a dominant factor in the performance of short and even intermediate maturity TIPS.

One useful way of looking at this is by isolating the impact of the more volatile components of CPI. Bryan and Meyer (2010) divide the CPI basket into “flexible price” and “sticky price” categories, leading to two separate measures

of inflation. Examples of flexible price items are gasoline, fruit and vegetables, and women’s apparel; examples of sticky price items are furniture, alcoholic beverages, and public transportation.

Figure 9 shows that during the period since the introduction of TIPS in 1997, there was a positive correlation between changes in 10-year TIPS break-even inflation and changes in three-month flexible price CPI inflation. It may seem surprising that 10-year inflation expectations are visibly influenced by realized three-month inflation; but this simply reflects the strong influence of carry on the trading behavior of TIPS, even 10-year maturity TIPS.

We can get a more refined view of the relationship if we run a vector autoregression analysis and examine the impulse-response functions. Some of the results from an analysis using 2003–2011 data on five-year TIPS (i.e., a shorter maturity, more strongly influenced by carry considerations) are shown in Figure 10. The impulse-response functions suggest that:

1. 5-year TIPS break-even inflation responds more strongly to flexible price CPI than

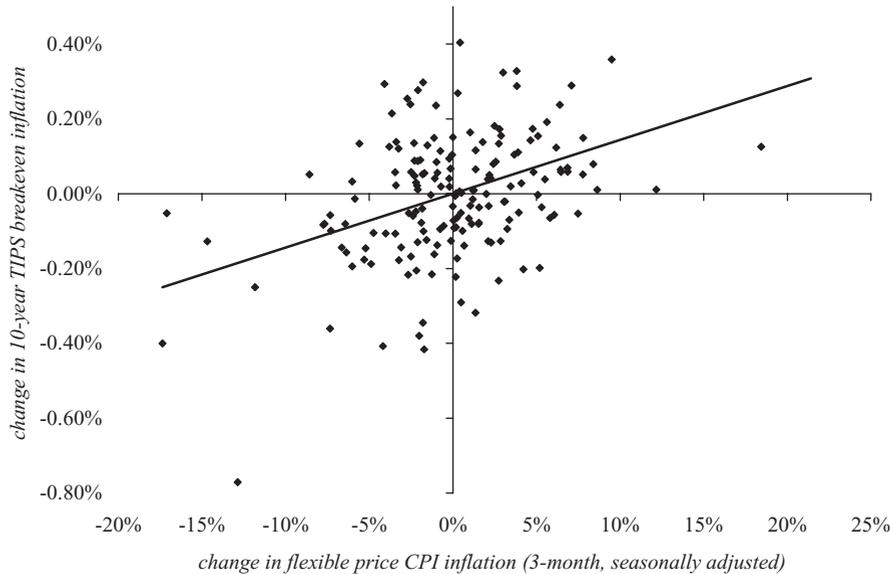


Figure 9 TIPS Performance and Flexible Price CPI

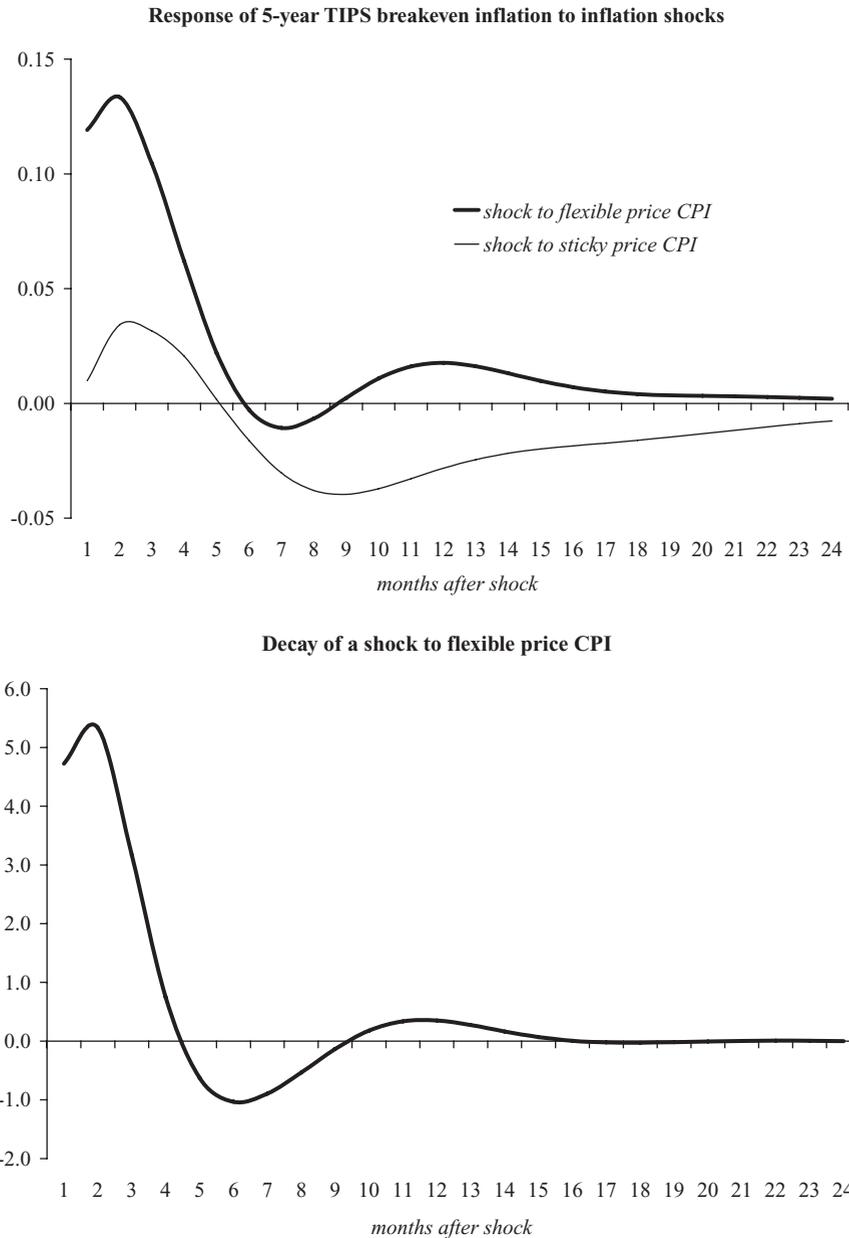


Figure 10 Five-Year TIPS Break-Even Inflation and Shocks to Flexible Price CPI Inflation

- to sticky price CPI, that is, the sources of shorter-term inflation volatility are more important; despite the fact that
2. Shocks to flexible price CPI inflation tend to be quite short-lived, dissipating after a couple of months and even tending to (partially) correct.

Since TIPS inflation accrual is based on non-seasonally-adjusted headline CPI inflation, a further aspect of TIPS carry is the strong seasonal pattern exhibited by CPI inflation. This needs to be analyzed separately. Seasonal factors have often been a source of market inefficiency in the past.

The TIPS Premium versus Survey-Based Real Yield Measures

As can be seen from Figure 1, TIPS real yields have usually (but not always) been higher than the real yields implied by subtracting consensus inflation forecasts from observed nominal Treasury yields. In other words, TIPS real yields usually incorporate an apparent “concession.” The historical behavior of this apparent real yield premium is shown in Figure 11, together with an estimate of its trend behavior (derived by applying a standard Hodrick-Prescott filter¹¹).

This premium has averaged around 40–50 bp, but has fluctuated quite a bit over time. It seems to mean revert to trend (heavy line in Figure 11) over about a 12-month period on average.

Why would this premium exist?

1. *Survey bias*: Economists’ forecasts of future inflation may be systematically biased (higher) relative to the market’s forecasts. This is more likely to have been true during the period of declining trend inflation from the mid-1980s to the mid-2000s; and indeed the real yield premium seems to have

decreased since then, though it has still been positive on average.

2. *Recalculation risk*: There may be a downward bias to the risk of future changes to the definition of CPI.¹²
3. *Liquidity*: TIPS are less liquid than nominals (i.e., they have wider and more uncertain bid/ask spreads, and greater market impact of large trades), so investors require a higher real yield to compensate for that.
4. *Tracking error*: TIPS aren’t in the standard bond indexes, so index-sensitive investors need to be compensated for the fact that owing TIPS leads to additional tracking error.
5. *Undesirable correlations*: TIPS tend to underperform in deflations/recessions, which is when investors most want bonds to do well; another kind of undesirable correlation is that TIPS liquidity tends to deteriorate in periods of general market stress.

The first two factors are difficult to quantify, but one could argue that their influence is probably fairly constant over time. It thus seems feasible to develop relative value measures conditioned on the remaining three factors, which are potentially more tractable.

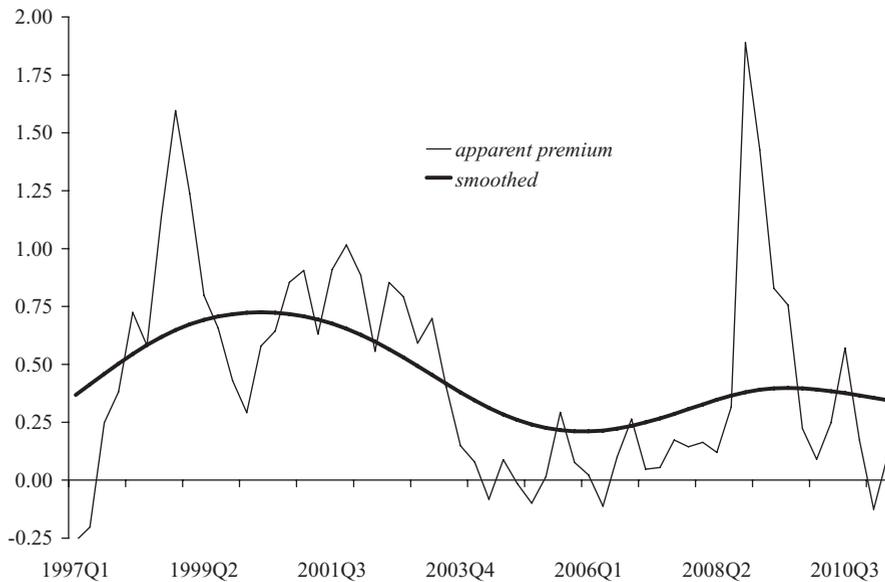


Figure 11 Apparent TIPS Real Yield Premium

The (il)liquidity premium turns out to be particularly important, since it exhibits the most time variation. It is difficult to estimate based on yield data alone;¹³ furthermore, as with all liquidity premiums, it is not fully determined by prevailing liquidity conditions (bid/ask spreads, quoted volumes, and market impact of trades) but is also influenced by the perceived risk that future liquidity conditions may differ from today's.

Modeling liquidity premiums is extremely difficult, but useful information can be extracted via model-free approaches. For example, Christensen and Gillan (2011) argue that the difference between TIPS break-even inflation and inflation swap rates provides a time-varying upper bound on the TIPS liquidity premium. This upper bound has typically fluctuated between 10 bp and 20 bp, but rose to over 100 bp in late 2008 during the financial crisis; it has been highly correlated with other measures of bond liquidity, such as the yield premium of off-the-run versus on-the-run nominal Treasuries.

KEY POINTS

- TIPS real yields are volatile. They are influenced by domestic growth, external balances, and the behavior of competing asset classes.
- TIPS real yields also reflect a modest and somewhat volatile inflation risk premium.
- There are different notions of *TIPS duration* corresponding to different aspects of TIPS interest rate risk.
- TIPS often do not play a significant role in efficient portfolios, and some investors may be better off regarding them as opportunistic rather than core investments.
- Market returns on TIPS are often driven by short-term inflation accrual, and this is best analyzed by breaking observed inflation down into suitable components.
- Survey-based measures of inflation and real yields often differ from those implied by the

TIPS market, and it is important for investors to understand the reasons for the divergence.

NOTES

1. The following discussion of risk factors expands on the account in Carmody and Mason (1996).
2. See Chapter 12 in Keynes (1936).
3. See Eichholtz, Naber, and Petri (1993).
4. See Brown and Schaefer (1996).
5. See Eichholtz, Naber, and Petri (1993).
6. See Carmody and Mason (1996).
7. See the citations in Grishchenko and Huang (2009).
8. See, for example, D'Amico, Kim, and Wei (2008) and Durham (2006).
9. See Roll (1996) for more details.
10. By the way, this provides an example of the tax clientele effects analyzed by Dybvig and Ross (1986).
11. The Hodrick-Prescott filter developed in Hodrick and Prescott (1997) is an econometric technique employed in macroeconomics in the analysis of time series data.
12. See the change in the calculation of CPI following the recommendations of the Boskin Commission (1996).
13. D'Amico, Kim, and Wei (2008).

REFERENCES

- Boskin Commission (1996). *Toward a More Accurate Measure of the Cost of Living: Final Report to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index*, December 4, 1996; available at <http://www.ssa.gov/history/reports/boskinrpt.html>
- Brown, R., and Schaefer, S. (1996). Ten years of the real term structure: 1984–1994. *Journal of Fixed Income* 6: 6–22.
- Bryan, M. F., and Meyer, B. (2010). Are some prices in the CPI more forward looking than others? We think so. Economic Commentary 05.19.2010, Federal Reserve Bank of Cleveland; available at <http://www.clevelandfed.org/Research/commentary/2010/2010-2.cfm>

- Carmody, S., and Mason, R. (1996). *Analysis of Australian Index-Linked Securities*. Deutsche Morgan Grenfell (Sydney) research report, June.
- Christensen, J., and Gillan, J. M. (2011). A model-independent maximum range for the liquidity correction of TIPS yields. Working Paper 2011-16, Federal Reserve Bank of San Francisco.
- D'Amico, S., Kim, D. H., and Wei, M. (2008). Tips from TIPS: The informational content of Treasury Inflation-Protected Security prices. Federal Reserve Board Working Paper 2008-30, February.
- Durham, J. B. (2006). An estimate of the inflation risk premium using a three-factor affine term structure model. Federal Reserve Board Working Paper 2006-42.
- Dybvig, P., and Ross, S. (1986). Tax clienteles and asset pricing. *Journal of Finance* 41: 751-771.
- Eichholtz, P., Naber, P., and Petri, V. (1993). Index-linked bonds in a liability framework. *Journal of Fixed Income* 3: 54-62.
- Grishchenko, O. V., and Huang, J. (2009). Inflation risk premium: Evidence from the TIPS market. 20th Anniversary Conference on Financial Economics and Accounting, Rutgers University.
- Hodrick, R., and Prescott, E. C. (1997). Postwar U.S. business cycles: An empirical investigation. *Journal of Money, Credit, and Banking* 29: 1-16.
- Judson, R., and Orphanides, A. (1999). Inflation, volatility and growth. *International Finance* 2(1): 117-138.
- Keynes, J. M. (1936). *The General Theory of Employment, Interest and Money*. London: Macmillan.
- Roll, R. (1996). US Treasury inflation-indexed bonds: The design of a new security. *Journal of Fixed Income* 6: 9-28.

Credit Risk Modeling

An Introduction to Credit Risk Models

DONALD R. VAN DEVENTER, PhD

Chairman and Chief Executive Officer, Kamakura Corporation

Abstract: Credit risk technology has evolved with advances in computer science and information technology. Traditional credit ratings date back to 1860, an era when the cost of collecting and analyzing corporate credit information was high. The commercial advantages of a central provider of credit risk analysis were high. With the advent of better computer technology and databases of corporate financial information and stock prices, quantitative approaches to credit risk assessment have become more popular and increasingly accurate. Credit scoring is a quantitative approach to retail credit assessment, but, in the corporate world, more and more credit analysts prefer a default probability with an explicit maturity to a “credit rating” or “credit score.”

This entry introduces the topic of *credit risk modeling* by first summarizing the key objectives of credit risk modeling. We then discuss ratings and credit scores, contrasting them with modern *default probability* technology. Next, we discuss why valuation, pricing, and hedging of credit risky instruments are even more important than knowing the default probability of the issuer of the security. We review some empirical data on the consistency of movements between common stock prices and credit spreads with some surprising results. Finally, we compare the accuracy of ratings, the *Merton model* of risky debt, and reduced form credit models.

KEY OBJECTIVES IN CREDIT RISK MODELING

In short, the objective of the credit risk modeling process is to provide an investor with practical tools to “buy low/sell high.”¹ Robert Merton, in

a 2002 story retold by van Deventer, Imai, and Mesler (2004), explained how Wall Street has worked for years to get investors to focus on expected returns, ignoring risk, in order to get investors to move into higher risk investments. In a similar vein, investment banks have tried to get potential investors in collateralized debt obligations (CDOs) to focus on “expected loss” instead of market value and the volatility of that market value on a CDO. The result, according to the Global Stability Report of the International Monetary Fund, was an estimated \$945 billion in global credit losses during the credit crisis that began in earnest in 2007.²

This means that we need more than a default probability. The default probability provides some help in the initial yes/no decision on a new transaction, but it is not enough information to make a well-informed yes/no, buy/sell decision, as we discuss below. Once the transaction is done, we have a number of very critical objectives from the credit risk modeling

process. We need to know the value of the portfolio, the risk of the portfolio (as measured most importantly by the random variation in its value), and the proper hedge of the risk if we deem the risk to be beyond our risk appetite. Indeed, the best single sentence test of a credit model is “What is the hedge?” If one cannot answer this question, the credit modeling effort falls far short of normal risk management standards. It is inconceivable that an interest rate risk manager could not answer this question. Why should we expect any less from a credit risk manager, who probably has more risk in his area of responsibility than almost anyone else? Indeed, stress testing with respect to macroeconomic factors is now standard under proposals from the European Central Bank and under the U.S. programs titled “Supervisory Capital Assessment Program” and “Comprehensive Capital Analysis and Review.” The latter programs, applied to the 19 largest financial institutions in the United States, focused on macro factors like home prices, real gross domestic product growth, and unemployment.

RATINGS AND “CREDIT SCORES” VERSUS DEFAULT PROBABILITIES

Rating agencies have played a major role in fixed income markets around the world since the origins of Standard & Poor’s in 1860. Even the “rating agencies” of consumer debt, the credit bureaus, play prominently in the banking markets of most industrialized countries. Why do financial institutions use ratings and credit scores instead of default probabilities? As a former banker myself, I confess that the embarrassing answer is “There is no good reason” to use a rating or a credit score as long as the *default probability modeling* effort is a sophisticated one and the inputs to that model are complete.

Ratings have a lot in common with interest accrual based on 360 days in a year. Both ratings and this interest accrual convention date

from an era that predates calculators and modern default probability technology. Why use a debt rating updated every 1–2 years when one can literally have the full term structure of default probabilities on every public company updated daily or in real time? In the past, there were good reasons for the reliance on ratings:

- Default probability formulas were not disclosed, so proper corporate governance would not allow reliance on those default probabilities.
- Default probability model accuracy was either not disclosed or disclosed in such a way that weak performance was disguised by selecting small sectors of the covered universe for testing.
- Default probability models relied on old technology, like the Merton model of risky debt and its variants, that has long been recognized as out of date.³
- Default probability models implausibly relied on a single input (the unobservable value of company assets), ignoring other obvious determinants of credit risk like cash balances, cash flow coverage, the charge card balance of the CEO of a small business, or the number of days past due on a retail credit.

With modern credit technology, none of these reasons are currently valid because there is a rich, modern credit technology available with full disclosure and an unconstrained ability to take useful explanatory variables. In this vein, ratings suffer from a number of comparisons to the modern credit model:

- Ratings are discrete with a limited number of grades. There are 21 Standard & Poor’s ratings grades, for example, running from AAA to D. Default probabilities are continuous and run (or should run) from 0 to 100%.
- Ratings are updated very infrequently and there are obvious barriers that provoke even later than usual response from the rating agencies, like the 2004 downgrade from AAA to AA- for Merck, a full three weeks after the

withdrawal of its major drug Vioxx crushed the company’s stock price. Another example is General Electric, first rated AAA in 1956, which was not downgraded until March 2009, a full four months after General Electric was forced to borrow under the Federal Reserve’s Commercial Paper Funding Facility.⁴ Default probabilities can adjust in real time if done right.

- Ratings have an ambiguous maturity, which we discuss in the next section. The full term structure of default probabilities is available and the obvious impact of the business cycle is observable: The full default probability term structure rises and falls through the business cycle, with short-term default probabilities rising and falling more dramatically than long-term default probabilities. Figure 1 illustrates this cyclical rise and fall during the credit crisis of 2007–2011 for Bank of America Corporation and Citigroup, two of the largest U.S. bank holding companies, using the reduced form model default probabilities discussed below and provided by Kamakura Corporation.

The cyclical rise and fall of default probabilities for both banks are very clear and show the impact of the credit crisis, which was at its height in 2007–2009. We take a longer-term view from 1990 to 2006 below.

Figure 2 shows clearly the joint rise in default probabilities in 1990–1991, a mini recession in 1994–1995, and the impact of the Russian debt crisis and high-tech crash in 1998–2002. By way of contrast, Standard & Poor’s only changed its ratings on Bank of America twice in the 1995–2006 period, once in 1996 and once in 2005.

What about consumer and small business “credit scores”? Like ratings and the interest accrual method mentioned above, these date from an era when there was limited understanding of credit risk in the financial community. Vendors of credit scores had two objectives in marketing a credit risk product: to make it simple enough for any banker to understand and to avoid angering consumers who might later learn how they are ranked under the credit measure. The latter concern is still, ironically, the best reason for the use of credit scores instead of default probabilities today on the retail side. From a

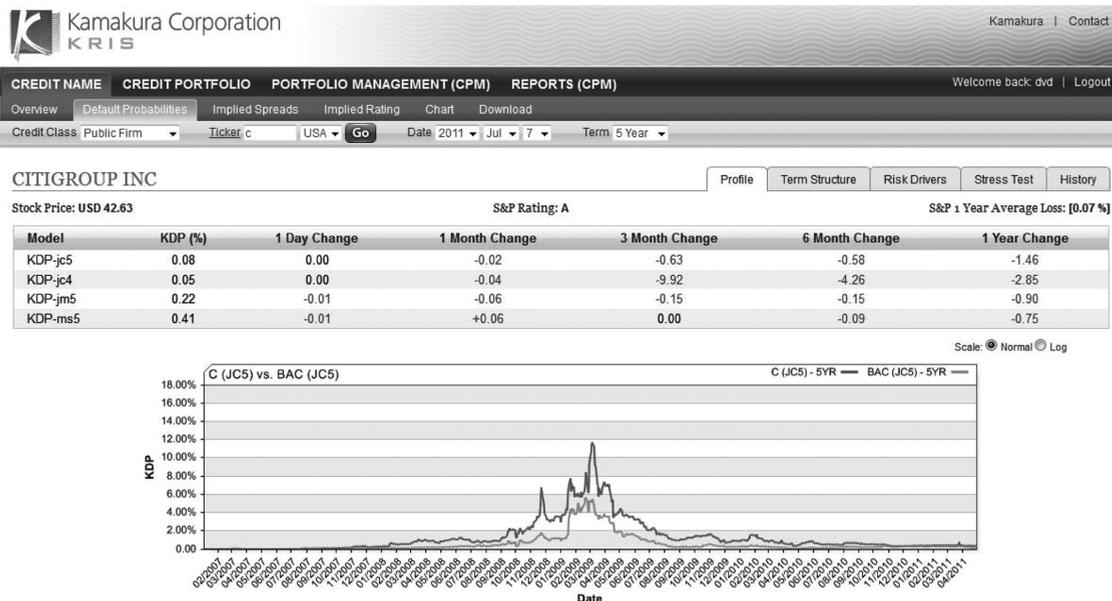


Figure 1 Five-Year Default Probabilities for Bank of America and Citigroup: January 1, 2007 to May 1, 2011

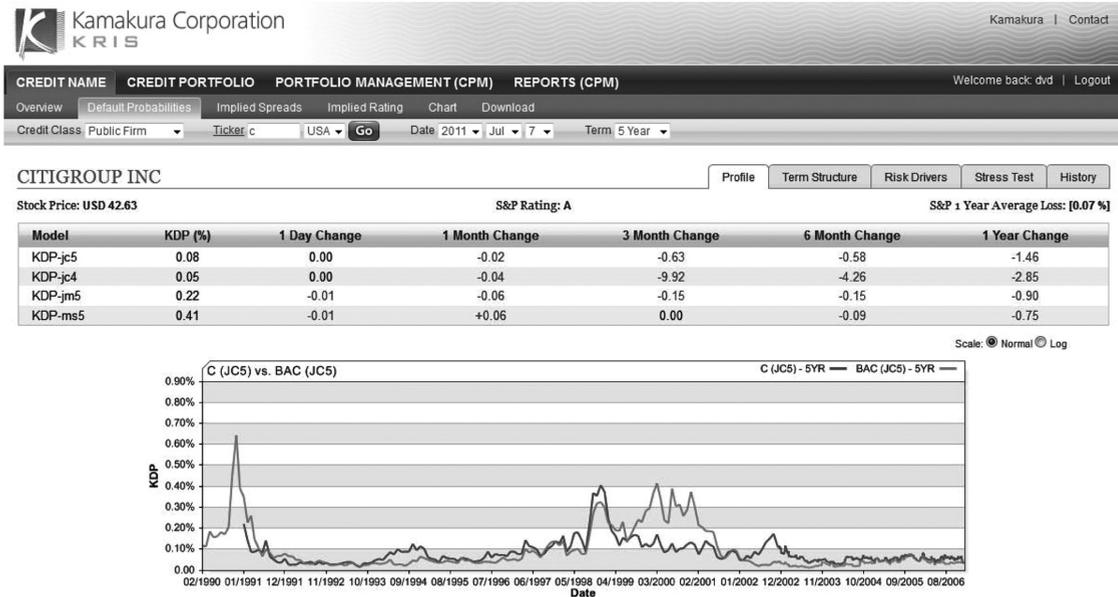


Figure 2 Five-Year Default Probabilities for Bank of America and Citigroup: January 1, 1990 to December 31, 2006

banker's perspective, though, the score hides information that is known to the credit score vendor. The credit scoring vendor is actually using the statistical techniques we describe below to derive a default probability for the consumer. They then hide it by scaling the default probability to run from some arbitrary range like 600 to 1,000 with 1,000 being best.⁵ One scaling that does this, for example, is the formula:

$$\text{Credit score} = 1,000 - 4 (\text{Consumer 1-year default probability})$$

This scaling formula hides the default probability that Basel II requires and modern bankers are forced to "undo" by analyzing the mapping of credit scores to defaults. This just wastes everyone's time for no good reason other than the desire to avoid angering retail borrowers with a cold-hearted default probability assessment.

The only time a rating or credit score can outperform a modern credit model is if there are variables missing in the credit model. Heading into the credit crisis as of December 31, 2006, for example, Citigroup had a roughly \$50 billion direct and indirect exposure to super senior

tranches of collateralized debt obligations, but these exposures were not reported in a quantitative form and therefore could not be used in a quantitative credit model. A judgmental rating in this case would be able to adjust for this risk if proper disclosure were made to the rating agencies. This, however, is a rare case and in general a first-class modeling effort will be consistently superior.⁶

WHAT "THROUGH THE CYCLE" REALLY MEANS

Financial market participants often comment that default probabilities span a specific period of time (30 days, 1 year, 5 years) while ratings are "through the cycle" ratings. What does "through the cycle" really mean?

Figure 3 provides the answer. It shows the term structure of default probabilities out for 10 years for Morgan Stanley on October 15, 2008, one month after the collapse of Lehman Brothers, and July 7, 2011. The July 7, 2011 term structure was quite low because business

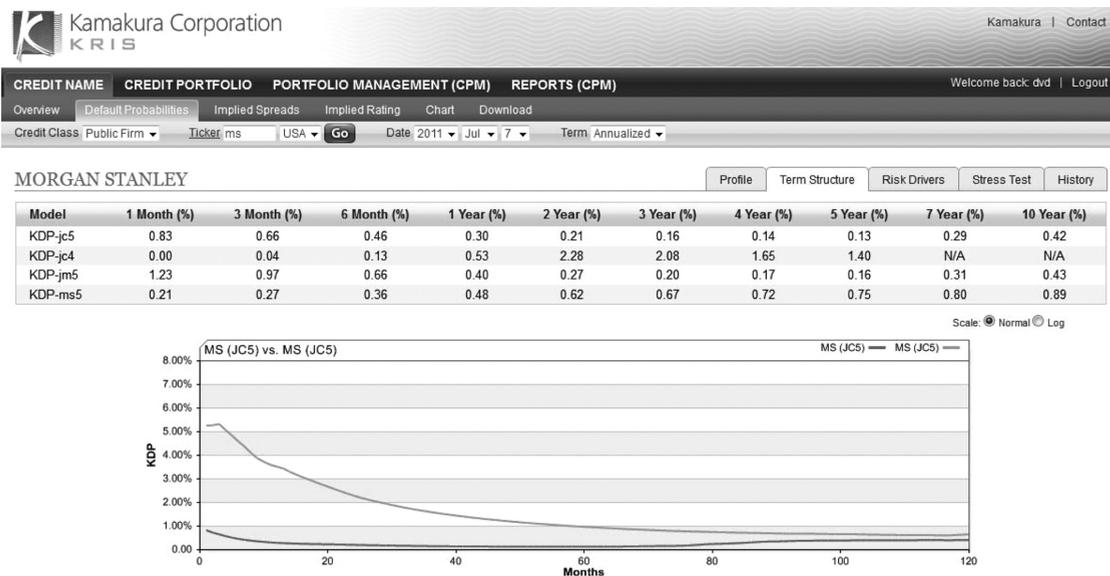


Figure 3 Term Structure of Default Probabilities for Morgan Stanley on October 15, 2008 and July 7, 2011

conditions at the time were excellent.⁷ Looking at the right-hand side of the curve, we can see that both default probability curves are converging and, if the graph is continued to a long enough maturity, both will hit about 42–50 basis points for a very long-term default probability.

This is consistent with the “long-run” default experience for both Morgan Stanley’s 2011 rating of A.⁸ Over the 15 years after being rated A, 2.77% of those formerly rated A defaulted. This is the same as a constant default rate over those 15 years of 18.7 basis points, a rate double the 8 basis point default rate in just the first one year after being rated A. Morgan Stanley is a higher than average risk for an A-rated company as it was forced to borrow as much as \$61.3 billion from the Federal Reserve on September 29, 2008.⁹ “Through the cycle” has a very simple meaning—it is a very long-term default probability that is totally consistent with the *term structure of default probabilities* of a well-specified model. What is the term? The major rating agencies are currently reporting about 30 years of historical experience, so the answer is 30 years.

VALUATION, PRICING, AND HEDGING

Earlier in this entry, we said the best one-sentence test of a credit model is “what is the hedge?” That statement is no exaggeration, because in order to be able to specify the hedge, we need to be able to value the risky credit (or portfolio of risky credits). If we can value the credits, we can price them as well. If we can value them, we can stress test that valuation as macroeconomic factors driving default probabilities shift. The pervasive impact of macroeconomic factors on default probabilities Figure 1 shows for Bank of America and Citigroup makes obvious what is documented by van Deventer and Imai (2003). The business cycle drives default risk (and valuations) up and down. With this valuation capability, we can meet one of the key objectives specified in this entry: We know the true value of everything we own and everything Wall Street wants us to buy or sell. We can see that the structured product offered at 103 is in reality only worth 98. This capability is essential to meet modern risk management

standards. Just as important, it is critical insurance against becoming yet another victim of Wall Street.

EMPIRICAL DATA ON CREDIT SPREADS AND COMMON STOCK PRICES

Before exploring the nature and performance of modern credit models, it is useful to look at the relationship between stock prices and credit spreads. Van Deventer and Imai (2003) print in its entirety a useful data series of new issue credit spreads compiled over a nine-year period beginning in the mid-1980s by First Interstate Bancorp. First Interstate at the time was the seventh largest bank holding company in the United States, one of the largest debt issuers in the United States, and a company whose rating ranged from AA to BB during the course of the data series. The credit spreads were the average credit spread quoted for a new issue of noncall debt of \$100 million by six investment banking firms, with the high and low quotations thrown out. Data were collected weekly for 427 weeks. No yield curve smoothing or secondary market bond prices were necessary to get the spreads, as the spreads themselves were the pricing quotation. These data, in the author's judgment, are much more reliable than the average credit default swap spread available since 2003 because of the extremely low volumes of credit default swap transactions reported by the Depository Trust and Clearing Corporation on www.dtcc.com.

Jarrow and van Deventer (1998, 1999) first used these data to test the implications of credit models. They reported the following findings on the relationship between credit spreads and equity prices:

- Stock prices and credit spreads moved in opposite directions during the week 172–184 times (depending on the maturity of the credit spread) of the 427 observations.

- Stock prices and credit spreads were both unchanged in only 1–3 observations.
- In total, only 40.7% to 43.6% of the observations were consistent with the Merton model (and literally any of its single factor variants) of risky debt.

This means that multiple variables are impacting credit spreads and stock prices, not the single variable (the value of company assets) that is the explanatory variable in any of the commercially available implementations of default probabilities that are Merton related. We address this issue in detail in our discussion of the Merton model¹⁰ and its variants in the following section. The summary data on the First Interstate stock price and credit spreads are reproduced in Table 1.

STRUCTURAL MODELS OF RISKY DEBT

Modern derivatives technology was the first place analysts turned in the mid-1970s as they sought to augment Altman's early work on corporate default prediction with an analytical model of default.¹¹ The original work in this regard was done by Black and Scholes (1973) and Merton (1974). This early work and almost all of the more recent extensions of it share a common framework:

- The assets of the firm are assumed to be perfectly liquid and are traded in efficient markets with no transactions costs.
- The amount of debt is set at time zero and does not vary.
- The value of the assets of the firm equal the sum of the equity value and the sum of the debt value, the original Modigliani and Miller assumptions.

All of the analysts using this framework conclude that the equity of the firm is some kind of option on the assets of the firm. An immediate implication of this is that one variable (except

Table 1 Analysis of Changes in First Interstate Bancorp Credit Spreads Stock Prices

	SPREAD 2 Years	SPREAD 3 Years	SPREAD 5 Years	SPREAD 7 Years	SPREAD 10 Years	Total
Total Number of Data Points	427	427	427	427	427	2135
Data Points Consistent with Merton						
Opposite Move in Stock Price and Spreads	179	178	183	172	184	896
Stock Price and Credit Spreads Unchanged	3	3	1	2	2	11
Total Consistent	182	181	184	174	186	907
Percent Consistent						
With Merton Model	42.6%	42.4%	43.1%	40.7%	43.6%	42.5%
Standard Deviation	2.4%	2.4%	2.4%	2.4%	2.4%	1.1%
Standard Deviations from 100% Consistency	-23.9	-24.1	-23.7	-24.9	-23.5	-53.8
Standard Deviations from 50% Consistency	-3.1	-3.2	-2.9	-3.9	-2.7	-7.0

Source: van Deventer and Imai (2003).

in the cases of random interest rates assumed below), the random value of company assets, completely determines stock prices, debt prices, and credit spreads. Except for the random interest rate versions of the model, this means that when the value of company assets rises, then stock prices should rise and credit spreads should fall. Table 1 rejects the hypothesis that this result is true by 23.5 to 24.9 standard deviations using the First Interstate data described earlier. In fact, as the First Interstate data show, stock prices and credit spreads move in the direction implied by various versions of the Merton model only 40.7% to 43.6% of the time. Van Deventer and Imai (2003) report on a similar analysis for a large number of companies with more than 20,000 observations and find similar results.

Given this inconsistency of actual market movements with the strongly restrictive assumption that only one variable drives debt and equity prices, why did analysts choose the structural models of risky debt in the first place? Originally, the models were implemented on the hope (and sometimes belief) that performance must be good. Later, once the performance of the model was found to be poor, this knowledge was known only to very large financial institutions who had an extensive credit model testing regime. One very large institution, for example, told the author in 2003 that it had known for years that the most popu-

lar commercial implementation of the Merton model of risky debt was less accurate than the market leverage ratio in the ordinal ranking of companies by riskiness. The firm was actively using this knowledge to arbitrage market participants who believed, but had not confirmed, that the Merton model of risky debt was accurate. We report on the large body of test results that began to enter the public domain in 1998 in a later section.

As analysts began to realize there were problems with the structural models of risky debt, active attempts were made to improve the model. We present in the following paragraphs a brief listing of the types of assumptions that can be used in the structural models of risky debt.¹²

Pure Black-Scholes/Merton Approach

The original Merton model assumes interest rates are constant and that equity is a European option on the assets of the firm. This means that bankruptcy can occur only at the maturity debt of the single debt instrument issued by the firm. Lando (2004, p. 14) notes a very important liability of the basic Merton model as the maturity of debt gets progressively shorter: "When the value of assets is larger than the face value of debt, the yield spreads go to zero as time

to maturity goes to 0 in the Merton model.” This is a critical handicap in trying to use this one-period model as a complete valuation framework. If credit spreads are unrealistic, we cannot achieve accuracy in our one-sentence credit model test: What’s the hedge?

We note here that allowing for various classes of debt is a very modest extension of the model. Allowing for subordinated debt does not change the probability of default. The implicit loss given default will simply be higher for the subordinated debt issue than it will for the senior debt issue.

Merton Model with Stochastic Interest Rates

The Merton model with stochastic interest rates was published by Shimko, Tejima, and van Deventer (1993). This modest extension of the original Merton framework simply combined Merton’s own model for options when interest rates are random with the structural credit risk framework. The model has the virtue of allowing two random factors (the risk-free short-term rate of interest and the value of company assets, which can have any arbitrary degree of correlation). It provides at least a partial explanation of the First Interstate results discussed above, but it shares most of the other liabilities of the basic Merton approach.

The Merton Model with Jumps in Asset Values

One of the most straightforward ways in which to make credit spreads more realistic is to assume that there are random jumps in the value of company assets, overlaid on top of the basic Merton assumption of geometric Brownian motion (i.e., normally distributed asset returns and lognormally distributed asset values). This model produces more realistic credit spread values, but Lando (2004, p. 27) concludes, “while the jump-diffusion model is

excellent for illustration and simulating the effects of jumps, the problems in estimating the model make it less attractive in practical risk management.”

Introducing Early Default in the Merton Structural Approach

In 1976, Black and Cox allowed default to occur prior to the maturity of debt if the value of company assets hits a deterministic barrier that can be a function of time. The value of equity is the equivalent of a “down and out” call option. When there are dividend payments, modeling gets much more complicated. Lando (2004, p. 33) summarizes key attributes of this modeling assumption: “While the existence of a default barrier increases the probability of default in a Black-Cox setting compared with that in a Merton setting, note that the bond holders actually take over the remaining assets when the boundary is hit and this in fact leads to higher bond prices and lower spreads.”

Other Variations on the Merton Model

Other extensions of the model summarized by Lando include

- A Merton model with continuous coupons and perpetual debt.
- Stochastic interest rates and jumps with barriers in the Merton model.
- Models of capital structure with stationary leverage ratios.

Ironically, all current commercial implementations of the Merton model for default probability estimation are minor variations on the original Merton model or extremely modest extensions of Black and Cox (1976). In short, at best 34-year-old technology is being used. Moreover, all current commercial implementations assume interest rates are constant, making failure of the “What’s the hedge test” a

certainty for fixed income portfolio managers, the primary users of default technology. All of the problems raised in the previous section on the First Interstate dataset remain for all current commercial implementations. That has much to do with the empirical results summarized below.

REDUCED-FORM MODELS OF RISKY DEBT

The many problems with the major variations on the Merton approach led Jarrow and Turnbull (1995) to elaborate on a reduced form of the original Merton model. In his options model for companies where the stock price is lognormally distributed, Merton allowed for a constant instantaneous default intensity. If the default event occurred, the stock price was assumed to go to zero. Merton derived the value of options on a defaultable common stock in a constant interest rates framework. Van Deventer (2006) shows how to use this Merton “reduced form” model to imply default probabilities from observable put and call options.

Jarrow and Turnbull adopted this default intensity approach as an alternative to the Merton structural approach. They did so under the increasingly popular belief that companies’ choices of capital structure vary dynamically with the credit quality of the firm, and that the assets they hold are often highly illiquid, contrary to the assumptions in the structural approach. Duffie and Singleton (1999), Jarrow (2001), and many others have dramatically increased the richness of the original Jarrow-Turnbull model to include the following features:

- Interest rates are random.
- An instantaneous default intensity is also random and driven by interest rates and one or more random macroeconomic factors.
- Bonds are traded in a less liquid market, and credit spreads have a “liquidity premium”

above and beyond the loss component of the credit spread.

- Loss given default can be random and driven by macroeconomic factors as well.

Default intensities and the full term structure of default probabilities can be derived in two ways:

- By implicit estimation, from observable bond prices, credit default swap prices, or options prices or any combination of them
- By explicit estimation, using a historical default database

The first commercial implementation on a sustained basis of the latter approach was the 2002 launch of the Kamakura Risk Information Services multiple models default probability service, which includes both Merton and reduced form models benchmarked in historical default data bases. The first commercial implementation of this approach for sovereign default risk assessment was also by Kamakura Risk Information Services in 2008.

In deriving default probabilities from historical data, financial economists have converged on a hazard rate modeling estimation procedure using logistic regression, where estimated default probabilities $P[t]$ are fitted to a historical database with both defaulting and nondefaulting observations and a list of explanatory variables X_i . Chava and Jarrow (2004) prove that the logistic regression is the maximum likelihood estimator when trying to predict a dependent variable that is either one (i.e., in the default case) or zero (in the “no default” case):

$$P[t] = 1/[1 + \exp(-\alpha - \sum_{i=1}^n \beta_i X_i)]$$

This simple equation makes obvious the most important virtue of the reduced form approach. The reduced form approach can employ any variable, without restriction, that improves the quality of default prediction, because any variable can contribute in the equation above including Merton default probabilities if they

have explanatory power. This means that the reduced form approach can never be worse than the Merton model because the Merton model can always be an input. The reverse is not true—the charge card balance of the chief executive officer is a well-known predictor of small business default, but the Merton default formulas do not have the flexibility to use this insight. Note also that the linear function in the denominator can be thought of as Altman’s 1968 z-score concept. In that sense, the reduced form/logistic regression approach has both Altman’s work and Merton’s work as ancestors.

In short, reduced form models can be the result of unconstrained variable selection among the full set of variables that add true economic explanatory power to default prediction. The Merton model, in any variation, is a constrained approach to default estimation because the mathematical formula for the model does not allow many potential explanatory variables to be used.

Most importantly, the logistic regression approach provides a solid opportunity to test whether in fact the Merton model does have the problems one would predict from the First Interstate data discussed above. We turn to that task now.

EMPIRICAL EVIDENCE ON MODEL PERFORMANCE

Shumway and Bharath (2008) conduct an extensive test of the Merton approach. They test two hypotheses. Hypothesis 1 is that the Merton model is a “sufficient statistic” for the probability of default, that is, a variable so powerful that in a logistic regression like the formula in the previous section no other explanatory variables add explanatory power. Hypothesis 2 is the hypothesis that the Merton model adds explanatory power even if common reduced form model explanatory variables are present. They specifically test modifications of the Merton structure partially disclosed by commercial

vendors of the Merton model. The Shumway and Bharath (2008) conclusions, based on all publicly traded firms in the United States (except financial firms) using quarterly data from 1980 to 2003 are as follows:¹³

- “We conclude that the . . . Merton model does not produce a sufficient statistic for the probability of default.”
- “Models 6 and 7 include a number of other covariates: the firm’s returns over the past year, the log of the firm’s debt, the inverse of the firm’s equity volatility, and the firm’s ratio of net income to total assets. Each of these predictors is statistically significant, making our rejection of hypothesis one quite robust. Interestingly, with all of these predictors included in the hazard model, the . . . Merton probability is no longer statistically significant, implying that we can reject hypothesis two.”
- “Looking at CDS implied default probability regressions and bond yield spread regressions, the . . . Merton probability does not appear to be a significant predictor of either quantity when our naïve probability, agency ratings and other explanatory variables are accounted for.”

These conclusions have been confirmed by Kamakura Corporation in five studies done in 2002, 2003, 2004, 2006, and 2011. The current Kamakura default database includes more than 1.76 million monthly observations on all public companies in North America from 1990 to December 2008, including 2,046 defaulting observations. Both hypotheses 1 and 2 were tested in the context of a “hybrid” model, which adds the Kamakura Merton implementation as an additional explanatory variable alongside the Kamakura reduced form model inputs. In every case, Kamakura agrees with Shumway and Bharath that hypothesis 1 can be strongly rejected. Kamakura has found 49 other variables that are statistically significant predictors of default even when Merton default probabilities are added as an explanatory variable.

Somewhat different from Shumway and Bharath, Kamakura finds that the Merton default probability has weak statistical significance when added as an explanatory variable to these other 49 variables, but the coefficient on the Merton default probability has the wrong sign; when Merton default probabilities rise, the predicted hybrid default probabilities fall. This is because Merton default probabilities are highly correlated with other variables like the market leverage ratio (which was mentioned above as out-predicting the commercial Merton implementation) and the ratio of total liabilities to total assets. It is an interesting econometric question whether the Merton input variable should be retained in such an event.

These findings were indirectly confirmed in Bohn, Arora, and Korablev (2005), in which Moody's for the first time releases quantitative test results on their Merton implementation. In that paper, the authors report on the relative accuracy of their proprietary Merton implementation compared to the more standard Merton theoretical implementation; they state that on a relatively easy data set (1996–2004 with small firms and financial institutions excluded) the proprietary Merton implementation has a receiver operating characteristics (ROC) accuracy ratio 7.5% higher than the standard Merton implementation.¹⁴ This puts the accuracy of the Moody's model more than 5% below that reported on a harder data set (all public firms of all sizes, including banks, 1990–2004) in the Kamakura Risk Information Services Technical Guide, Version 4.1 (2005) and again in the Kamakura Risk Information Services Guide, Version 5.0 (2010) on data spanning 1990–2008. The accuracy is also well below reduced form model accuracy published in Bharath and Shumway (2008), Campbell, Hilscher, and Szilagyi (2008), Hilscher and Wilson (2011), van Deventer and Imai (2003), and van Deventer, Imai, and Mesler (2004). The standard Merton accuracy ratio reported by Bohn, Arora, and Korablev (2005) is identical to that reported by Kamakura on a harder data set. It is

not surprising that there were no comparisons to reduced-form models using logistic regression in Bohn, Arora, and Korablev.

KEY POINTS

- Ratings date from the founding of a predecessor of Standard & Poor's in 1860. The very existence of ratings as a credit assessment tool dates from an era when computers did not exist and the electronic transmission of financial information was impossible.
- Because of this history, ratings are extremely simple ordinal rankings of firms or other counterparties by a small number of ratings grades, 21 grades in the case of the U.S. rating agencies.
- Ratings have no explicit maturity and no explicit default probability associated with them.
- For consumer credit risk assessment, "credit scores" are similar to ratings in that they are an ordinal risk measure, they have no maturity, and they have no explicit default probability associated with the score. While some credit bureaus state that credit scores rank the risk of a 90-day past due experience over 24 months, they are used on the full spectrum of credits from charge cards to 30-year mortgages.
- Unlike ratings, which have both qualitative and quantitative inputs to the process, the creation of credit scores is fully automated and based on a sophisticated statistical process.
- In the modern era, there is no need for either ratings or credit scores if the credit analyst has access to best in class default probabilities for a full term structure of time horizons for each counterparty.
- The ratings debate about "point in time" and "through the cycle" is a distinction without a difference. All ratings reflect information as of the ratings announcement date, as do default probabilities, so they are in that sense a "point in time." The longest term default probability

is the best measure of long-term risk and the shortest term default probability is the best measure of short-term risk. The longest default probability is “through the cycle” if the maturity is long enough. The maturity of the rating has never been clearly articulated by the rating agencies themselves.

- The first attempts at measuring default probabilities were based on the early work by Robert Merton nearly 40 years ago. Merton’s theory is simple and has intuitive appeal.
- The Merton model has not been accurate in practical use because it is based on assumptions that are simply not true: that common stock prices and bond prices are driven by only one factor, the value of company assets, and that company assets are perfectly liquid.
- A modern reduced form approach will always be more accurate than the Merton approach because the reduced form approach can employ any input that makes economic sense and improves accuracy.
- Logistic regression is the maximum likelihood estimator for prediction of a variable that has a zero (no default) or one (default) value.
- Reduced form default models were introduced by Jarrow based on an early continuous time default model by Merton. Empirical evidence suggests reduced form models are more accurate than ratings and the Merton approach in predicting default.
- Reduced form default models were first launched commercially in 2002 for public firms and in 2008 for sovereigns. They are also in wide use for predicting default of retail and small business clients.

NOTES

1. For a detailed discussion of the objectives of the credit risk modeling process, see van Deventer and Imai (2003).
2. *Financial Times*, April 8, 2008.
3. For evidence in this regard, see Bharath and Shumway (2008), Campbell, Hilscher, and Szilagyi (2008), and Hilscher and Wilson (2011).
4. The exact amounts, dates, and terms of borrowing are available at www.frb.gov.
5. Typically, the range of credit scores runs from 300 to 850 in the United States. There are differences by region and by vendor in the range used.
6. For examples, see Hilscher and Wilson (2011) and Kamakura Corporation press releases on March 15, 2006 and March 8, 2011.
7. The Kamakura Corporation troubled company index measures the percent of public firms that are “troubled,” defined as firms with annualized 1 month default risk over 1%. This index was only 6% in July 2011, and it was near 25% in October 2008.
8. See Table 24 in Standard & Poor’s (2011).
9. See “Case Studies in Liquidity Risk: Morgan Stanley,” Kamakura blog, www.kamakuraco.com, May 31, 2011.
10. See Merton (1974).
11. See Altman (1968).
12. For a summary of the extensions of the model, see Chapter 2 in Lando (2004).
13. Quotations are from an unpublished 2004 version of the paper, rather than the final 2008 published version, as some of the author’s insights were removed during the editorial process.
14. The difference is 15% on the equivalent cumulative accuracy profile basis, which is scaled from 0 to 100, compared to a 50–100 scale for the ROC accuracy ratio.

REFERENCES

- Altman, E. I. (1968). Financial bankruptcies, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance* 23, 589–609.
- Bharath, S., and Shumway, T. (2008). Forecasting default with the Merton distance to default model. *Review of Financial Studies* 21, 1339–1369.
- Black, F., and Cox, J. C. (1976). Valuing corporate securities: Some effects of bond indenture provisions. *Journal of Finance* 31: 351–367.

- Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.
- Bohn, J., Arora, N., and Korabev, I. (2005). *Power and Level Validation of the EDFtm Credit Measure in North America*. Moody's KMV memorandum, March 18.
- Campbell, J. Y., Hilscher, J., and Szilagyi, J. (2008). In search of distress risk. *Journal of Finance* 63, 2899–2939.
- Duffie, D., and Singleton, K. (1999). Modeling term structures of defaultable bonds. *Review of Financial Studies* 12: 197–226.
- Hilscher, J., and Wilson, M. (2011). *Credit Ratings and Credit Risk*. Working Paper, Brandeis University.
- Jarrow, R. A. (2001). Technical guide: Default probabilities implicit in debt and equity prices. *Kamakura Corporation Technical Guide*.
- Jarrow, R. A., Mesler, M., and van Deventer, D. R. (2006). *Kamakura Default Probabilities Technical Report, Kamakura Risk Information Services, Version 4.1*. Kamakura Corporation memorandum (January).
- Jarrow, R. A., Klein, S., Mesler, M., and van Deventer, D. R. (2010). *Kamakura Default Probabilities Technical Report, Kamakura Risk Information Services, Version 5.0*. Kamakura Corporation memorandum (September).
- Jarrow, R. A., and Turnbull, S. (1995). Pricing derivatives on financial securities subject to credit risk. *Journal of Finance* 50: 53–85.
- Jarrow, R. A., and van Deventer, D. R. (1998). Integrating interest rate risk and credit risk in asset and liability management. In *Asset and Liability Management: The Synthesis of New Methodologies*. London, UK: Risk Publications.
- Jarrow, R. A., and van Deventer, D. R. (1999). Practical usage of credit risk models in loan portfolio and counterparty exposure management: An update. In D. Shimko (ed.), *Credit Risk Models and Management*. London, UK: Risk Publications.
- Lando, D. (2004). *Credit Risk Modeling*. Princeton, NJ: Princeton University Press.
- Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29: 449–470.
- Shimko, D. C., Tejima, N., and van Deventer, D. R. (1993). The pricing of risky debt when interest rates are stochastic. *Journal of Fixed Income* 3, 58–66.
- Standard & Poor's. (2011). *Default, Transition and Recovery: 2010 Annual Global Default Study and Ratings Transitions*. March 30.
- van Deventer, D. R. (2006). Asset and liability management in enterprise wide risk management perspective. In M. Ong (ed.), *Risk Management: A Modern Perspective*. London: Elsevier Academic Press.
- van Deventer, D. R. and Imai, K. (2003). *Credit Risk Models and the Basel Accords*. Hoboken, NJ: John Wiley & Sons.
- van Deventer, D. R., Imai, K., and Mesler, M. (2004). *Advanced Financial Risk Management: Tools and Techniques for Integrated Credit Risk and Interest Rate Risk Management*. Hoboken, NJ: John Wiley & Sons.

Default Correlation in Intensity Models for Credit Risk Modeling

ABEL ELIZALDE, PhD

Credit Derivatives Strategy, J.P. Morgan

Abstract: The two primary types of credit risk models that seek to statistically describe default processes are the reduced-form model and the structural model. The most extended types of reduced-form models are the intensity models. There are three main approaches to incorporate credit risk correlation among firms within the framework of reduced models. The first approach, the conditionally independent defaults models, introduces credit risk dependence among firms through the dependence of the firms' default intensity processes on a common set of state variables. Contagion models extend the conditionally independent defaults approach to account for default clustering (periods in which the firms' credit risk is increased and in which the majority of the defaults take place). Finally, default dependencies can also be accounted for using copula functions. The copula approach takes as given the marginal default probabilities of the different firms and plugs them into a copula function, which provides the model with the default dependence structure.

There are two primary types of models in the literature that attempt to describe default processes for debt obligations and other defaultable financial instruments, usually referred to as structural and *reduced-form* (or *intensity*) models.

Structural models use the evolution of firms' structural variables, such as asset and debt values, to determine the time of default. Merton's model (1974) was the first modern model of default and is considered the first structural model. In Merton's model, a firm defaults if, at the time of servicing the debt, its assets are below its outstanding debt. A second approach within the structural framework was introduced by Black and Cox (1976). In this approach defaults occur as soon as a firm's asset value falls below a certain threshold. In con-

trast to the Merton approach, default can occur at any time.

Reduced-form models do not consider the relation between default and firm value in an explicit manner. Intensity models represent the most extended type of reduced-form models.¹ In contrast to structural models, the time of default in intensity models is not determined via the value of the firm, but it is the first jump of an exogenously given jump process. The parameters governing the default hazard rate are inferred from market data.

Structural default models provide a link between the credit quality of a firm and the firm's economic and financial conditions. Thus, defaults are endogenously generated within the model instead of exogenously given as in the

reduced approach. Another difference between the two approaches refers to the treatment of recovery rates: Whereas reduced models exogenously specify recovery rates, in structural models the value of the firm's assets and liabilities at default will determine recovery rates.

This entry focuses on the intensity approach, analyzing various models and reviewing the three main approaches to incorporate credit risk correlation among firms within the framework of reduced-form models.

PRELIMINARIES

In this section, we fix the information and probabilistic framework we need to develop the theory of reduced-form models. After presenting the basic features of reduced models and the motivation of the default intensity through Poisson processes, we apply these concepts to the specification of single firm default probabilities and to the valuation formulas for defaultable and default-free bonds. Finally, we analyze the different treatments the recovery rate has received in the literature.

Information Framework

For the purposes of this investigation, we shall always assume that economic uncertainty is modeled with the specification of a filtered probability space $\Pi = (\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbf{P})$, where Ω is the set of possible states of the economic world, and \mathbf{P} is a probability measure. The filtration (\mathcal{F}_t) represents the flow of information over time. $\mathcal{F} = \sigma(\bigcup_{t \geq 0} \mathcal{F}_t)$ is a σ -algebra, a family of events at which we can assign probabilities in a consistent way.² Before continuing with the exposition, let us make some remarks about the choice of the probability space.

First, we assume, as a starting point, that we can fix a unique physical or real probability measure $\bar{\mathbf{P}}$, and we consider the filtered probability space $\tilde{\Pi} = (\Omega, \mathcal{F}, (\mathcal{F}_t), \bar{\mathbf{P}})$. The choice of

the probability space will vary in some respects, according to the particular problems under consideration. In the rest of the entry, as we indicated above, we shall regularly make use of a probability measure \mathbf{P} , that will be assumed to be equivalent to $\bar{\mathbf{P}}$. The choice of \mathbf{P} then varies according to the context.

The model for the default-free term structure of interest rates is given by a non-negative, bounded and (\mathcal{F}_t) -adapted default-free short-rate process r_t . The money market account value process is given by:

$$\beta_t = \exp\left(\int_0^t r_s ds\right) \quad (1)$$

For our purposes we shall use the class of equivalent probability measures \mathbf{P} , where non-dividend-paying (NDP) asset processes discounted by the money market account are $((\mathcal{F}_t), \mathbf{P})$ -martingales, that is, where \mathbf{P} is an equivalent probability measure that uses the money market account as numeraire.³ Such an equivalent measure is called a risk neutral measure, because under this probability measure the investors are indifferent between investing in the money market account or in any other asset. There are different scenarios under which the transition from the physical to the equivalent (or risk neutral) probability measure can usually be accomplished.

We present a mathematical framework that will embody essentially all models used throughout this entry. Nevertheless, more general frameworks can be considered. On our probability space Π we assume that there exists an R^J -valued Markov process $X_t = (X_{1,t}, \dots, X_{J,t})'$ or background process, that represents J economy-wide variables, either state (observable) or latent (not observable).⁴ There also exist I counting processes, $N_{i,t}$, $i = 1, \dots, I$, initialized at 0, that represent the default processes of the I firms in the economy such that the default of the i th firm occurs when $N_{i,t}$ jumps from 0 to 1. $(\mathcal{G}_{X,t})$ and $(\mathcal{G}_{i,t})$, where $\mathcal{G}_{X,t} = \sigma(X_s, 0 \leq s \leq t)$ and $\mathcal{G}_{i,t} = \sigma(N_{i,t}, 0 \leq s \leq t)$, represent the filtrations

generated by X_t and $N_{i,t}$ respectively. The filtration $(\mathcal{G}_{X,t})$ represents information about the development of general market variables and all the background information, whereas $(\mathcal{G}_{i,t})$ only contains information about the default status of firm i .

The filtration (\mathcal{F}_t) contains the information generated by both the information contained in the state variables and the default processes:

$$(\mathcal{F}_t) = (\mathcal{G}_{X,t}) \vee (\mathcal{G}_{1,t}) \vee \dots \vee (\mathcal{G}_{I,t}) \quad (2)$$

We also define the filtrations $(\mathcal{F}_{i,t})$, $i = 1, \dots, I$, as

$$(\mathcal{F}_{i,t}) = (\mathcal{G}_{X,t}) \vee (\mathcal{G}_{i,t}) \quad (3)$$

which only accumulate the information generated by the state variables and the default status of each firm.

Poisson and Cox Processes

Poisson processes provide a convenient way of modeling default arrival risk in intensity-based default risk models.⁵ In contrast to structural models, the time of default in intensity models is not determined via the value of the firm, but instead is taken to be the first jump of a point process (for example, a Poisson process). The parameters governing the default intensity (associated with the probability measure \mathbf{P}) are inferred from market data.

First, we recall the formal definition of Poisson and Cox processes. Consider an increasing sequence of stopping times $(\tau_h < \tau_{h+1})$. We define a counting process associated with that sequence as a stochastic process N_t given by

$$N_t = \sum_h \mathbf{1}_{\{\tau_h \leq t\}} \quad (4)$$

A (homogeneous) Poisson process with intensity $\lambda > 0$ is a counting process whose increments are independent and satisfy

$$\mathbf{P}[N_t - N_s = n] = \frac{1}{n!} (t - s)^n \lambda^n \exp(-(t - s)\lambda) \quad (5)$$

for $0 \leq s \leq t$, that is, the increments $N_t - N_s$ are independent and have a Poisson distribution with parameter $\lambda(t - s)$ for $s \leq t$.

So far, we have considered only the case of homogeneous Poisson processes where the default intensity is a constant parameter λ , but we can easily generalize it allowing the default intensity to be time dependent $\lambda_t = \lambda(t)$, in which case we would talk about unhomogeneous Poisson processes.

If we consider stochastic default intensities, the Poisson process would be called a Cox process. For example, we can assume λ_t follows a diffusion process of the form

$$d\lambda_t = \mu(t, \lambda_t)dt + \sigma(t, \lambda_t)dW_t \quad (6)$$

where W_t is a Brownian motion. We can also assume that the intensity is a function of a set of state variables (economic variables, interest rates, currencies, etc.) X_t , that is, $\lambda_t = \lambda(t, X_t)$.

The fundamental idea of the intensity-based framework is to model the default time as the first jump of a Poisson process. Thus, we define the default time to be

$$\tau = \inf \{t \in \mathbb{R}^+ \mid N_t > 0\} \quad (7)$$

The survival probabilities in this setup are given by

$$\mathbf{P}[N_t = 0] = \mathbf{P}[\tau > t] = E \left[\exp \left(- \int_0^t \lambda_s ds \right) \right] \quad (8)$$

The intensity, or hazard rate, is the conditional default arrival rate, given no default:

$$\lim_{h \rightarrow 0} \frac{\mathbf{P}[\tau \in (t, t+h] \mid \tau > t]}{h} = \frac{f(t)}{1 - F(t)} = \lambda_t \quad (9)$$

where

$$F(t) = \mathbf{P}[\tau \leq t] \quad (10)$$

and $f(t)$ is the density of F .

The functions $p(t, T) = \mathbf{P}[\tau \leq T \mid \tau > t]$ and $s(t, T) = \mathbf{P}[\tau > T \mid \tau > t]$ are the risk neutral default and survival probabilities from time t to time T respectively, where $0 \leq t \leq T$. Note

that $s(t, T) = 1 - p(t, T)$, and if we fix $t = 0$, then $p(0, T) = F(T)$.

The hazard or intensity rate λ_t is the central element of reduced form models, and represents the instantaneous default probability, that is, the (very) short-term default risk.

Pricing Building Blocks

This section reviews the pricing of risk-free and defaultable zero-coupon bonds, which together with the default/survival probabilities constitute the building blocks for pricing credit derivatives and defaultable instruments.

We assume a perfect and arbitrage-free capital market, where the money market account value process β_t is given by (1). Since our probability measure \mathbf{P} takes as numeraire the money market account process, the value of any NDP-asset discounted by the money market account follows an $((\mathcal{F}_t), \mathbf{P})$ -martingale. Using the previous property, the price at time t of a default-free zero coupon bond with maturity T and face value of 1 unit is given by

$$\begin{aligned} P(t, T) &= \beta_t E \left[\frac{P(T, T)}{\beta_T} \mid \mathcal{F}_t \right] \\ &= E \left[\exp \left(- \int_t^T r_s ds \right) \mid \mathcal{F}_t \right] \end{aligned} \quad (11)$$

From the previous section we know that the survival probability $s(t, T)$ in the risk-neutral measure can be expressed as

$$\begin{aligned} s(t, T) &= \mathbf{P}[\tau > T \mid \tau > t] \\ &= E \left[\exp \left(- \int_t^T \lambda_s ds \right) \mid \mathcal{F}_t \right] \end{aligned} \quad (12)$$

Consider a defaultable zero coupon bond issued by firm i with maturity T and face value of M units that, in case of default at time $\tau < T$, generates a recovery payment of R_τ units. R_t is an (\mathcal{F}_t) -adapted stochastic process, with $R_t = 0$ for all $t > T$.⁶ The price of the defaultable coupon bond at time t , ($0 \leq t \leq T$), is given

by

$$\begin{aligned} Q(t, T) &= \beta_t E \left[\frac{Q(T, T)}{\beta_T} \mid \mathcal{F}_t \right] \\ &= \beta_t E \left[\frac{M \mathbf{1}_{\{\tau > T\}}}{\beta_T} \mid \mathcal{F}_t \right] + \beta_t E \left[\frac{R_\tau}{\beta_\tau} \mid \mathcal{F}_t \right] \end{aligned} \quad (13)$$

which can be expressed as⁷

$$\begin{aligned} Q(t, T) &= E \left[\exp \left(- \int_t^T (r_s + \lambda_s) ds \right) M \mid \mathcal{F}_t \right] + \\ &E \left[\int_t^T R_s \lambda_s \exp \left(- \int_t^s (r_u + \lambda_u) du \right) ds \mid \mathcal{F}_t \right] \end{aligned} \quad (14)$$

assuming $\tau > t$ and all the technical conditions that ensure that the expectations are finite.⁸ This expression has to be evaluated considering the treatment of the recovery payment and any other assumptions about the correlations between interest rates, intensities, and recoveries. The first term represents the expected discounted value of the payment of M units at time T , taking into account the possibility that the firm may default and the M units not received, through the inclusion of the hazard or intensity rate (instantaneous probability of default) in the discount rate. The second term represents the expected discounted value of the recovery payment using the risk-free rate plus the intensity rate as discount factor. The first integral in the second term of the previous expression, from t to T , makes reference to the fact that default can happen at any time between t and T . Thus, for each $s \in (t, T]$, we discount the value of the recovery rate R_s times the instantaneous probability of default at time s given that no default has occurred before, which is given by the intensity λ_s .

Recovery Rates

Recovery rates refer to how we model, after a firm defaults, the value that a debt instrument has left.⁹ In terms of the recovery rate parametrization, three main specifications have

been adopted in the literature. The first one considers that the recovery rate is an exogenous fraction of the face value of the defaultable bond (recovery of face value, RFV).¹⁰ Jarrow and Turnbull (1997) consider the recovery rate to be an exogenous fraction of the value of an equivalent default-free bond (recovery of treasury, RT). Finally, Duffie and Singleton (1999a) fix a recovery rate equal to an exogenous fraction of the market value of the bond just before default (recovery of market value, RMV).

The RMV specification has gained a great deal of attention in the literature thanks to, among others, Duffie and Singleton (1999a). Consider a zero-coupon defaultable bond, which pays M at maturity T if there is no default prior to maturity and whose payoff in case of default is modeled according to the RMV assumption. They show that this bond can be priced as if it were a default-free zero-coupon bond, by replacing the usual short-term interest rate process r_t with a default-adjusted short rate process $\pi_t = r_t + \lambda_t L_t$. L_t is the expected loss rate in the market value if default were to occur at time t , conditional on the information available up to time t :

$$R_\tau = (1 - L_\tau)Q(\tau-, T) \tag{15}$$

$$Q(\tau-, T) = \lim_{\substack{s \rightarrow \tau \\ s \leq \tau}} Q(s, T) \tag{16}$$

where τ is the default time, $Q(\tau-, T)$ the market price of the bond just before default, and R_τ the value of the defaulted bond. Duffie and Singleton (1999a) show that (14) can be expressed as

$$Q(t, T) = E \left[\exp \left(- \int_t^T \pi_s ds \right) M \mid \mathcal{F}_t \right] \tag{17}$$

This expression shows that discounting at the adjusted rate π_t accounts for both the probability and the timing of default, and for the effect of losses at default. But the main advantage of the previous pricing formula is that, if the mean loss rate $\lambda_t L_t$ does not depend on the value of the defaultable bond, we can apply well-known term structure processes to model π_t instead of r_t to

price defaultable debt. One of the main drawbacks of this approach is that since $\lambda_t L_t$ appears multiplied in π_t , in order to be able to estimate λ_t and L_t separately using data of defaultable instruments, it is not enough to know defaultable bond prices alone. We would need to have available a collection of bonds that share some, but not all, default characteristics, or derivative securities whose payoffs depend, in different ways, on λ_t and L_t . In case λ_t and L_t are not separable, we shall have to model the product $\lambda_t L_t$ (which represents the short-term credit spread).¹¹ This identification problem is the reason why most of the empirical work that tries to estimate the default intensity process from defaultable bond data uses an exogenously given constant, that is, $L_t = L$ for all t , recovery rate.¹²

The previous valuation formula allows one to introduce dependencies between short-term interest rates, default intensities, and recovery rates (via state variables, for example).

From a pricing point of view, the above pricing formula allows us to include the case in which, as is often seen in practice after a default takes place, a firm reorganizes itself and continues with its activity. If we assume that after each possible default the firm is reorganized and the bondholders lose a fraction L_t of the predefault bond's market value, Giesecke (2002a) shows that letting L_t be a constant, that is, $L_t = L$ for all t , the price of a default risky zero-coupon bond is, as in the case with no reorganization, given by (17).

Another advantage of this framework is that it allows one to consider liquidity risk by introducing a stochastic process l_t as a liquidity spread in the adjusted discount process π_t ; that is, $\pi_t = r_t + \lambda_t L_t + l_t$.

SINGLE ENTITY

The aim of this section is to develop some tools in the modeling of intensity processes, in order to build the models for *default correlation*. In case we consider a deterministic specification

for default intensities, it is natural to think of time dependent intensities, in which $\lambda_t = \lambda(t)$, where $\lambda(t)$ is usually modeled as either a constant, linear, or quadratic polynomial of the time to maturity.¹³

The treatment of default-free interest rates, the recovery rate, and the intensity process differentiates each intensity model.

It is interesting to note that the difference between the pricing formulas of default-free zero-coupon bonds and survival probabilities in the intensity approach lies in the discount rate:

$$P(0, t) = E \left[\exp \left(- \int_0^t r_s ds \right) \right] \quad (18)$$

$$s(0, t) = E \left[\exp \left(- \int_0^t \lambda_s ds \right) \right] \quad (19)$$

This analogy between intensity-based default risk models and interest rate models allows us to apply well-known short-rate term models to the modeling of default intensities.

Schönbucher (2003) enumerates several characteristics that an ideal specification of the interest rate r_t and the default intensity λ_t should have. First, both r_t and λ_t should be stochastic. Second, the dynamics of r_t and λ_t should be rich enough to include correlation between them. Third, it is desirable to have processes for r_t and λ_t that remain positive at all times. And finally, the easier the pricing of the pricing building blocks, the better.

We start with a general framework, making use of the Markov process $X_t = (X_{1,t}, \dots, X_{J,t})'$, which represents J state variables. The most general process for X_t that we shall consider is called a basic affine process, which is an example of an affine jump diffusion given by

$$dX_{j,t} = \kappa_j (\theta_j - X_{j,t}) dt + \sigma_j \sqrt{X_{j,t}} dW_{j,t} + dq_{j,t} \quad (20)$$

for $j = 1, \dots, J$, where $W_{j,t}$ is an $((\mathcal{F}_t), \mathbf{P})$ -Brownian motion. κ_j and θ_j represent the mean reversion rate and level of the process, and σ_j is a constant affecting the volatility of the process. $dq_{j,t}$ denotes any jump that occurs at time t of

a pure-jump process $q_{j,t}$, independent of $W_{j,t}$, whose jump sizes are exponentially distributed with mean μ_j and whose jump times are independent Poisson random variables with intensity of arrival γ_j (jump times and jump sizes are also independent). By modeling the jump size as an exponential random variable, we restrict the jumps to be positive. This process is called a basic affine process with parameters $(\kappa_j, \theta_j, \sigma_j, \mu_j, \gamma_j)$.¹⁴

Making r_t and λ_t dependent on a set of common stochastic factors X_t , one can introduce randomness and correlation in the processes of r_t and λ_t . Moreover, if we use basic affine processes for the common factors X_t , we can make use of the following results, which will yield closed-form solutions for the building blocks we examined in the previous section.¹⁵

1. For any discount-rate function $\varphi : R^J \rightarrow R$ and any function $g : R^J \rightarrow R$, if X_t is a Markov process (which holds in the case of basic affine process), then

$$E \left[\exp \left(- \int_s^t \varphi(X_u) du \right) g(X_t) \mid \mathcal{F}_s \right] = H(X_s) \quad (21)$$

for $0 \leq s \leq t$ and for some function $H : R^J \rightarrow R$.

2. Defining an affine function as constant plus linear, if $\varphi(x)$ and $g(x)$ are affine functions ($\varphi(x) = a_0 + a_1 x_1 + \dots + a_J x_J$ and $g(x) = b_0 + b_1 x_1 + \dots + b_J x_J$) then, as shown by Duffie, Pan, and Singleton (2000), if X_t is an affine jump-diffusion process, it is verified that $H(X_s)$ can be expressed in closed form by

$$H(X_s) = \exp(\alpha(s, t) + \theta(s, t) \cdot X_s) \quad (22)$$

for some coefficients $\alpha(s, t), \theta_1(s, t), \dots, \theta_J(s, t)$ which are functions, also in closed form, of the parameters of the model.¹⁶

Observing that our pricing building blocks $P(t, T)$, $s(t, T)$ and $Q(t, T)$ are special cases of the previous expressions, one realizes the gains in terms of tractability achieved by the use of

affine processes in the modelling of the default term structure.¹⁷ In order to make use of this tractability the state variables X_t should follow affine processes, and the specification for the risk-adjusted rate π_t should be an affine function of the state variables.

Consider the case in which the $X_{1,t}, \dots, X_{J,t}$ follow (20). If we eliminate the jump component from the process of $X_{j,t}$

$$dX_{j,t} = \kappa_j (\theta_j - X_{j,t}) dt + \sigma_j \sqrt{X_{j,t}} dW_{j,t} \quad (23)$$

we obtain the CIR process, and eliminating the square root of $X_{j,t}$

$$dX_{j,t} = \kappa_j (\theta_j - X_{j,t}) dt + \sigma_j dW_{j,t} \quad (24)$$

we end up with a Vasicek model.

Various reduced-form models differ from each other in their choices of the state variables and the processes they follow. In the models we consider below, the intensity and interest rate are linear, and therefore affine, functions of X_t , where X_t are basic affine processes.¹⁸

One can consider expressions for r_t and λ_t of the general form

$$r_t = a_{0,r}(t) + a_{1,r}(t) X_{1,t} + \dots + a_{J,r}(t) X_{J,t} \quad (25)$$

$$\lambda_t = a_{0,\lambda}(t) + a_{1,\lambda}(t) X_{1,t} + \dots + a_{J,\lambda}(t) X_{J,t} \quad (26)$$

for some deterministic (possibly time-dependent) coefficients $a_{j,r}$ and $a_{j,\lambda}$, $j = 1, \dots, J$. This type of model allows us to treat r_t and λ_t as stochastic processes, to introduce correlations between them, and to have analytically tractable expressions for pricing the building blocks. A simple example could be

$$dr_t = \kappa_r (\theta_r - r_t) dt + \sigma_r \sqrt{r_t} dW_{r,t} \quad (27)$$

$$d\lambda_t = \kappa_\lambda (\theta_\lambda - \lambda_t) dt + \sigma_\lambda \sqrt{\lambda_t} dW_{\lambda,t} + dq_{\lambda,t} \quad (28)$$

$$dW_{r,t} dW_{\lambda,t} = \rho dt \quad (29)$$

in which the state variables are r_t and λ_t themselves, whose Brownian motions are correlated.

Duffie (2005) presents an extensive review of the use of affine processes for credit risk modeling using intensity models, and applies such models to price different credit derivatives (credit default swaps, credit guarantees, spread options, lines of credit, and ratings-based step-up bonds.)

Default Times Simulation

Letting U be a uniform (0,1) random variable independent of $(\mathcal{G}_{X,t})$, the time of default is defined by

$$\tau = \inf \left\{ t > 0 \mid \exp \left(- \int_0^t \lambda_s ds \right) \leq U \right\} \quad (30)$$

Equivalently, we can let η be an exponentially distributed random variable with parameter 1 and independent of $(\mathcal{G}_{X,t})$ and define the default time as

$$\tau = \inf \left\{ t > 0 \mid \int_0^t \lambda_s ds \geq \eta \right\} \quad (31)$$

Once we have specified and calibrated the dynamics of λ_t , we can easily simulate default times using a simple procedure based on the two previous definitions. First, we simulate a realization u of a uniform $[0, 1]$ random variable U and choose τ such that $\exp(-\int_0^\tau \lambda_s ds) = u$. Equally, we can simulate a random variable η exponentially distributed with parameter 1 and fix τ such that $\int_0^\tau \lambda_s ds = \eta$.

DEFAULT CORRELATION

This section reviews the different approaches to model the default dependence between firms in the reduced-form approach. With the tools provided in the previous section we can calculate the survival or default probability of a given firm in a given time interval. The next natural question to ask ourselves concerns the default or survival probability of more than one firm. If we are currently at time t ($0 \leq t \leq T$) and no default has occurred so far, what is the probability

that $n \geq 1$ different firms default before time T ? or, what is the probability that they all survive until time T ?

Schönbucher (2003), again, points out some properties that any good approach to model dependent defaults should verify. First, the model must be able to produce default correlations of a realistic magnitude. Second, it has to do it by keeping the number of parameters introduced to describe the dependence structure under control, without growing dramatically with the number of firms. Third, it should be a dynamic model, able to model the number of defaults as well as the timing of defaults. Fourth, since it is clear from the default history that there are periods in which defaults may cluster, the model should be capable of reproducing these periods. And fifth, the easier the calibration and implementation of the model, the better.

We can distinguish three different approaches to model default correlation in the literature of intensity credit risk modeling. The first approach introduces correlation in the firms' default intensities making them dependent on a set of common variables X_t and on a firm specific factor. These models have received the name of conditionally independent defaults (CID) models, because conditioned to the realization of the state variables X_t , the firm's default intensities are independent as are the default times that they generate. Apparently, the main drawback of these models is that they do not generate sufficiently high default correlations. However, Yu (2002a) indicates that this is not a problem of the model per se, but rather an indication of the lack of sophistication in the choice of the state variables.

Two direct extensions of the CID approach try to introduce more default correlation in the models. One is the possibility of joint jumps in the default intensities (Duffie and Singleton 1999b) and the other is the possibility of default-event triggers that cause joint defaults (Duffie and Singleton 1999b, Kijima 2000, and Kijima and Muromachi 2000).

The second approach to model default correlation, *contagion models*, relies on the works by Davis and Lo (1999) and Jarrow and Yu (2001). It is based on the idea of default contagion in which, when a firm defaults, the default intensities of related firms jump upwards. In these models default dependencies arise from direct links between firms. The default of one firm increases the default probabilities of related firms, which might even trigger the default of some of them.

The last approach to model default correlation makes use of *copula functions*. A copula is a function that links univariate marginal distributions to the joint multivariate distribution with auxiliary correlating variables. To estimate a joint probability distribution of default times, we can start by estimating the marginal probability distributions of individual defaults, and then transform these marginal estimates into the joint distribution using a copula function. Copula functions take as inputs the individual probabilities and transform them into joint probabilities, such that the dependence structure is completely introduced by the copula.

Measures of Default Correlation

The complete specification of the default correlation will be given by the joint distribution of default times. Nevertheless, we can specify some other measures of default correlation. Consider two firms A and B that have not defaulted before time t ($0 \leq t \leq T$). We denote the probabilities that firms A and B will default in the time interval $[t, T]$ by p_A and p_B respectively. Denote p_{AB} the probability of both firms defaulting before T , and τ_A and τ_B the default times of each firm. The linear correlation coefficient between the default indicator random variables $\mathbf{1}_A = \mathbf{1}_{\{\tau_A \leq T\}}$ and $\mathbf{1}_B = \mathbf{1}_{\{\tau_B \leq T\}}$ is given by

$$\rho(\mathbf{1}_{\{\tau_A \leq T\}}, \mathbf{1}_{\{\tau_B \leq T\}}) = \frac{p_{AB} - p_A p_B}{\sqrt{p_A(1-p_A)p_B(1-p_B)}} \quad (32)$$

In the same way we can define the linear correlation of the random variables $\mathbf{1}_{\{\tau_A > T\}}$ and $\mathbf{1}_{\{\tau_B > T\}}$. Another measure of default dependence between firms is the linear correlation between the random variables τ_A and τ_B , $\rho(\tau_A, \tau_B)$.

The conclusions extracted from the comparison of linear default correlations should be viewed with caution because they are covariance-based and hence they are only the natural dependence measures for joint elliptical random variables.¹⁹ Default times, default events, and survival events are not elliptical random variables, and hence these measures can lead to severe misinterpretations of the true default correlation structure.²⁰

The previous correlation coefficients, when estimated via a risk neutral intensity model, are based on the risk neutral measure. However, when we calculate the correlation coefficients using empirical default events, the correlation coefficients are obtained under the physical measure. Jarrow, Lando, and Yu (2001) and Yu (2002a, b) provide a procedure for computing physical default correlation through the use of risk neutral intensities.

Conditionally Independent Default Models

From now on, we consider $i = 1, \dots, I$ different firms and denote by $\lambda_{i,t}$ and τ_i their default intensities and default times respectively.

In CID models, firms' default intensities are independent once we fix the realization of the state variables X_t . The default correlation is introduced through the dependence of each firm's intensity on the random vector X_t . A firm-specific factor of stochasticity $\lambda_{i,t}^*$, independent across firms, completes the specification of each firm's default intensity:

$$\lambda_{i,t} = a_{0,\lambda_i} + a_{1,\lambda_i} X_{1,t} + \dots + a_{J,\lambda_i} X_{J,t} + \lambda_{i,t}^* \tag{33}$$

where a_{j,λ_i} are some deterministic coefficients, for $j = 1, \dots, J$ and $i = 1, \dots, I$.²¹

Since default times are continuously distributed, this specification implies that the probability of having two or more simultaneous defaults is zero.

Let us consider an example of a CID model based on Duffee (1999). The default-free interest rate is given by

$$r_t = a_{r,0} + X_{1,t} + X_{2,t} \tag{34}$$

where $a_{r,0}$ is a constant coefficient, and $X_{1,t}$ and $X_{2,t}$ are two latent factors (unobservable, interpreted as the slope and level of the default-free yield curve). After having estimated the latent factors $X_{1,t}$ and $X_{2,t}$ from default-free bond data, Duffee (1999) uses them to model the intensity process of each firm i as

$$\lambda_{i,t} = a_{0,\lambda_i} + a_{1,\lambda_i} (X_{1,t} - \bar{X}_1) + a_{2,\lambda_i} (X_{2,t} - \bar{X}_2) + \lambda_{i,t}^* \tag{35}$$

$$d\lambda_{i,t}^* = \kappa_i (\theta_i - \lambda_{i,t}^*) dt + \sigma_i \sqrt{\lambda_{i,t}^*} dW_{i,t} \tag{36}$$

where $W_{1,t}, \dots, W_{I,t}$ are independent Brownian motions, a_{0,λ_i} , a_{1,λ_i} , and a_{2,λ_i} are constant coefficients, and \bar{X}_1 and \bar{X}_2 are the sample means of $X_{1,t}$ and $X_{2,t}$.

The intensity of each firm i depends on the two common latent factors $X_{1,t}$ and $X_{2,t}$, and on an idiosyncratic factor $\lambda_{i,t}^*$, independent across firms. The coefficients a_{0,λ_i} , a_{1,λ_i} , a_{2,λ_i} , κ_i , θ_i and σ_i are different for each firm. In Duffee's model, $\lambda_{i,t}^*$ captures the stochasticity of intensities and the coefficients a_{1,λ_i} and a_{2,λ_i} , $i = 1, \dots, I$, capture the correlations between intensities themselves, and between intensities and interest rates.

Duffee (1999), Zhang (2003), Driessen (2005), and Elizalde (2005b) propose, and estimate, different CID models.²²

The literature on credit risk correlation has criticized the CID approach, arguing that it generates low levels of default correlation when compared with empirical default correlations. However, Yu (2002a) suggests that this apparent low correlation is not a problem of the approach itself but a problem of the choice of state or latent variables, owing to the

inability of a limited set of state variables to fully capture the dynamics of changes in default intensities. In order to achieve the level of correlation seen in empirical data, a CID model must include among the state variables the evolution of the stock market, corporate and default-free bond markets, as well as various industry factors.

According to Yu, the problem of low correlation in Duffee's model may arise because of the insufficient specification of the common factor structure, which may not capture all the sources of common variation in the model, leaving them to the idiosyncratic component, which in turn would not be independent across firms. In fact, Duffee finds that idiosyncratic factors are statistically significant and correlated across firms. As long as the firms' credit risk depend on common factors different from the interest rate factors, Duffee's specification is not able to capture all the correlation between firms' default probabilities. Xie, Wu, and Shi (2004) estimate Duffee's model for a sample of U.S. corporate bonds and perform a careful analysis of the model pricing errors. A principal component analysis reveals that the first factor explains more than 90% of the variation of pricing errors. Regressing bond pricing errors with respect to several macroeconomic variables, they find that returns on the S&P 500 index explain around 30% of their variations. Therefore, Duffee's model leaves out some important aggregate factors that affect all bonds.

Driessen (2005) proposes a model in which the firms' hazard rates are a linear function of two common factors, two factors derived from the term structure of interest rates, a firm idiosyncratic factor, and a liquidity factor. Yu also examines the model of Driessen (2005), finding that the inclusion of two new common factors elevates the default correlation.

Finally, Elizalde (2005b) shows that any firm's credit risk is, to a very large extent, driven by common risk factors affecting all firms. The study decomposes the credit risk of a sample of corporate bonds (14 U.S. firms, 2001–2003)

into different unobservable risk factors. A single common factor accounts for more than 50% of all (but two) of the firms' credit risk levels, with an average of 68% across firms. Such factor represents the credit risk levels underlying the U.S. economy and is strongly correlated with main U.S. stock indexes. When three common factors are considered (two of them coming from the term structure of interest rates), the model explains an average of 72% of the firms' credit risk.²³

Default Times Simulation

In the CID approach, to simulate default times we proceed as we did in the single entity case. Once we know the realization of the state variables X_t , we simulate a set of I independent unit exponential random variables η_1, \dots, η_I , which are also independent of $(\mathcal{G}_{X,t})$. The default time of each firm $i = 1, \dots, I$ is defined by

$$\tau_i = \inf \left\{ t > 0 \mid \int_0^t \lambda_{i,s} ds \geq \eta_i \right\} \quad (37)$$

Thus, once we have simulated η_i , τ_i will be such that

$$\int_0^{\tau_i} \lambda_{i,s} ds = \eta_i \quad (38)$$

Joint Jumps/Joint Defaults Duffie and Singleton (1999b) proposed two ways out of the low correlation problem. One is the possibility of joint jumps in the default intensities, and the other is the possibility of default-event triggers that cause joint defaults.²⁴

Duffie and Singleton develop an approach in which firms experience correlated jumps in their default intensities. Assume that the default intensity of each firm follows the following process:

$$d\lambda_{i,t} = \kappa_i (\theta_i - \lambda_{i,t}) dt + dq_{i,t} \quad (39)$$

which consists of a deterministic mean reversion process plus a pure jump process $q_{i,t}$, whose intensity of arrival is distributed as a Poisson random variable with parameter γ_i and

whose jump size follows an exponential random variable with mean μ (equal for all firms $i = 1, \dots, I$).²⁵ Duffie and Singleton introduce correlation to the firm's jump processes, keeping unchanged the characteristics of the individual intensities. They postulate that each firm's jump component consists of two kinds of jumps, joint jumps and idiosyncratic jumps. The joint jump process has a Poisson intensity γ_c and an exponentially distributed size with mean μ . Individual default intensities experience a joint jump with probability p_i . That is, a firm suffers a joint jump with Poisson intensity of arrival of $p_i\gamma_c$. In order to keep the total jump in each firm's default intensity with intensity of arrival γ_i and size μ , the idiosyncratic jump (independent across firms) is set to have an exponentially distributed size μ and intensity of arrival h_i , such that $\gamma_i = p_i\gamma_c + h_i$.

Note that if $p_i = 0$ the jumps are only idiosyncratic jumps, implying that default intensities and hence default times are independent across firms. If $p_i = 1$ and $h_i = 0$ all firms have the same jump intensity, which does not mean that default times are perfectly correlated, since the size of the jump is independent across firms. Only if we additionally assume that μ goes to infinity we obtain identical default times.

The second alternative considers the possibility of simultaneous defaults triggered by common credit events, at which several obligors can default with positive probability. Imagine there exist $m = 1, \dots, M$ common credit events, each one modeled as a Poisson process with intensity $\lambda_{m,t}^c$. Given the occurrence of a credit event m at time t , each firm i defaults with probability $p_{i,m,t}$. If, given the occurrence of a common shock, the firm's default probability is less than one, this common shock is called nonfatal shock, whereas if this probability is one, the common shock is called fatal shock. In addition to the common credit events, each entity can experience default through an idiosyncratic Poisson process with intensity $\lambda_{i,t}^*$, which is independent across firms. Therefore, the total

intensity of firm i is given by

$$\lambda_{i,t} = \lambda_{i,t}^* + \sum_{m=1}^M p_{i,m,t} \lambda_{m,t}^c \quad (40)$$

Consider a simplified version of this setting with two firms, constant idiosyncratic intensities λ_1^* and λ_2^* , and one common and fatal event with constant intensity λ^c . In this case firm i 's survival probability is given by

$$s_i(t, T) = \exp\left(-\left(\lambda_i^* + \lambda^c\right)(T - t)\right) \quad (41)$$

Denoting by $s(t; T_1, T_2)$ the joint survival probability, given no default until time t , that firm 1 does not default before time T_1 and firm 2 does not default before time T_2 , then

$$\begin{aligned} s(t; T_1, T_2) &= \exp(-\lambda_1^*(T_1 - t) - \lambda_2^*(T_2 - t) \\ &\quad - \lambda^c \max\{T_1 - t, T_2 - t\}) = \\ &= \exp(-(\lambda_1^* + \lambda^c)(T_1 - t) \\ &\quad - (\lambda_2^* + \lambda^c)(T_2 - t) + \lambda^c \\ &\quad \min\{T_1 - t, T_2 - t\}) \end{aligned} \quad (42)$$

which can be expressed as

$$\begin{aligned} s(t; T_1, T_2) &= s_1(t, T) s_2(t, T) \\ &\quad \min\{\exp(\lambda^c(T_1 - t)), \exp(\lambda^c(T_2 - t))\} \end{aligned} \quad (43)$$

This expression for the joint survival probability explicitly includes individual survival probabilities and a term that introduces the dependence structure. This is the approach followed by copula functions, which couple marginal probabilities into joint probabilities. In fact, the above example is a special case of copula function, called Marshall-Olkin copula.

The relationship between joint survival and default probabilities is given by

$$\begin{aligned} s(t; T_1, T_2) &= 1 - p_1(t, T_1) - p_2(t, T_2) \\ &\quad + p(t; T_1, T_2) \end{aligned} \quad (44)$$

where $p(t; T_1, T_2)$ represents the joint default probability, given no default until time t , that firm 1 defaults before time T_1 and firm 2

defaults before time T_2 . Obviously the case with multiple common shocks is more troublesome in terms of notation and calibration because, for every possible common credit event, an intensity must be specified and calibrated.²⁶

Duffie and Singleton (1999b) propose algorithms to simulate default times within these two frameworks. The criticisms that the joint credit event approach has received stem from the fact that it is unrealistic that several firms default at exactly the same time, and also from the fact that after a common credit event that makes some obligors default, the intensity of other related obligors that do not default does not change at all.

Although theoretically appealing, the main drawback of these two last models has to do with their calibration and implementation. To the best of my knowledge there is not a single paper that carries out an empirical calibration and implementation of a model like the ones presented in this section. The same applies to the contagion models presented in the next section.

Contagion Mechanisms

Contagion models take CID models one step further, introducing into the model two empirical facts: that the default of one firm can trigger the default of other related firms and that default times tend to concentrate in certain periods of time, in which the default probability of all firms is increased. The last model examined in the previous section (joint credit events) differs from contagion mechanisms in that if an obligor does not experience a default, its intensity does not change due to the default of any related obligor. The literature of default contagion includes two approaches: the infectious defaults model of Davis and Lo (1999), and the model proposed by Jarrow and Yu (2001), which we shall refer to as the propensity model. The main issues to be resolved concerning these two models are associated with difficulties in their calibration to market prices.

The Davis-Lo Infectious Defaults Model

The model developed by Davis and Lo (1999) has two versions, a static version that only considers the number of defaults in a given time period,²⁷ and a dynamic version in which the timing of default is also incorporated.²⁸

In the dynamic version of the model, each firm has an initial hazard rate of $\lambda_{i,t}$, for $i = 1, \dots, I$, which can be constant, time dependent, or follow a CID model. When a default occurs, the default intensity of all remaining firms is increased by a factor $a > 1$, called the enhancement factor, to $a\lambda_{i,t}$. This augmented intensity remains for an exponentially distributed period of time, after which the enhancement factor disappears ($a = 1$). During the period of augmented intensity, the default probabilities of all firms increase, reflecting the risk of default contagion.

The Jarrow-Yu Propensity Model

In order to account for the clustering of default in specific periods, Jarrow and Yu (2001) extend CID models to account for counterparty risk: the risk that the default of a firm may increase the default probability of other firms with which it has commercial or financial relationships. This allows them to introduce extra-default dependence in CID models to account for *default clustering*. In a first attempt, Jarrow and Yu assume that the default intensity of a firm depends on the status (default/not default) of the rest of the firms (symmetric dependence). However, symmetric dependence introduces a circularity in the model, which they refer to as looping defaults, which makes it extremely difficult and troublesome to construct and derive the joint distribution of default times.

Jarrow and Yu restrict the structure of the model to avoid the problem of looping defaults. They distinguish between primary firms ($1, \dots, K$) and secondary firms ($K + 1, \dots, I$). First, they derive the default intensity of primary firms, using a CID model. The primary

firm intensities $\lambda_{1,t}, \dots, \lambda_{K,t}$ are $(\mathcal{G}_{X,t})$ -adapted and do not depend on the default status of any other firm. If a primary firm defaults, this increases the default intensities of secondary firms, but not the other way around (asymmetric dependence). Thus, secondary firms' default intensities are given by

$$\lambda_{i,t} = \hat{\lambda}_{i,t} + \sum_{j=1}^K a_{i,t}^j \mathbf{1}_{\{\tau_j \leq t\}} \quad (45)$$

for $i = K + 1, \dots, I$ and $j = 1, \dots, K$, where $\hat{\lambda}_{i,t}$ and $a_{i,t}^j$ are $(\mathcal{G}_{X,t})$ -adapted. $\hat{\lambda}_{i,t}$ represents the part of secondary firm i 's hazard rate independent of the default status of other firms.

Default intensities of primary firms $\lambda_{1,t}, \dots, \lambda_{K,t}$ are $(\mathcal{G}_{X,t})$ -adapted, whereas default intensities of secondary firms $\lambda_{K+1,t}, \dots, \lambda_{I,t}$ are adapted with respect to the filtration $(\mathcal{G}_{X,t}) \vee (\mathcal{G}_{1,t}) \vee \dots \vee (\mathcal{G}_{K,t})$.

This model introduces a new source of default correlation between secondary firms, and also between primary and secondary firms, but it does not solve the drawbacks of low correlation between primary firms, which CID models apparently imply, because the setting for primary firms is, after all, only a CID model.²⁹

Default Times Simulation First we simulate the default times for the primary firms exactly as in the case of CID. Then, we simulate a set of $I-K$ independent unit exponential random variables $\eta_{K+1}, \dots, \eta_I$ (independent of $(\mathcal{G}_{X,t}) \vee (\mathcal{G}_{1,t}) \vee \dots \vee (\mathcal{G}_{K,t})$), and define the default time of each secondary firm $i = K + 1, \dots, I$ as

$$\tau_i = \inf \left\{ t > 0 \mid \int_0^t \lambda_{i,s} ds \geq \eta_i \right\} \quad (46)$$

Copulas

In CID and contagion models the specification of the individual intensities includes all the default dependence structure between firms. In contrast, the copula approach separates individual default probabilities from the credit risk dependence structure. The copula function

takes as inputs the marginal probabilities and introduces the dependence structure to generate joint probabilities.

Copulas were introduced in 1959 and have been extensively applied to model, among others, survival data in areas such as actuarial science.³⁰

In the rest of this section we review copula theory and its use in the credit risk literature. To make notation simple, assume we are at time $t = 0$ and take $s_i(t)$ and $p_i(t)$ (or $F_i(t)$) to be the survival and default probabilities, respectively, of firm $i = 1, \dots, I$ from time 0 to time $t > 0$. Then

$$F_i(t) = \mathbf{P}[\tau_i \leq t] = 1 - s_i(t) = 1 - \mathbf{P}[\tau_i > t] \quad (47)$$

where τ_i denotes the default time of firm i .

A copula function transforms marginal probabilities into joint probabilities. In case we model default times, the joint default probability is given by

$$\begin{aligned} F(t_1, \dots, t_I) &= \mathbf{P}[\tau_1 \leq t_1, \dots, \tau_I \leq t_I] \\ &= \mathbf{C}^d(F_1(t_1), \dots, F_I(t_I)) \end{aligned} \quad (48)$$

and if we model survival times, the joint survival probability takes the form

$$\begin{aligned} s(t_1, \dots, t_I) &= \mathbf{P}[\tau_1 > t_1, \dots, \tau_I > t_I] \\ &= \mathbf{C}^s(s_1(t_1), \dots, s_I(t_I)) \end{aligned} \quad (49)$$

where \mathbf{C}^d and \mathbf{C}^s are two different copulas.³¹

The copula function takes as inputs the marginal probabilities without considering how we have derived them. Thus, the intensity approach is not the only framework with which we can use copula functions to model the default dependence structure between firms. Any other approach to model marginal default probabilities, such as the structural approach, can use copula theory to model joint probabilities.

Description

An intuitive definition of a copula function is as follows:³²

Copula Function A function $C : [0, 1]^I \rightarrow [0, 1]$ is a copula if there are uniform random variables U_1, \dots, U_I taking values in $[0, 1]$ such that C is their joint distribution function.

A copula function C has uniform marginal distributions, that is,

$$C(1, \dots, 1, u_i, 1, \dots, 1) = u_i \quad (50)$$

for all $i = 1, \dots, I$ and $u_i \in [0, 1]$.

This definition is used, for example, by Schönbucher (2003).³³ The copula function C is the joint distribution of a set of I uniform random variables U_1, \dots, U_I . Copula functions allow one to separate the modeling of the marginal distribution functions from the modeling of the dependence structure. The choice of the copula does not constrain the choice of the marginal distributions. Sklar (1959) showed that any multivariate distribution function F can be written in the form of a copula function. The following theorem is known as Sklar's theorem:

Sklar's Theorem Let Y_1, \dots, Y_I be random variables with marginal distribution functions F_1, \dots, F_I and joint distribution function F . Then there exists an I -dimensional copula C such that $F(y_1, \dots, y_I) = C(F_1(y_1), \dots, F_I(y_I))$ for all (y_1, \dots, y_I) in R^I . Moreover, if each F_i is continuous, then the copula C is unique.

We shall consider the default times of each firm τ_1, \dots, τ_I as the marginal random variables whose joint distribution function will be determined by a copula function. If Y is a random variable with distribution function F , then the random variable U , defined as $U = F(Y)$, is a uniform $[0, 1]$ random variable. Denoting by t_i the realization of each τ_i ,³⁴

$$\begin{aligned} F(t_1, \dots, t_I) &= \mathbf{P}[\tau_1 \leq t_1, \dots, \tau_I \leq t_I] \\ &= C(F_1(t_1), \dots, F_I(t_I)) \end{aligned} \quad (51)$$

The marginal distribution of the default time τ_i will be given by

$$\begin{aligned} F_i(t_i) &= F(\infty, \dots, \infty, t_i, \infty, \dots, \infty) \\ &= \mathbf{P}[\tau_1 \leq \infty, \dots, \tau_i \leq t_i, \dots, \tau_I \leq \infty] = \\ &= C(F_1(\infty), \dots, F_i(t_i), \dots, F_I(\infty)) \\ &= C(1, \dots, F_i(t_i), \dots, 1) \end{aligned} \quad (52)$$

In the bivariate case, the relationship between the copula C^d and the survival copula C^s , which satisfies $s(t_1, t_2) = C^s(s_1(t_1), s_2(t_2))$, is given by³⁵

$$\begin{aligned} C^s(u_1, u_2) &= u_1 + u_2 - 1 \\ &+ C^d(1 - u_1, 1 - u_2) \end{aligned} \quad (53)$$

Nelsen (1999) points out that C^s is a copula and that it couples the joint survival function $s(\cdot, \dots, \cdot)$ to its univariate margins $s_1(\cdot), \dots, s_I(\cdot)$ in a manner completely analogous to the way in which a copula connects the joint distribution function $F(\cdot, \dots, \cdot)$ to its margins $F_1(\cdot), \dots, F_I(\cdot)$. When modeling credit risk using the copula framework we can specify a copula for either the default times or the survival times.

Measures of the Dependence Structure The dependence between the marginal distributions linked by a copula is characterized entirely by the choice of the copula. If C_1 and C_2 are two I -dimensional copula functions we say that C_1 is smaller than C_2 , denoted by $C_1 < C_2$, if $C_1(u) \leq C_2(u)$ for all $u \in [0, 1]^I$.

The Fréchet-Hoeffding copulas, C^- and C^+ , are two reference copulas given by³⁶

$$C^- = \max\{u_1 + \dots + u_I + 1 - I, 0\} \quad (54)$$

$$C^+ = \min\{u_1, \dots, u_I\} \quad (55)$$

satisfying $C^- < C < C^+$ for any copula C . However, this is a partial ordering in the sense that not every pair of copulas can be compared in this way.

In order to compare any two copulas, we would need an index to measure the dependence structure between two random variables introduced by the choice of the copula function. Linear (Pearson) correlation coefficient ρ is the

most used measure of dependence; however, it harbors several drawbacks, which makes it not very suitable to compare copula functions. For example, linear correlation depends not only on the copula but also on the marginal distributions.

We focus on four dependence measures that depend only on the copula function, not in the marginal distributions: Kendall's tau, Spearman's rho, and upper/lower tail dependence coefficients.

First, we introduce the concept of concordance:

Concordance Let (y_1, y_2) and (\hat{y}_1, \hat{y}_2) be two observations from a vector (Y_1, Y_2) of continuous random variables. Then, (y_1, y_2) and (\hat{y}_1, \hat{y}_2) are said to be concordant if $(y_1 - \hat{y}_1)(y_2 - \hat{y}_2) > 0$ and discordant if $(y_1 - \hat{y}_1)(y_2 - \hat{y}_2) < 0$.

Kendall's tau and Spearman's rho are two measures of concordance:

Kendall's Tau Let (Y_1, Y_2) and (\hat{Y}_1, \hat{Y}_2) be IID random vectors of continuous random variables with the same joint distribution function given by the copula \mathbf{C} (and with marginals F_1 and F_2). Then, Kendall's tau of the vector (Y_1, Y_2) (and thus of the copula \mathbf{C}) is defined as the probability of concordance minus the probability of discordance; that is,

$$\tau = \mathbf{P}[(Y_1 - \hat{Y}_1)(Y_2 - \hat{Y}_2) > 0] - \mathbf{P}[(Y_1 - \hat{Y}_1)(Y_2 - \hat{Y}_2) < 0] \quad (56)$$

Spearman's Rho Let (Y_1, Y_2) , (\hat{Y}_1, \hat{Y}_2) and $(\hat{\hat{Y}}_1, \hat{\hat{Y}}_2)$ be IID random vectors of continuous random variables with the same joint distribution function given by the copula \mathbf{C} (and with marginals F_1 and F_2). Then, Spearman's rho of the vector (Y_1, Y_2) (and thus of the copula \mathbf{C}) is defined as

$$\rho_S = 3(\mathbf{P}[(Y_1 - \hat{Y}_1)(Y_2 - \hat{\hat{Y}}_2) > 0] - \mathbf{P}[(Y_1 - \hat{Y}_1)(Y_2 - \hat{\hat{Y}}_2) < 0]) \quad (57)$$

Both Kendall's tau and Spearman's rho³⁷ take values in the interval $[0, 1]$ and can be defined in terms of the copula function by

$$\tau = 4 \iint_{[0,1]^2} \mathbf{C}(u, v) d\mathbf{C}(u, v) - 1 \quad (58)$$

$$\rho_S = 12 \iint_{[0,1]^2} uv d\mathbf{C}(u, v) - 3 = 12 \iint_{[0,1]^2} \mathbf{C}(u, v) dudv - 3 \quad (59)$$

The Fréchet-Hoeffding copulas take the two extreme values of Kendall's tau and Spearman's rho: If the copula of the vector (Y_1, Y_2) is \mathbf{C}^- then $\tau = \rho_S = -1$, and if it has copula \mathbf{C}^+ then $\tau = \rho_S = 1$. The product copula \mathbf{C}^P represents independent random variables, that is, if Y_1, \dots, Y_I are independent random variables, their copula is given by \mathbf{C}^P , such that $\mathbf{C}^P(u_1, \dots, u_I) = u_1 \dots u_I$. For a vector (Y_1, Y_2) of independent random variables, $\tau = \rho_S = 0$. Kendall's tau and Spearman's rho are equal for a given copula \mathbf{C} and its associated survival copula \mathbf{C}^s .

Kendall's tau and Spearman's rho are measures of global dependence. In contrast, tail dependence coefficients between two random variables (Y_1, Y_2) are local measures of dependence, as they refer to the level of dependence between extreme values, that is, values at the tails of the distributions $F_1(Y_1)$ and $F_2(Y_2)$.

Tail Dependence Let (Y_1, Y_2) be a random vector of continuous random variables with copula \mathbf{C} (and with marginals F_1 and F_2). Then, the coefficient of upper tail dependence of the vector (Y_1, Y_2) (and thus of the copula \mathbf{C}) is defined as

$$\lambda_U = \lim_{u \nearrow 1} \mathbf{P}[Y_1 > F_1^{-1}(u) \mid Y_2 > F_2^{-1}(u)] \quad (60)$$

where F_i^{-1} represents the inverse function of F_i , provided the limit exists. We say that the random vector (and thus the copula \mathbf{C}) has upper tail dependence if $\lambda_U > 0$. Similarly, the coefficient of lower tail dependence of the vector (Y_1, Y_2) (and thus

of the copula \mathbf{C} is defined as

$$\lambda_L = \lim_{u \searrow 0} \mathbf{P} [Y_1 < F_1^{-1}(u) \mid Y_2 < F_2^{-1}(u)] \quad (61)$$

We say that the random vector (and thus the copula \mathbf{C}) has lower tail dependence if $\lambda_L > 0$.

Upper (lower) tail dependence measures the probability that one component of the vector (Y_1, Y_2) is extremely large (small) given that the other is extremely large (small). As in the case of Kendall's tau and Spearman's rho, tail dependence is a copula property and can be expressed as³⁸

$$\lambda_U = \lim_{u \nearrow 1} \frac{1 + \mathbf{C}(u, u) - 2u}{1 - u} \quad (62)$$

$$\lambda_L = \lim_{u \searrow 0} \frac{\mathbf{C}(u, u)}{u} \quad (63)$$

The upper (lower) coefficient of tail dependence of the copula \mathbf{C} is the lower (upper) coefficient of tail dependence of its associated survival copula \mathbf{C}^s .

Consider the random vector (τ_1, τ_2) of default times for two firms. The coefficient of upper (lower) tail dependence represents the probability of long-term survival (immediate joint death). The existence of default clustering periods implies that a copula to model joint default (survival) probabilities should have lower (upper) tail dependence to capture those periods.

Examples of Copulas Here, we review some of the most used copulas in default risk modeling. The first two copulas, normal and Student t copulas, belong to the elliptical family of copulas. We also present the class of Archimedean copulas and the Marshall-Olkin copula.³⁹

1. Elliptical Copulas The I -dimensional normal copula with correlation matrix Σ is given by

$$\mathbf{C}(u_1, \dots, u_I) = \Phi_{\Sigma}^I(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_I)) \quad (64)$$

where Φ_{Σ}^I represents an I -dimensional normal distribution function with covariance matrix Σ ,

and Φ^{-1} denotes the inverse of the univariate standard normal distribution function.

Normal copulas are radially symmetric ($\lambda_U = \lambda_L$), tail independent ($\lambda_U = \lambda_L = 0$), and their concordance order depends on the linear correlation parameter ρ :

$$\mathbf{C}^- < \mathbf{C}_{\rho=-1} < \mathbf{C}_{\rho<0} < \mathbf{C}_{\rho=0} = \mathbf{C}^P < \mathbf{C}_{\rho>0} < \mathbf{C}_{\rho=1} = \mathbf{C}^+ \quad (65)$$

As with any other copula, the normal copula allows the use of any marginal distribution. We can express the linear correlation coefficients for a normal copula (ρ) in terms of both Kendall's tau (τ) and Spearman's rho (ρ_S) in the following way:

$$\rho = 2 \sin\left(\frac{\pi}{6} \rho_S\right) = \sin\left(\frac{\pi}{2} \tau\right) \quad (66)$$

Another elliptical copula is the t -copula. Letting X be a random vector distributed as an I -dimensional multivariate t -student with v degrees of freedom, mean vector μ (for $v > 1$) and covariance matrix $\frac{v}{v-2} \Sigma$ (for $v > 2$), we can express X as

$$X = \mu + \frac{\sqrt{v}}{\sqrt{S}} Z \quad (67)$$

where S is a random variable distributed as an χ^2 with v degrees of freedom and Z is an I -dimensional normal random vector, independent of S , with zero mean and linear correlation matrix Σ . The I -dimensional t -copula of X can be expressed as

$$\mathbf{C}(u_1, \dots, u_I) = t_{v,R}^I(t_v^{-1}(u_1), \dots, t_v^{-1}(u_I)) \quad (68)$$

where $t_{v,R}^I$ represents the distribution function of $\frac{\sqrt{v}}{\sqrt{S}} Y$, where Y is an I -dimensional normal random vector, independent of S , with zero mean and covariance matrix R . t_v^{-1} denotes the inverse of the univariate t -student distribution function with v degrees of freedom and $R_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$.

The t -copula is radially symmetric and exhibits tail dependence given by

$$\lambda_U = \lambda_L = 2 - 2t_{v+1} \left(\frac{(v+1)(1-\rho)}{(1+\rho)} \right)^{\frac{1}{2}} \quad (69)$$

where ρ is the linear correlation of the bivariate t-distribution.

2. Archimedean Copulas An I -dimensional Archimedean copula function \mathbf{C} is represented by

$$\mathbf{C}(u_1, \dots, u_I) = \phi^{-1}(\phi(u_1) + \dots + \phi(u_I)) \tag{70}$$

where the function $\phi : [0, 1] \rightarrow R^+$, called the generator of the copula, is invertible and satisfies $\phi'(u) < 0$, $\phi''(u) > 0$, $\phi(1) = 0$, $\phi(0) = \infty$. An Archimedean copula is entirely characterized by its generator function. Relevant Archimedean copulas are the Clayton, Frank, Gumbel, and Product copulas, whose generator functions are given by:

Copula	Generator $\phi(u)$
Clayton	$\frac{u^{-\theta}-1}{\theta}$ for $\theta \geq 0$
Frank	$-\ln \frac{e^{-\theta u}-1}{e^{-\theta}-1}$ for $\theta \in R \setminus \{0\}$
Gumbel	$(-\ln u)^\theta$ for $\theta \geq 1$
Product	$-\ln u$

The Clayton copula has lower tail dependence but not upper tail dependence. The Gumbel copula has upper tail dependence but not lower tail dependence. The Frank copula does not exhibit either upper or lower tail dependence.

Archimedean copulas allow for a great variety of different dependence structures, and the ones presented above are especially interesting because they are one-parameter copulas. In particular, the larger the parameter θ (in absolute value), the stronger the dependence structure. The Clayton, Frank, and Gumbel copulas are ordered in θ (i.e., $\mathbf{C}_{\theta_1} < \mathbf{C}_{\theta_2}$ for all $\theta_1 \leq \theta_2$). Unlike the Gumbel copula, which does not allow for negative dependence, Clayton and Frank copulas are able to model continuously the whole range of dependence between the lower Fréchet-Hoeffding copula, the product copula and the upper Fréchet-Hoeffding copula. Copulas with this property are called inclusive or comprehensive copulas. Frank copulas are the

only radially symmetric Archimedean copulas ($\mathbf{C} = \mathbf{C}^s$).

For Archimedean copulas, tail dependence and Kendall's tau coefficients can be expressed in terms of the generator function

$$\tau = 1 + 4 \int_0^1 \frac{\phi(u)}{\phi'(u)} du \tag{71}$$

$$\lambda_U = 2 - 2 \lim_{u \rightarrow 0} \frac{\phi^{-1}'(2u)}{\phi^{-1}'(u)} \tag{72}$$

$$\lambda_L = 2 \lim_{u \rightarrow \infty} \frac{\phi^{-1}'(2u)}{\phi^{-1}'(u)} \tag{73}$$

provided the derivatives and limits exist.

Archimedean copulas are interchangeable, which means that the dependence between any two (or more) random variables does not depend on which random variables we choose. In terms of credit risk analysis, this imposes an important restriction on the dependence structure since the default dependence introduced by an Archimedean copula is the same between any group of firms.

3. Marshall-Olkin Copula This copula was already mentioned when we dealt with joint defaults in intensity models. In its bivariate specification the Marshall-Olkin copula is given by

$$\begin{aligned} \mathbf{C}(u_1, u_2) &= \min \left\{ u_1^{1-\alpha_1} u_2, u_1 u_2^{1-\alpha_2} \right\} \\ &= u_1 u_2 \min \left\{ u_1^{-\alpha_1}, u_2^{-\alpha_2} \right\} \end{aligned} \tag{74}$$

for $\alpha_1, \alpha_2 \in (0, 1)$.⁴⁰

Copulas for Default Times

Within the reduced-form approach, we can distinguish two approaches to introduce default dependence using copulas. The first one, which we will refer to as Li's approach, was introduced by Li (1999) and represents one of the first attempts to use copula theory systematically in credit risk modeling. Li's approach takes as inputs the marginal default (survival) probabilities of each firm and derives the joint probabilities using a copula function.⁴¹ Although Li (1999) studies the case of a normal copula,

any other copula can be used within this framework.

If we are using a copula function as a joint distribution for default (survival) times, the simulated vector (u_1, \dots, u_I) of uniform $[0, 1]$ random variables from the copula will correspond to the default F_1, \dots, F_I (survival s_1, \dots, s_I) marginal distributions. Once we have simulated the vector (u_1, \dots, u_I) , we use it to derive the implied default times τ_1, \dots, τ_I such that $\tau_i = F_i^{-1}(u_i)$, or $\tau_i = s_i^{-1}(u_i)$ in the survival case, for $i = 1, \dots, I$.

The second approach was introduced by Schönbucher and Schubert (2001), and here we shall call it the Schönbucher-Schubert (SS) approach. In the algorithm to draw a default time in the case of a single firm, we simulated a realization u_i of a uniform $[0, 1]$ random variable U_i independent of $(\mathcal{G}_{X,t})$, and defined the time of default of firm i as τ_i such that

$$\exp\left(-\int_0^{\tau_i} \lambda_i(s) ds\right) = u_i \quad (75)$$

where λ_i is the default intensity process of firm i . The idea of the SS approach is to link the default thresholds U_1, \dots, U_I with a copula.

Schönbucher and Schubert consider that the processes $\lambda_1, \dots, \lambda_I$ are $(\mathcal{F}_{i,t})$ -adapted⁴² and call them pseudo default intensities. Thus, λ_i is the default intensity if investors only consider the information generated by the background filtration $(\mathcal{G}_{X,t})$ and by the default status of firm i , $(\mathcal{G}_{i,t})$. However, investors are not restricted to the information represented by $(\mathcal{F}_{i,t})$ as they also observe the default status of the rest of the firms. Therefore, λ_i is not the density of default with respect to all the information investors have available, as represented by (\mathcal{F}_t) , but rather with respect to a smaller information set.

To calculate the default (or survival) probabilities conditional to all the information that investors have available, (\mathcal{F}_t) , we cannot define those probabilities in terms of the pseudo default intensities $\lambda_1, \dots, \lambda_I$. We have to find the

“real” intensities implied by the investors’ information set. The difference between pseudo and real intensities lies in the fact that real intensities, in addition to all the information considered by pseudo intensities, include information about the default status of all firms. The default thresholds’ copula function includes this information in the SS approach.

In order to find the “real” default intensities h_1, \dots, h_I , which are (\mathcal{F}_t) -adapted, we need to combine both the pseudo default intensities and the copula function, which links the default thresholds. The pseudo default intensity λ_i includes information about the state variables and the default situation of firm i , and only coincides with the “real” default intensity h_i in cases of independent default or when the information of the market is restricted to $(\mathcal{F}_{i,t})$.⁴³

The simulation of the default times in this approach is exactly the same as in Li’s approach. The only difference with the SS approach is that it allows us to recover the dynamics of the “real” default intensities h_1, \dots, h_I , which include the default contagion effects implicit in the default threshold copula. In contrast to the models of Jarrow and Yu (2001) and Davis and Lo (1999), the SS approach allows the contagion effects to arise endogenously through the use of the copula.

Schönbucher (2003) calls the SS approach a dynamic approach in the sense that it considers the dynamics of the “real” default intensities h_1, \dots, h_I , as opposed to Li’s approach, which only considers the dynamics of the pseudo default intensities.

As Schönbucher and Schubert (2001) point out, this setup is very general, and the reader has freedom to choose the specification of the default intensities. We can introduce default correlation by both correlating the default intensities, for example with a CID model, and by using any of the copula approaches we have just presented.

In an extension of the SS approach, Rogge and Schönbucher (2003) propose not to use the normal or t -copulas but Archimedean copulas,

arguing that normal and t -copulas do not imply a realistic dynamic process for default intensities.

Galiani (2003) provides a detailed analysis of the use of copula functions to price multiname credit derivatives using both a normal and Student t copula.

Choosing and Calibrating the Copula

Once we have reviewed how to use copula theory in the context of joint default probabilities, we have to choose a copula and estimate its parameters. In order to choose a copula we should consider aspects such as the dependence structure each copula involves as well as the number of parameters we need to estimate.

Since the normal copula presents neither lower nor upper tail dependence, the use of multivariate normal distributions to model default (or price) behavior has been strongly criticized for not assigning enough probability to the occurrence of extreme events and, among them, the periods of default clustering. The use of the t -copula is the natural answer to the lack of tail dependence, since, subject to the degrees of freedom and covariance matrix, this copula exhibits tail dependence. The main problem in using a normal or t -copula is the number of parameters we have to estimate, which grows with the dimensionality of the copula.⁴⁴

Archimedean copulas are especially attractive because there exists a large number of one-parameter Archimedean copulas⁴⁵ which allows for a great variety of dependence structures. The disadvantage of Archimedean copulas is that they may impose too much dependence structure in the sense that, as they are interchangeable copulas, the dependence between any group of firms is the same independently of the firms we consider.

In case we decide to use an Archimedean copula, Genest and Rivest (1993) propose a procedure for identifying the Archimedean copula that best fits the data.⁴⁶ The problem is that they consider only the bivariate case and that,

as we shall see later, we need a sample of the marginal random variables (the random variables X_1, \dots, X_I whose marginal distributions we link to the copula function) that is available if we are modeling equity returns, but not if we are modeling default times. More generally, Fermanian and Scaillet (2004) discuss the issue of choosing the copula that best fits a given data set, using goodness-of-fit tests.

According to Durrleman, Nikeghbali, and Roncalli (2000):

There does not exist a systematic rigorous method for the choice of the copula: nothing can tell us that the selected family of copula will converge to the real structure dependence underlying the data. This can provide biased results since according to the dependence structure selected the obtained results might be different.

Jouanin et al. (2001) use the term model risk to denote this uncertainty in the choice of the copula.

Assuming we manage to select a copula function, we now face the estimation of its parameters. The main problem of the use of copula theory to model credit risk is the scarcity of default data from which to calibrate the copula.

We cannot rely on multiname credit derivatives, such as i^{th} -to-default products, to calibrate the copula because, in most cases, they are not publicly traded and also because of their lack of liquidity.

Imagine that, instead of fitting a copula to default times, we are fitting a copula to daily stock returns for I different firms. Let Y_1, \dots, Y_I be random variables denoting the daily returns of firms $i = 1, \dots, I$ with marginal distribution functions F_1, \dots, F_I and joint distribution function F . Sklar's theorem proves that there exists an I -dimensional copula C such that $F(y_1, \dots, y_I) = C(F_1(y_1), \dots, F_I(y_I))$ for all (y_1, \dots, y_I) in R^I . In this case, we have available, for each day, a sample of the random vector Y_1, \dots, Y_I that we can use to estimate the parameters of the copula. We would have to estimate the parameters of the marginal

distribution functions F_1, \dots, F_I and then estimate the parameters of the copula. Since, in our application to default times, we already have the marginal distributions, determined by the specification of the marginal default intensities, we are left with the estimation of the copula parameters. Providing we have a large sample of the random variables Y_1, \dots, Y_I , we can estimate the copula parameters in several ways.⁴⁷

If the copula is differentiable we can always use maximum likelihood to estimate the parameters.⁴⁸ De Matteis (2001) mentions that this parametric method may be convenient when we work with a large data set, but in case there are outliers or if the marginal distributions are heavy tailed, a nonparametric approach may be more suitable.

A nonparametric approach would involve the use of the sample version of a dependence measure, such as Kendall's tau or Spearman's rho (or both),⁴⁹ to calibrate the copula parameters. However, this nonparametric approach is restricted to the bivariate case, and we would need to have at least the same sample dependence measures as copula parameters.⁵⁰

The estimation methods exposed above rely on the availability of a large sample of the random variables Y_1, \dots, Y_I . However, this is not the case when we work with default times. We do not have available a large sample of default times for the I firms. In fact, we do not have a single realization of the default times random vector.

One solution is to assume that the marginal default (survival) probabilities and the marginal distributions of the equity returns share the same copula, that is, share the same dependence structure, and use equity returns to estimate the copula parameters. But this shortcut has its own drawbacks. We need to fit a copula to a set of given marginal distributions for the default (survival) times, which are characterized by a default intensity for each firm. Ideally we should estimate the parameters of the copula function using default times data. However, we rarely have enough default times data

available such as to properly estimate the parameters of the copula function. In those cases, we must rely on other data sources to calibrate the copula function. For example, a usual practice is to calibrate the copula using equity data of the different firms. However, the dependence of the firms' default probabilities will probably differ from the dependence in the evolution of their equity prices.

Another way of dealing with the estimation of the copula parameters is, as Jouanin et al. (2001) propose, through the use of "original methods that are based on the practice of the credit market rather than mimicking statistical methods that are never used by practitioners." They suggest a method based on Moody's diversity score.⁵¹ The diversity score or binomial expansion technique consists of transforming a portfolio of (credit dependent) defaultable bonds on an equivalent portfolio of uncorrelated and homogeneous credits assumed to mimic the default behavior of the original portfolio, using the so-called diversity score parameter, which depends on the degree of diversification of the original portfolio. We then match the first two moments of the number of defaults within a fixed time horizon for both the original and the transformed portfolio. Since the original portfolio assumes default dependence, the distribution of the number of defaults will depend on the copula parameters. In the transformed portfolio, that is, independent defaults, the number of defaults follows a binomial distribution with some probability p . Matching the first two moments of the number of defaults in both portfolios, we would extract an estimation for the probability p and for the copula parameters.⁵² However, Moody's diversity score approach has its own drawbacks. Among others, it is a static model with a fixed time horizon, that is, it does not consider when defaults take place but only the number of defaults within the fixed time horizon. In fact, the Committee on the Global Financial System (Bank for International Settlements) suggests, in its last report,⁵³ that diversity scores "are a fairly crude measure of

the degree of diversification in a portfolio of credits.”

Similarly to the choice of the copula function, there does not exist a rigorous method to estimate the parameters of the copula. We can talk about parameter risk which, together with the model risk mentioned earlier, are the principal problems we face if we use the copula approach in the modeling of dependent defaults.

KEY POINTS

- There are two primary types of models in the literature that attempt to describe default processes: structural and reduced-form models. Intensity models represent the most extended type of reduced-form models. In contrast to structural models, the time of default in intensity models is not determined via the value of the firm, but it is the first jump of an exogenously given jump process. The fundamental idea of the intensity-based framework is to model the default time as the first jump of a Poisson process. The default intensity of the Poisson process, also referred to as the hazard rate, can be deterministic (constant or time dependent) or stochastic.
- We review three different ways of introducing default correlations among firms in the framework of intensity models: the conditionally independent defaults (CID) approach, contagion models, and copula functions.
- CID models generate credit risk dependence among firms through the dependence of the firms' intensity processes on a common set of state variables. Firms' default rates are independent once we fix the realization of the state variables. Different CID models differ in their choices of the state variables and the processes they follow. Extensions of CID models introduce joint jumps in the firms' default processes or common default events.
- Contagion models extend the CID approach to account for the empirical observation of de-

fault clustering (periods in which firms' credit risk increases simultaneously and in which the majority of defaults take place). They are based on the idea that, when a firm defaults, the default intensities of related firms jump (upwards), that is, the default of one firm increases the default probabilities of other firms (to the point of potentially causing the default of some of them). These models include, on the specification of default intensities, the existence of contagion sources among firms, which can be explained by either their commercial/financial relationships or simply by their common exposure to the economy.

- In CID and contagion models the specification of the individual intensities includes all the default dependence structure between firms. In contrast, the copula approach separates individual default probabilities from the credit risk dependence structure.
- A copula is a function that links univariate marginal distributions to the joint multivariate distribution function. The copula approach takes as given the marginal default probabilities of the different firms and plugs them into a copula function, which provides the model with the dependence structure to generate joint default probabilities. This approach separates the modeling and estimation of the individual default probabilities, determined by the default intensity processes, from the modeling and calibration or estimation of the device that introduces the credit risk dependence, the copula.

NOTES

1. Brody, Hughston, and Macrina (2007) present an alternative reduced-form model based on the amount and precision of the information received by market participants about the firm's credit risk. Such a model does not require the use of default intensities; it belongs to the reduced-form approach because (like intensity models) it relies on market prices of defaultable

instruments as the only source of information about the firms' credit risk.

2. An event is a subset of Ω , namely a collection of possible outcomes. σ -algebras are models for information, and filtrations models for flows of information.
3. For some applications we may wish to enlarge this category of probability measures by relaxing the martingale condition to a local martingale condition, though this point does not concern us in what follows.
4. Given the filtered probability space Π , an (\mathcal{F}_t) -adapted process X_t is a Markov process with respect to (\mathcal{F}_t) if

$$E[f(X_t) | \mathcal{F}_s] = E[f(X_t) | X_s]$$

with probability one, for all $0 \leq s \leq t$, and for every bounded function f . This means that the conditional distribution at time s of X_t , given all available information, depends only on the current state X_s .

5. For a more detailed exposition see Lando (1994) and Chapter 5 in Schönbucher (2003).
6. This specification of the recovery rate incorporates all possible ways of dealing with recovery payments considered in the literature. Here we consider a continuous version of the recovery rate, that is, R_t is measured and received precisely at the default time. In the discrete version of the recovery rate, R_t is measured and received on the first date after default among a prespecified list $T_1 < \dots < T_n$ of times, where T_n is the maturity date T .
7. See Hughston and Turnbull (2001).
8. See proof on Lando (1994, Proposition 3.1) and Bielecki and Rutkowski (2002, Proposition 8.2).
9. For an extensive review of the treatment of recovery rates see Chapter 6 in Schönbucher (2003).
10. Houweling and Vorst (2001) consider the RFV specification for pricing credit default swaps.

11. See Duffie and Singleton (1999a) and Jarrow (1999).
12. There exist some empirical works that, under some specifications of λ_t and r_t , find that the value of the recovery rate does not substantially affect the results, as long as the recovery rate lies within a logical interval. See, for instance, Houweling and Vorst (2001) and Elizalde (2005a).
13. See Houweling and Vorst (2001) and Elizalde (2005a) for a comparison of different specifications of (deterministic) time-dependent intensity rates.
14. For a detailed description of affine processes see Duffie and Kan (1996), Duffie (1998), Duffie, Pan, and Singleton (2000), Duffie, Filipovic, and Schachermayer (2002), and Appendix A in Duffie and Singleton (2003). An affine jump-diffusion process is a jump-diffusion process for which the drift vector, instantaneous covariance matrix, and jump intensities all have affine dependence on the state vector. If X_t is a Markov process in some space state $D \subset \mathcal{R}^d$, X_t is an affine jump-diffusion if it can be expressed as

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t + dq_t$$

where W_t is an (\mathcal{F}_t) -Brownian motion in \mathcal{R}^d , $\mu: D \rightarrow \mathcal{R}^d$, $\sigma: D \rightarrow \mathcal{R}^d$ and q is a pure jump process whose jumps have a fixed probability distribution ν on \mathcal{R}^d and arrive with intensity $\{\psi(X_t): t \geq 0\}$, for some constant $\psi: D \rightarrow [0, \infty)$. That is, the drift vector μ , instantaneous covariance matrix $\sigma\sigma'$ and jump intensities ψ all have affine dependence on the state vector X_t . Intuitively this means that, conditional on the path of X_t , the jump times of q are the jump times of a Poisson process with time varying-intensity $\{\psi(X_t): t \geq 0\}$, and that the size of the jump at time T is independent of $\{X_s: 0 \leq s \leq t\}$ and has the probability distribution ν .

15. See Duffie and Kan (1996) and Duffie, Pan, and Singleton (2000).

16. For the basic affine model, the coefficients can be calculated explicitly. See Duffie and Garleanu (2001), Appendix A in Duffie and Singleton (2003), and Duffie, Pan, and Singleton (2000) for details and extensions.
17. Duffie, Pan, and Singleton (2000) developed a similar closed-form expression to the second term of the price of a defaultable zero-coupon bond $Q(t, T)$

$$E \left[\int_t^T R_s \lambda_s \exp \left(- \int_t^s (r_u + \lambda_u) du \right) ds \mid \mathcal{F}_t \right]$$

Using their expression, the pricing of defaultable zero-coupon bonds with constant recovery of face value reduces to the computation of a one-dimensional integral of a known function.

18. Several versions of the modeling of r_t and λ_t in this framework can be found in Duffie, Schroder, and Skidas (1996), Duffee (1999), Duffie and Singleton (1999, 2003), Kijima (2000), Kijima and Muromachi (2000), Duffie and Garleanu (2001), Bielecki and Rutkowski (2002), and Schönbucher (2003). For the estimation of an affine process intensity model without jumps see Duffee (1998) and Duffie, Pedersen, and Singleton (2003).
19. If Y is an n -dimensional random vector and, for some $\mu \in R^n$ and some $n \times n$ nonnegative definite, symmetric matrix Σ , the characteristic function $\psi_{Y-\mu}(t)$ of $Y - \mu$ is a function of the quadratic form $t^T \Sigma t$, $\psi_{Y-\mu}(t) = \phi(t^T \Sigma t)$. We say that Y has an elliptical distribution with parameters μ , Σ and ϕ . For example, normal and Student t distributions are elliptical distributions. For a more detailed treatment of elliptical distributions see Bingham and Kiesel (2002) and references cited therein.
20. See Embrechts, McNeal, and Straumann (1999) and Embrechts, Lindskog, and McNeil (2001).
21. We can always consider a model such as

$$d\lambda_{i,t} = \kappa_i (\theta_i - \lambda_{i,t}) dt + \sigma_i \sqrt{\lambda_{i,t}} dW_{i,t} + dq_{i,t}$$

- for each firm i , and introduce correlation via the Brownian motions $W_{1,t}, \dots, W_{I,t}$.
22. Duffee (1999), Driessen (2005), and Elizalde (2005b) use latent variables instead of state variables. Collin-Dufresne, Goldstein, and Martin (2001) show that financial and economic variables cannot explain the correlation structure of intensity processes. Latent factors are modeled as affine diffusions and estimated through a maximum likelihood procedure based on the Kalman filter.
 23. While Driessen (2005) considers that all firms with the same rating are affected in the same way by common factors, Elizalde (2005b) allows for the effect of each common factor to differ across firms, which increases the flexibility of the credit risk correlation structure.
 24. See also Kijima (2000), Kijima and Muromachi (2000), and Giesecke (2002b).
 25. This is a basic affine process with parameters $(\kappa_i, \theta_i, \sigma_i = 0, \mu, \gamma_i)$.
 26. See Embrechts, Lindskog, and McNeil (2001) and Giesecke (2002b).
 27. Extending the diversity score (or binomial expansion technique) of Moody's.
 28. This dynamic version is introduced in Davis and Lo (2001).
 29. See Yu (2002a) and Frey and Backhaus (2003) for an extension of the Jarrow and Yu (2001) model.
 30. See Sklar (1959) and Frees and Valdez (1998).
 31. Note that $F_i(t_i) = F(\infty, \dots, t_i, \dots, \infty)$ and $s_i(t_i) = s(0, \dots, t_i, \dots, 0)$.
 32. For a more detailed description of copula theory see Joe (1997), Frees and Valdez (1998), Nelsen (1999), Costinot, Roncalli, and Teiletche (2000), Embrechts, Lindskog, and McNeil (2001), De Matteis (2001), and Georges et al. (2001).
 33. A more formal definition would be the following (Frey, McNeil, and Nyfeler 2001): An I -dimensional copula \mathbf{C} is a function $\mathbf{C} : [0, 1]^I \rightarrow [0, 1]$ with the following properties:

- Grounded: For all $u \in [0, 1]^I$, $\mathbf{C}(u) = 0$ if at least one coordinate $u_j = 0$, $j = 1, \dots, I$.
- Reflective: If all coordinates of u are 1 except u_j then $\mathbf{C}(u) = u_j$, $j = 1, \dots, I$.
- I -increasing: The \mathbf{C} -volume of all hypercubes with vertices in $[0, 1]^I$ is positive, *i.e.*

$$\sum_{i_1=1}^2 \dots \sum_{i_I=1}^2 (-1)^{i_1+\dots+i_I} \mathbf{C}(u_{1,i_1}, \dots, u_{I,i_I}) \geq 0$$

for all $(u_{1,1}, \dots, u_{I,1})$ and $(u_{1,2}, \dots, u_{I,2})$ in $[0, 1]^I$ with $u_{j,1} \leq u_{j,2}$ for all $j = 1, \dots, I$.

34. We will use \mathbf{C}^d , or simply \mathbf{C} , to denote the copula function of default times and \mathbf{C}^s for the copula function of survival times.
35. See Georges et al. (2001) for a complete characterization of the relation between default and survival copulas.
36. \mathbf{C}^- is always a copula, but \mathbf{C}^+ is only a copula for $I \geq 3$.
37. A simple interpretation of Spearman's rho is the following. Let (Y_1, Y_2) be a random vector of continuous random variables with the same joint distribution function H (whose margins are F_1 and F_2) and copula C , and consider the random variables $U = F(Y_1)$ and $V = F(Y_2)$. Then, we can write the Spearman's rho coefficient of (Y_1, Y_2) as

$$\begin{aligned} \rho_S(Y_1, Y_2) &= 12 \iint_{[0,1]^2} \mathbf{C}(u, v) \, dudv - 3 \\ &= 12\mathbf{E}[UV] - 3 = \frac{\mathbf{E}[UV] - \frac{1}{4}}{\frac{1}{12}} = \\ &= \frac{\text{Cov}(U, V)}{\sqrt{\text{Var}(U)\text{Var}(V)}} = \rho(U, V) \\ &= \rho(F_1(Y_1), F_2(Y_2)) \end{aligned}$$

where ρ denotes the Pearson or linear correlation coefficient. So the Spearman's rho of the vector (Y_1, Y_2) is the Pearson correlation of the random variables $F_1(Y_1)$ and $F_2(Y_2)$.

38. Since $\mathbf{P}[Y_1 > F_1^{-1}(u) \mid Y_2 > F_2^{-1}(u)]$ can be written as

$$\frac{1 - \mathbf{P}[Y_1 \leq F_1^{-1}(u)] - \mathbf{P}[Y_2 \leq F_2^{-1}(u)] + \mathbf{P}[Y_1 \leq F_1^{-1}(u), Y_2 \leq F_2^{-1}(u)]}{1 - \mathbf{P}[Y_2 \leq F_2^{-1}(u)]}$$

we can express λ_U as

$$\lambda_U = \lim_{u \nearrow 1} \frac{1 + \mathbf{C}(u, u) - 2u}{1 - u}$$

39. See Embrechts, Lindskog, and McNeil (2001) and Nelsen (1999) for a more detailed description.
40. For a multivariate version of the Marshall-Olkin copula, see Embrechts, Lindskog, and McNeil (2001).
41. Li (1999) considers a copula that links individual survival probabilities to model the joint survival probability. However, as we have explained previously, this can be done exactly in the same way if we consider default probabilities instead of survival probabilities.
42. Remember that $(\mathcal{F}_{i,t}) = (\mathcal{G}_{X,t}) \vee (\mathcal{G}_{i,t})$ is the information generated by the state variables plus the information generated by the default status of firm i .
43. This distinction between pseudo and real default intensities can also be found in Gregory and Laurent (2002).
44. If we are considering I firms, the number of parameters of the normal copula will be $\frac{I(I-1)}{2}$, and we have to add the degrees of freedom parameter in the case of the t -copula.
45. See Nelsen (1999).
46. See also De Matteis (2001) and Frees and Valdez (1998).
47. For a more detailed description of copula parameters estimation see Frees and Valdez (1998), Bouy e et al. (2000), Durrleman, Nikeghbali, and Roncalli (2000), De Matteis (2001), and Patton (2002).
48. We have to distinguish the case in which we estimate the parameters of the marginal distributions and the copula function altogether from the case in which we first estimate the parameters of the marginal distributions and then, using those parameters, we estimate the parameters of the copula function. The latter approach is

called inference functions for margins or the IFM method.

49. Imagine we have N random samples of a bivariate vector (Y_1, Y_2) , let us denote them by (y_1^n, y_2^n) , $n = 1, \dots, N$. The sample estimators of Kendall's tau ($\hat{\tau}$) and Spearman's rho ($\hat{\rho}_s$) are given by:

$$\hat{\rho}_s = \frac{12}{N(N^2 - 1)} \sum_{n=1}^N \left(\text{rank}(y_1^n) - \frac{n(n+1)}{2} \right) \times \left(\text{rank}(y_2^n) - \frac{n(n+1)}{2} \right)$$

$$\hat{\tau} = \frac{c - d}{c + d} = \frac{2}{N(N-1)} \sum_{n < m} \text{sign}[(y_1^n - y_1^m)(y_2^n - y_2^m)]$$

where c and d are the number of concordant and discordant pairs, respectively.

50. In some cases analytical expressions for the dependence measures are available. Otherwise we have to use a root-finding procedure.
51. For a detailed description of the diversity score method see Cifuentes, Murphy, and O'Connor (1996), Cifuentes and O'Connor (1996), and Cifuentes and Wilcox (1998).
52. See Jouanin et al. (2001).
53. See Committee on the Global Financial System (2003).

REFERENCES

- Bielecki, T. R., and Rutkowski, M. (2002). *Credit Risk: Modeling, Valuation and Hedging*. Springer Finance.
- Bingham, N. H., and Kiesel, R. (2002). Semiparametric modelling in finance: Theoretical foundations. *Quantitative Finance* 2, 241–250.
- Black, F., and Cox, J. C. (1976). Valuing corporate securities: Some effects of bond indenture provisions. *Journal of Finance* 31, 351–367.
- Bouyé, E., Durrleman, V., Nikeghbali, A., Riboulet, G., and Roncalli, T. (2000). Copulas for finance: A reading guide and some applications. Working Paper, Groupe de Recherche Opérationnelle, Crédit Lyonnais, France.
- Brody, D., Hughston, L., and Macrina, A. (2007). Beyond hazard rates: A new framework for credit-risk modelling, *Advances in Mathematical Finance* Edited by Fu, M. C., Jarrow, R. A., Yen, J., and Elliott, R. J. pp: 231–257.
- Cifuentes, A., Murphy, E., and O'Connor, G. (1996). Emerging market collateralized bond obligations: An overview. Special Report, Moody's Investor Service.
- Cifuentes, A., and O'Connor, G. (1996). The binomial expansion method applied to CBO/CLO analysis. Special Report, Moody's Investor Service.
- Cifuentes, A., and Wilcox, C. (1998). The double binomial method and its application to a special case of CBO structures. Special Report, Moody's Investor Service.
- Collin-Dufresne, P., Goldstein, R. S., and Martin, J. S. (2001). The determinants of credit spread changes. *Journal of Finance* 56, 2177–2207.
- Committee on the Global Financial System (2003). Credit risk transfer. Bank for International Settlements.
- Costinot, A., Roncalli, T., and Teiletche, J. (2000). Revisiting the dependence between financial markets with copulas. Working Paper, Groupe de Recherche Opérationnelle, Crédit Lyonnais, France.
- Davis, M., and Lo, V. (1999). Modelling default correlation in bond portfolios. Working Paper, Tokyo-Mitsubishi International.
- Davis, M., and Lo, V. (2001). Infectious defaults. *Quantitative Finance* 1, 382–386.
- De Matteis, R. (2001). Fitting copulas to data. Diploma Thesis, Institute of Mathematics of the University of Zurich.
- Drissen, J. (2005). Is default event risk priced in corporate bonds? *Review of Financial Studies* 18, 165–195.
- Duffee, G. R. (1998). The relation between Treasury yields and corporate bond yield spreads. *Journal of Finance* 53, 65–79.
- Duffee, G. R. (1999). Estimating the price of default risk. *Review of Financial Studies* 12, 197–226.
- Duffie, D. (2005). Credit risk modelling with affine processes. *Journal of Banking and Finance* 29, 2751–2802.
- Duffie, D., Filipovic, D., and Schachermayer, W. (2003). Affine processes and applications in finance. *Annals of Applied Probability* 13, 984–1053.
- Duffie, D., and Garleanu, N. (2001). Risk and valuation of collateralized debt obligations. Working Paper, Graduate School of Business, Stanford University.

- Duffie, D., and Kan, R. (1996). A yield factor model of interest rates. *Mathematical Finance* 6, 379–406.
- Duffie, D., Pan, J., and Singleton, K. J. (2000). Transform analysis and asset pricing for affine jump diffusions. *Econometrica* 68, 1343–1376.
- Duffie, D., Pedersen, L. H., and Singleton, K. J. (2003). Modeling sovereign yield spreads: A case study of Russian debt. *Journal of Finance* 58, 119–160.
- Duffie, D., Schroder, M., and Skidas, C. (1996). Recursive valuation of defaultable securities and the timing of resolution of uncertainty. *Annals of Applied Probability* 6, 1075–1090.
- Duffie, D., and Singleton, K. J. (1999a). Modeling term structures of defaultable bonds. Working Paper, Graduate School of Business, Stanford University.
- Duffie, D., and Singleton, K. J. (1999b). Simulating correlated defaults. Working Paper, Graduate School of Business, Stanford University.
- Duffie, D., and Singleton, K. J. (2003). Credit risk: Pricing, measurement and management. *Princeton Series in Finance*.
- Durrleman, V., Nikeghbali, A., and Roncalli, T. (2000). Which copula is the right one? Working Paper, Groupe de Recherche Opérationnelle, Crédit Lyonnais, France.
- Embrechts, P., Lindskog, F., and McNeil, A. (2001). Modelling dependence with copulas and applications to risk management. Working Paper, Department of Mathematics, ETHZ, Zurich.
- Embrechts, P., McNeal, A., and Straumann, D. (1999). Correlation and dependence in risk management: Properties and pitfalls. Working Paper, Risklab, ETHZ, Zurich.
- Elizalde, A. (2005a). Credit default swap valuation: An application to Spanish firms. Available at www.abelelizalde.com.
- Elizalde, A. (2005b). Do we need to worry about credit risk correlation? *Journal of Fixed Income* 15: 3, 42–59.
- Fermanian, J., and Scaillet, O. (2004). Some statistical pitfalls in copula modelling for financial applications. FAME Research Paper No. 108.
- Frees, E. W., and Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal* 2, 1–25.
- Frey, R., and Backhaus, J. (2003). Interacting defaults and counterparty risk: A Markovian approach. Working Paper, Department of Mathematics, University of Leipzig.
- Frey, R., McNeil, A. J., and Nyfeler, M. A. (2001). Copulas and credit models. Working Paper, Department of Mathematics, ETHZ, Zurich.
- Galiani, S. (2003). Copula functions and their application in pricing and risk managing multivariate credit derivative products. Master Thesis, King's College London.
- Genest, C., and Rivest, L. P. (1993). Statistical inference procedures for bivariate archimedean copulas. *Journal of the American Statistical Association* 88, 1034–1043.
- Georges, P., Lamy, A. G., Nicolas, E., Quibel, G., and Roncalli, T. (2001). Multivariate survival modelling: A unified approach with copulas. Working Paper, Groupe de Recherche Opérationnelle, Crédit Lyonnais, France.
- Giesecke, K. (2002a). Credit risk modelling and valuation: An introduction. Working Paper, Humboldt-Universität, Berlin.
- Giesecke, K. (2002b). An exponential model for dependent defaults. Working Paper, Humboldt-Universität, Berlin.
- Gregory, J., and Laurent, J. P. (2002). Basket default swaps, CDO's and factor copulas. Working Paper, BNP Paribas.
- Houweling, P., and Vorst, A. C. F. (2001). An empirical comparison of default swap pricing models. Working Paper, Erasmus University Rotterdam.
- Hughston, L. P., and Turnbull, S. (2001). Credit risk: Constructing the basic building blocks. *Economic Notes* 30, 281–292.
- Hull, J., and White, A. (2001). Valuing credit default swaps II: Modelling default correlations. *Journal of Derivatives* 8, 12–22.
- Jarrow, R. (1999). Estimating recovery rates and pseudo default probabilities implicit in debt and equity prices. *Financial Analysts Journal* 57, 75–92.
- Jarrow, R., Lando, D., and Turnbull, S. (1997). A Markov model for the term structure of credit risk spreads. *Review of Financial Studies* 10, 481–523.
- Jarrow, R., Lando, D., and Yu, F. (2001). Default risk and diversification: Theory and applications. Working Paper, University of California, Irvine.
- Jarrow, R., and Turnbull, S. (1995). Pricing derivatives on financial securities subject to credit risk. *Journal of Finance* 50, 53–86.
- Jarrow, R., and Yu, F. (2001). Counterparty risk and the pricing of defaultable securities. *Journal of Finance* 56, 1765–1799.

- Jamshidian, F. (2002). Valuation of credit default swaps and swaptions. Working Paper, NIB Capital Bank.
- Joe, H. (1997). Multivariate models and dependence concepts. *Monographs on Statistics and Applied Probability* 73, Chapman and Hall, London.
- Jouanin, J. F., Rapuch, G., Riboulet, G., and Roncalli, T. (2001). Modelling dependence for credit derivatives with copulas. Working Paper, Groupe de Recherche Opérationnelle, Crédit Lyonnais, France.
- Kijima, M. (2000). Valuation of a credit swap of the basket type. *Review of Derivatives Research* 4, 81–97.
- Kijima, M., and Muromachi Y. (2000). Credit events and the valuation of credit derivatives of basket type. *Review of Derivatives Research* 4, 55–79.
- Lando, D. (1994). On Cox processes and credit risky securities. Chapter 3, Ph.D. Thesis, Cornell University.
- Li, D. X. (1999). On default correlation: A copula function approach. Working Paper 99–07, The Risk Metrics Group.
- Longstaff, F., and Schwartz, E. (1995). A simple approach to valuing risky and floating rate debt. *Journal of Finance* 50, 781–819.
- Madam, D., and Unal, H. (1998). Pricing the risks of default. *Review of Derivatives Research* 2, 121–160.
- Marshall, A., and Olkin, I. (1988). Families of multivariate distributions. *Journal of the American Statistical Association* 83, 834–841.
- Merton, R. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29, 449–470.
- Nelsen, R. B. (1999). An introduction to copulas. *Lecture Notes in Statistics* 139, Springer-Verlag, New York.
- Patton, A. J. (2002). Estimation of copula models for time series of possibly different lengths. Working Paper, University of California, San Diego.
- Rogge, E., and Schönbucher, P. J. (2003). Modelling dynamic portfolio credit risk. Working Paper, Department of Statistics, Bonn University.
- Schönbucher, P. J. (2003). *Credit derivatives pricing models*. Wiley Finance.
- Schönbucher, P. J., and Schubert, D. (2001). Copula-dependent default risk in intensity models. Working Paper, Department of Statistics, Bonn University.
- Sklar, A. (1959). Fonctions de repartition a n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* 8, 229–231.
- Xie, Y., Wu, C., and Shi, J. (2004). Do macroeconomic variables matter for the pricing of default risk? Evidence from the residual analysis of the reduced-form model pricing errors. Presented at 2004 FMA Annual Meeting Program.
- Yu, F. (2002a). Correlated default in reduced-form models. Working Paper, University of California, Irvine.
- Yu, F. (2002b). Modelling expected return on defaultable bonds. *Journal of Fixed Income* 12, 69–81.
- Zhang, F. X. (2003). What did the credit market expect of Argentina default? Evidence from default swap data. Working Paper, Federal Reserve Board, Division of Research and Statistics, Washington.

Structural Models in Credit Risk Modeling

ABEL ELIZALDE, PhD

Credit Derivatives Strategy, J.P. Morgan

Abstract: Structural models and reduced-form models are the two primary types of credit risk models that seek to statistically describe default processes. Structural models use the evolution of firms' structural variables, such as asset and debt values, to model the time of default. In contrast, reduced-form models do not consider structural variables in an explicit manner when modeling default processes; instead, they model default as an exogenously driven process. Structural models include first passage models, liquidity process models, and state dependent models.

In this entry we review the structural approach for credit risk modeling, both considering the case of a single firm and the case with default dependencies between firms. In the single firm case, we review the *Merton (1974) model* and first passage models, examining their main characteristics and extensions. Liquidation process models extend first passage models to account for the possibility of a lengthy liquidation process, which might or might not end up in default. Finally, we review *structural models* with state-dependent cash flows (recession vs. expansion) or debt coupons (rating-based). The estimation of structural models is also addressed in this entry, covering the different ways proposed in the literature. Finally, we present some approaches to model default dependencies between firms within the structural approach. These approaches account for two types of *default correlations*: cyclical default correlation and contagion effects.

REVIEW OF STRUCTURAL MODELS

Structural models use the evolution of firms' structural variables, such as asset and debt values, to determine the time of default. Merton's model (1974) was the first modern model of default and is considered the first structural model. In Merton's model, a firm defaults if, at the time of servicing the debt, its assets are below its outstanding debt. A second approach, within the structural framework, was introduced by Black and Cox (1976). In this approach defaults occur as soon as the firm's asset value falls below a certain threshold. In contrast to the Merton approach, default can occur at any time.

Reduced form models do not consider the relation between default and firm value in an explicit manner. In contrast to structural models, the time of default in intensity models is not

determined via the value of the firm, but it is the first jump of an exogenously given jump process. The parameters governing the default hazard rate are inferred from market data.¹

Structural default models provide a link between the credit quality of a firm and the firm's economic and financial conditions. Thus, defaults are endogenously generated within the model instead of exogenously given as in the reduced approach. Another difference between the two approaches refers to the treatment of recovery rates: Whereas reduced models exogenously specify recovery rates, in structural models the value of the firm's assets and liabilities at default will determine recovery rates.

The structural literature on credit risk starts with the paper by Merton (1974), who applies the option pricing theory developed by Black and Scholes (1973) to the modeling of a firm's debt. In Merton's model, the firm's capital structure is assumed to be composed by equity and a zero-coupon bond with maturity T and face value of D . The firm's equity is simply a European call option with maturity T and strike price D on the asset value and, therefore, the firm's debt value is just the asset value minus the equity value. This approach assumes a very simple and unrealistic capital structure and implies that default can only happen at the maturity of the zero-coupon bond.

Black and Cox (1976) introduced the first of the so-called *first passage models* (FPM). First passage models specify default as the first time the firm's asset value hits a lower barrier, allowing default to take place at any time. When the default barrier is exogenously fixed, as in Black and Cox (1976) and Longstaff and Schwartz (1995), it acts as a safety covenant to protect bondholders. Alternatively it can be endogenously fixed as a result of the stockholders' attempt to choose the default threshold that maximizes the value of the firm.²

Structural models have considered interest rates both as nonstochastic processes³ and as stochastic processes.^{4,5}

In first passage models, by definition, default occurs the first time the asset value goes below a certain lower threshold, that is, the firm is liquidated immediately after the default event. In contrast with first passage models, a new set of models has been put forward, supported by recent theoretical and empirical research, where a default event does not immediately cause liquidation but it represents the beginning of a process, the liquidation process, which might or might not cause liquidation after it is completed. This practice is consistent, for example, with Chapter 11 of the U.S. Bankruptcy Law, where firms filing for bankruptcy are granted a court-supervised grace period (up to several years) aimed at sorting out their financial problems in order to, if possible, avoid liquidation. We label those models *liquidation process models* (LPM).

State dependent models (SDM) represent, together with LPM, two recent efforts to incorporate into structural models different real-life phenomena. Although theoretically they make good sense, they lack empirical research testing their performance. SDM assume that some of the parameters governing the firm's ability to generate cash flows or its funding costs are state dependent, where states can represent the business cycle (recession vs. expansion) or the firm's external rating.

After the single firm case, we review some structural models for default correlations, in order to account for both cyclical default correlation⁶ as well as credit risk contagion effects.⁷ We will finish the default correlation section mentioning the so-called factor models.⁸

We concentrate on the review of the dynamics of the processes that generate the default times, without paying attention to the valuation formulas for defaultable bonds that each model generates. The aim of this entry is to serve as an introduction and guide to the literature of structural credit risk models. We provide an extensive list of references for each model specification and possible extensions or related papers.

SINGLE FIRM

We denote the physical and risk-neutral probability measures as $\bar{\mathbf{P}}$ and \mathbf{P} respectively, and assume an arbitrage-free market.⁹ Unless otherwise stated, all probabilities and expectations are taken under the risk-neutral measure. The model for the default-free term structure of interest rates is given by a short-rate process r_t .

Merton's Model

Merton (1974) makes use of the Black and Scholes (1973) option pricing model to value corporate liabilities. This is a straightforward application only if we adapt the firm's capital structure and the default assumptions to the requirements of the Black-Scholes model. Let us assume that the capital structure of the firm is comprised by equity and by a zero-coupon bond with maturity T and face value of D , whose values at time t are denoted by E_t and $z(t, T)$ respectively, for $0 \leq t \leq T$. The firm's asset value V_t is simply the sum of equity and debt values. Under these assumptions, equity represents a call option on the firm's assets with maturity T and strike price of D . If at maturity T the firm's asset value V_T is enough to pay back the face value of the debt D , the firm does not default and shareholders receive $V_T - D$. Otherwise ($V_T < D$) the firm defaults, bondholders take control of the firm, and shareholders receive nothing. Implicit in this argument is the fact that the firm can only default at time T . This assumption is important to be able to treat the firm's equity as a vanilla European call option, and therefore apply the Black-Scholes pricing formula.

The rest of the assumptions Merton (1974) adopts are the inexistence of transaction costs, bankruptcy costs, taxes, or problems with indivisibilities of assets; continuous time trading; unrestricted borrowing and lending at a constant interest rate r ; no restrictions on the short selling of the assets; the value of the firm is invariant under changes in its capital structure

(Modigliani-Miller theorem), and that the firm's asset value follows a diffusion process.

The firm's asset value is assumed to follow a diffusion process given by

$$dV_t = rV_t dt + \sigma_V V_t dW_t \quad (1)$$

where σ_V is the (relative) asset volatility and W_t is a Brownian motion.¹⁰

The payoffs to equityholders and bondholders at time T under the assumptions of this model are respectively, $\max\{V_T - D, 0\}$ and $V_T - E_T$, that is,

$$E_T = \max\{V_T - D, 0\} \quad (2)$$

$$z(T, T) = V_T - E_T \quad (3)$$

Applying the Black-Scholes pricing formula, the value of equity at time t ($0 \leq t \leq T$) is given by

$$\begin{aligned} E_t(V_t, \sigma_V, T - t) \\ = e^{-r(T-t)} \left[e^{r(T-t)} V_t \Phi(d_1) - D \Phi(d_2) \right] \end{aligned} \quad (4)$$

where $\Phi(\cdot)$ is the distribution function of a standard normal random variable and d_1 and d_2 are given by

$$d_1 = \frac{\ln\left(\frac{e^{r(T-t)} V_t}{D}\right) + \frac{1}{2} \sigma_V^2 (T - t)}{\sigma_V \sqrt{T - t}} \quad (5)$$

$$d_2 = d_1 - \sigma_V \sqrt{T - t} \quad (6)$$

The probability of default at time T is given by

$$P[V_T < D] = \Phi(-d_2) \quad (7)$$

Therefore, the value of the debt at time t is $z(t, T) = V_t - E_t$.

In order to implement Merton's model we have to estimate the firm's asset value V_t , its volatility σ_V (both unobservable processes), and we have to transform the debt structure of the firm into a zero-coupon bond with maturity T and face value D .

The maturity T of the zero-coupon bond can be chosen either to represent the maturity structure of the debt, for example as the Macaulay duration of all the liabilities, or simply as a required time horizon (for example, in case we

are pricing a credit derivative with some specific maturity).

Criticisms and Extensions

The main advantage of Merton's model is that it allows us to directly apply the theory of European options pricing developed by Black and Scholes (1973). But to do so the model needs to make the necessary assumptions to adapt the dynamics of the firm's asset value process, interest rates, and capital structure to the requirements of the Black-Scholes model. There is a trade-off between realistic assumptions and ease of implementation, and Merton's model opts for the latter one. All extensions to this model introduce more realistic assumptions trying to end up with a model not too difficult to implement and with closed, or at least numerically feasible, solutions for the expressions of the debt value and the default probabilities. Merton (1974) presents some extensions to the model, in order to account for coupon bonds, callable bonds, stochastic interest rates, and relaxing the assumption that the Modigliani-Miller theorem holds.

One problem of Merton's model is the restriction of default time to the maturity of the debt, ruling out the possibility of an early default, no matter what happens with the firm's value before the maturity of the debt. If the firm's value falls down to minimal levels before the maturity of the debt but it is able to recover and meet the debt's payment at maturity, the default would be avoided in Merton's approach.

Another handicap of the model is that the usual capital structure of a firm is much more complicated than a simple zero-coupon bond. Geske (1977, 1979) considers the debt structure of the firm as a coupon bond, in which each coupon payment is viewed as a compound option and a possible cause of default. At each coupon payment, the shareholders have the option either to make the payment to bondholders,¹¹ obtaining the right to control the firm until the next coupon, or to not make the payment, in which case the firm defaults. Geske also extends the model to consider character-

istics such as sinking funds, safety covenants, debt subordination, and payout restrictions.

The assumption of a constant and flat term structure of interest rates is another major criticism the model has received. Jones et al. (1984, p. 624) suggest that "there exists evidence that introducing stochastic interest rates, as well as taxes, would improve the model's performance." Stochastic interest rates allow us to introduce correlation between the firm's asset value and the short rate, and have been considered, among others, by Ronn and Verma (1986), Kim, Ramaswamy, and Sundaresan (1993), Nielsen et al. (1993), Longstaff and Schwartz (1995), Briys and de Varenne (1997), and Hsu, Saá-Requejo, and Santa-Clara (2004).

Another characteristic of Merton's model, which will also be present in some of the FPM, is the predictability of default. Since the firm's asset value is modeled as a geometric Brownian motion and default can only happen at the maturity of the debt, it can be predicted with increasing precision as the maturity of the debt comes near. As a result, in this approach default does not come as a surprise, which makes the models generate very low short-term credit spreads.¹² As we shall review, introducing jumps in the process followed by the asset value has been one of the solutions considered to this problem.

Delianedis and Geske (2001) study the proportion of the credit spread that, in a corporate bond data set, is explained by default risk, using the Merton (1974) and Geske (1977) frameworks. They conclude that it only explains a small fraction of the credit spreads; the rest is attributable to taxes, jumps, liquidity, and market risk factors. They also include a jump component in the Merton model finding that (p. 24) "while jumps may explain a portion of the residual spread it is unlikely that jumps can explain it entirely."

First Passage Models

First passage models were introduced by Black and Cox (1976) extending the Merton model to

the case when the firm may default at any time, not only at the maturity date of the debt.

Consider, as in the previous section, that the dynamics of the firm's asset value under the risk-neutral probability measure \mathbf{P} are given by the diffusion process

$$dV_t = rV_t dt + \sigma_V V_t dW_t \quad (8)$$

and that there exists a lower level of the asset value such that the firm defaults once it reaches this level. Although Black and Cox (1976) considered a time-dependent default threshold, let us assume first a constant default threshold $K > 0$. If we are at time $t \geq 0$, default has not been triggered yet and $V_t > K$, then the time of default τ is given by

$$\tau = \inf \{s \geq t \mid V_s \leq K\} \quad (9)$$

Using the properties of the Brownian motion W_t , in particular the reflection principle, we can infer the default probability from time t to time T :¹³

$$P[\tau \leq T \mid \tau > t] = \Phi(h_1) + \exp\left\{2\left(r - \frac{\sigma_V^2}{2}\right) \ln\left(\frac{K}{V_t}\right) \frac{1}{\sigma_V^2}\right\} \Phi(h_2) \quad (10)$$

where

$$h_1 = \frac{\ln\left(\frac{K}{e^{r(T-t)}V_t}\right) + \frac{\sigma_V^2}{2}(T-t)}{\sigma_V\sqrt{T-t}} \quad (11)$$

$$h_2 = h_1 - \sigma_V\sqrt{T-t} \quad (12)$$

FPM have been extended to account for stochastic interest rates, bankruptcy costs, taxes, debt subordination, strategic default, time-dependent and stochastic default barriers, jumps in the asset value process, and so on. Although these extensions introduce more realism into the model, they increment its analytical complexity.¹⁴

The default threshold, always positive, can be interpreted in various ways. We can think of it as a safety covenant of the firm's debt, which allows the bondholders to take control of the company once its asset value has reached this level. The safety covenant would

act as a protection mechanism for the bondholders against an unsatisfactory corporate performance. In this case, the default threshold would be deterministic, although possibly time dependent, and exogenously fixed when the firm's debt is issued. Kim, Ramaswamy, and Sundaresan (1993) and Longstaff and Schwartz (1995) assume an exogenously given constant default threshold K . Black and Cox (1976) consider a time-dependent default barrier given by $e^{-\gamma(T-t)}K$. A particular case of the Black and Cox default threshold specification is to consider $\gamma = r$, that is, to consider a default barrier equal to the face value of the debt discounted at the risk-free interest rate. In that case, the default threshold can be made stochastic if the model considers a stochastic process for the interest rate, as in Briys and de Varenne (1997).

Longstaff and Schwartz (1995) choose a constant default threshold and point out that "since it is the ratio of V_t to K , rather than the actual value of K , that plays the major role in our analysis, allowing a more general specification for K simply makes the model more complex without providing additional insight into the valuation of risky debt."

Hsu, Saá-Requejo, and Santa-Clara (2004) suggest that V_t and K do not matter directly to the valuation of default risky bonds but only through their ratio, which is a measure of the solvency of the firm. They model the default threshold as a stochastic process, which together with the stochastic process assumed for the firm's asset value, allow them to obtain the stochastic process of the ratio $\frac{V_t}{K}$. The dynamics of the ratio $\frac{V_t}{K}$ are used to price corporate bonds.

The default threshold can also be chosen endogenously by the stockholders to maximize the value of the equity.¹⁵ The literature has also considered the possibility of negotiation processes between stockholders and bondholders when the firm goes near the point of financial distress, from which the default threshold is determined.¹⁶

Similar to the description of the choice of the face-value of the zero-coupon in the Merton model, in FPM the default threshold can be

calculated as a weighted average of short and long-term debts.

Interest rates can be considered either as a constant or as a stochastic process.¹⁷ The stochasticity of interest rates allows the model to introduce correlation between asset value and interest rates, and to make the default threshold stochastic, in the cases when it is specified as the discounted value of the face value of the debt. Nielsen et al. (1993) and Longstaff and Schwartz (1995) consider a Vasicek process for the interest rate, correlated with the firm's asset value:

$$dV_t = (c - d)V_t dt + \sigma_V V_t dW_t \quad (13)$$

$$dr_t = (a - br_t) dt + \sigma_r d\bar{W}_t \quad (14)$$

$$d\bar{W}_t dW_t = \rho dt \quad (15)$$

where \bar{W}_t and W_t are correlated Brownian motions. Other specifications for the stochastic process of the short rate have been considered. For example Kim, Ramaswamy, and Sundaresan (1993) suggest a CIR process

$$dr_t = (a - br_t) dt + \sigma_r \sqrt{r_t} d\bar{W}_t \quad (16)$$

and Briys and de Varenne (1997) a generalized Vasicek process

$$dr_t = (a(t) - b(t)r_t) dt + \sigma_r(t) d\bar{W}_t \quad (17)$$

Hsu, Saá-Requejo, and Santa-Clara (2004) consider both the case of independence between risk-free interest rates and the default generating mechanism (given by the dynamics of the ratio $\frac{V_t}{K}$) and the case of correlation between both processes, specifying the risk-free rate as a CIR process. They present an interesting empirical illustration of the model, covering the calibration of the risk-free rate process and the estimation of the model's parameter through the generalized method of moments.

Drawbacks and Extensions

The principal drawback of FPM is the analytical complexity that they introduce, which is increased if we consider stochastic interest rates or endogenous default thresholds. This mathe-

matical complexity makes it difficult to obtain closed form expressions for the value of the firm's equity and debt, or even for the default probability, forcing us to make use of numerical procedures.

The empirical testing of FPM and structural models in general has not been very successful.¹⁸ Eom, Helwege, and Huang (2003), who carry out an empirical analysis of five models (Merton, Geske, Leland and Toft, Longstaff and Schwartz, and Collin-Dufresne and Goldstein), conclude that (p. 502)

Using estimates from the implementations we consider most realistic, we agree that the five structural bond pricing models do not accurately price corporate bonds. However, the difficulties are not limited to the underprediction of spreads. . . . they all share the same problem of inaccuracy, as each has a dramatic dispersion of predicted spreads.

Zhou (1997, Abstract) indicates that "the empirical application of a diffusion approach has yielded very disappointing results."

Another drawback of the structural models presented before is the so-called predictability of defaults. Generally, structural models consider continuous diffusion processes for the firm's asset value and complete information about the asset value and default threshold. In this setting, the actual distance from the asset value to the default threshold tells us the nearness of default, in such a way that if we are far away from default the probability of default in the short-term is close to zero, because the asset value process needs time to reach the default point. The knowledge of the distance of default and the fact that the asset value follows a continuous diffusion process makes default a predictable event, that is, default does not come as a surprise.

This predictability of defaults makes the models generate short-term credit spreads close to zero. In contrast, it is observed in the market that even short-term credit spreads are bounded from below, incorporating the possibility of an unexpected default or deterioration in the firm's credit quality.¹⁹

The same characteristics of the structural models that imply the predictability of default also imply predictability of recovery. In models that do not consider strategic defaults, the bondholders get the remaining value of the firm in case of default, which is precisely the value of the default threshold at default. Thus, if we assume complete information about the asset value and default threshold, the recovery rate is also a predictable quantity.

Essentially, two ways out of the predictability effects of structural models have been proposed in the literature. The predictability of default comes from the assumption of investors' perfect knowledge of the firm's asset value and default threshold. In practice, it is not possible to deduce from the capital structure of the firm neither the value of the firm V_t , its volatility σ_V , nor the level of the default threshold. If we consider incomplete information about either the firm value process, the default threshold (or both), investors can only infer a distribution function for these processes, which makes defaults impossible to predict. These considerations can be found, among others, in Duffie and Lando (2001), Giesecke (2005), and Jarrow and Protter (2004).²⁰

The second way consists in incorporating jumps in the dynamics of the firm value, which implies that the asset value of the firm can suddenly drop, reducing drastically the distance of default (between the asset value and default threshold), or even causing a default if the drop is sufficiently high. Thus, default is not a predictable event anymore, the default probabilities for short maturities do not tend to zero, and so the credit spreads generated. Zhou (1997, 2001a) and Hilberink and Rogers (2002) deal with structural models in which the firm's asset value incorporates a jump component. While Zhou extends the Longstaff and Schwartz (1995) model considering a lognormally distributed jump component, Hilberink and Rogers (2002) opt for an extension of Leland (1994) and Leland and Toft (1996) using Levy processes, which only allow for down-

ward jumps in the firm's value. Both models avoid the problem of default predictability implying positive credit spreads for short maturities. Another characteristic of jump models is that they convert the recovery payment at default in a random variable, since the value of the firm can drop suddenly below the default threshold, whereas if the firm's value follows a diffusion process without jumps, the value of the firm at default, that is, what bondholders get, is always equal to the default threshold because of the continuity of the firm's value path.

Fouque, Sircar and Solna (2006) consider the effect of introducing stochastic volatility in FPM, finding that it increases short-term spreads.

Davydenko (2005) criticizes existing structural models because they obviate the liquidity reasons as the main determinants of default for some firms, particularly the ones with high external financing costs (p. 2):

Several default triggers have been proposed in structural models of debt pricing. Most models assume that a firm defaults when the market value of its assets falls below a certain boundary (Black and Cox, 1976; Leland, 1994). This default boundary may correspond to an exogenous net-worth covenant, or to the endogenously determined threshold at which equityholders are no longer willing to service debt obligations. Should the firm find itself in a liquidity crisis while its asset value is still above the boundary, equityholders in these models will always be willing and able to avoid default by raising outside financing. This approach contrasts with the assumption that firms default when current assets fall short of current obligations, due to either a minimum cash-flow covenant, or market frictions precluding the firm from raising sufficient new external financing (Kim et al., 1993; Anderson and Sundaresan, 1996). Models incorporating both value- and liquidity-based defaults are rare, and little empirical evidence is available to motivate the choice of the default trigger. If, in reality, default is triggered by different factors for different firms, existing models are likely to lack accuracy in predictions.

Davydenko (2005), using a sample of U.S. (speculative rating-grade) bond issuers from 1996 to 2003, shows that the importance of liquidity shortages in triggering default for a

particular firm depends on the firm's cost of external financing (p. 2): "firms with low costs of external financing default when the continuation value of assets is low. By contrast, if external funds are costly, a liquidity crisis may force reorganization even if the going-concern surplus is still substantial."²¹ Moreover, the author presents empirical evidence against the view that default is triggered when the asset value crosses a particular threshold.

Therefore, empirical evidence suggests that structural models need to be theoretically extended in order to incorporate the possibility of the firms defaulting because of liquidity shortages and high funding costs.

Estimation

The literature provides several ways of calibrating V_t and σ_V . The first method makes use of Ito's lemma to obtain a system of two equations in which the only two unknown variables are V_t and σ_V .²² Assume the firm's equity value follows a geometric Brownian motion under \mathbf{P} , with volatility σ_E :

$$dE_t = r E_t dt + \sigma_E E_t dW_t \quad (18)$$

Since the value of the equity is a function of time and of the value of the assets, $E_t = f(V_t, t)$, we can apply Ito's lemma to get

$$\begin{aligned} dE_t = & \left[\frac{\delta f(V_t, t)}{\delta t} + \frac{\delta f(V_t, t)}{\delta V_t} V_t r \right. \\ & \left. + \frac{1}{2} \frac{\delta^2 f(V_t, t)}{(\delta V_t)^2} (V_t \sigma_V)^2 \right] dt \\ & + \frac{\delta f(V_t, t)}{\delta V_t} V_t \sigma_V dW_t \end{aligned} \quad (19)$$

Comparing the coefficients multiplying the Brownian motion in the two previous equations we obtain the following identity

$$\sigma_E E_t = \frac{\delta f(V_t, t)}{\delta V_t} V_t \sigma_V \quad (20)$$

Noting that $\frac{\delta f(V_t, t)}{\delta V_t} = \frac{\delta E_t}{\delta V_t} = \Phi(d_1)$ and rearranging we obtain the first equation of the system:²³

$$\sigma_V = \frac{E_t}{V_t} \sigma_E \Phi(d_1) \quad (21)$$

The second equation results simply from matching the theoretical value of equity with the observed market price (\hat{E}_t):

$$E_t(V_t, \sigma_V, T - t) = \hat{E}_t \quad (22)$$

As we mentioned before, the only two unknowns in the system formed by the last two equations are V_t and σ_V .²⁴

Duan (1994) points out some drawbacks of the previous method. First, the method considers the equity volatility as constant and independent of the firm's asset value and time. Second, he claims that the first equation is redundant since it is used to derive the second equation. And third, the traditional method does not provide us with distribution functions, or even confidence intervals, for the estimates of V_t and σ_V .

Duan (1994) proposes another method of estimating V_t and σ_V , based on maximum likelihood estimation using equity prices and the one-to-one relationship between equity and asset levels given by (4).²⁵ Duan et al. (2004) follow the maximum likelihood approach introduced by Duan (1994) but, unlike previous works, they take into account the survivorship issue, by incorporating into the likelihood function the fact that a firm survived. They argue that (p. 3), "In the credit risk setting, it is imperative for analysts to recognize the fact that a firm in operation has by definition survived so far. Estimating a credit risk model using the sample of equity prices needs to reflect this reality, or runs the risk of biasing the estimator."

Duan and Fulop (2005) extend Duan's (1994) maximum likelihood estimation method to account for the fact that observed equity prices might be contaminated by trading noises. They find that taking into account trading noises generates lower estimates for the asset volatility σ_V

and therefore overestimates the firms' default probabilities.

Bruche (2005) describes how structural models can be estimated using a simulated maximum likelihood procedure, which allows us to use data on any of the firm's traded claims (bonds, equity, CDS, ...) as well as balance sheet information to improve the efficiency of the estimation. The paper explores the possibility of considering that not only equity, but the rest of the claims used in the estimation procedure can be priced with noise, showing that (p. 3) "even small amounts of noise can have serious consequences for estimation results when they are ignored."

A different way of estimating V_t and σ_V , which can be found in Jones et al. (1984), consists simply of estimating the asset value as the sum of the equity market value, the market value of traded debt, and the estimated value of nontraded debt. Provided with a time series for V_t we can estimate its volatility σ_V .

Hull, Nelken, and White (2004) propose a way to estimate the model's parameters from implied volatilities of options on the company's equity, avoiding the need to estimate σ_E and to transform the firm's debt structure into a zero-coupon bond. Using as inputs two equity implied volatilities and an estimate of the firm's debt maturity T , their model provides us with an estimate of σ_V and the leverage ratio $\frac{De^{-r(T-t)}}{V_t}$, which allows us to calculate E_t and the probability of default. We should note that to calculate the value of the debt $z(t, T) = V_t - E_t$ we still need an estimate for V_t .²⁶

We still have to estimate the default threshold K . Sundaram (2001) indicates that (p. 7) "default tends to occur in practice when the market value of the firm's assets drops below a critical point that typically lies below the book value of all liabilities, but above the book value of short-term liabilities." Thus, one approach is to choose a value for D between those two limits. Davydenko (2005) estimates the default threshold to be around 72% of the firm's face value of debt.

Liquidation Process Models

In FPM default occurs the first time the asset value goes below a certain lower threshold, that is, the firm is liquidated immediately after the default event; the default event corresponding to the crossing of the asset value through the lower barrier. In contrast with FPM, a new set of models considers the case where the default event does not immediately cause liquidation but it represents the beginning of a process, the liquidation process, which might or might not cause liquidation after it is completed. As explained earlier, we refer to these models as liquidation process models (LPM).

The distinction between the terms *default event* and *liquidation* must be clear to understand LPM and their differences with FPM. A default event takes place when the firm's asset value V_t goes below the lower threshold K (which can be exogenous, constant, time dependent, stochastic, or endogenously derived). A default event signals the beginning of a financially distressed period, which will not necessarily lead to liquidation. Liquidation takes place when the firm is actually liquidated, its activity stops, and its remainings are distributed among its claimholders.

In FPM described above the default event does coincide with liquidation.²⁷ However, as pointed out by Couderc and Renault (2005, p. 2), most liquidations "do not arise suddenly but are rather the conclusion of a long lasting process." As pointed out by Moraux (2004, p. 3): "Empirical studies in USA have found that additional 'survival' periods beyond the main default event last up to 3 years (Altman-Eberhart (1994), Betker (1995), Hotchkiss (1995)). Helwege (1999) reports that the longest default of modern US junk bond market is seven years long."²⁸

The fact that the liquidation process can take quite a while implies that when empirically studying the causes of liquidation past information shows up as a significant explanatory variable, together, of course, with contemporaneous

information, because it comprises information about the liquidation process. Information here refers to the firms' financial variables as well as financial markets, business cycle, credit markets, and default cycle indicators. Couderc and Renault (2005) use a database containing the rating history of over ten thousand firms for the period 1981–2003 and analyze, using duration models, whether past values of several financial markets (business cycle, credit markets, and default cycle) are relevant in explaining default probabilities in addition to their contemporaneous values. Their results show the critical importance of past information in default probabilities.

LPM extend FPM to account for the fact that the liquidation time takes place after (sometimes quite a lot after) the occurrence of a default event. François and Morellec (2004), Moraux (2004), and Galai, Raviv, and Wiener (2005) put forward a theoretical LPM.

François and Morellec (2004) argue that while in most of FPM the default event leads to an immediate liquidation of the firm's assets, firms in financial distress have several options to deal with their distress. First, under Chapter 7 of the U.S. Bankruptcy Code, they can liquidate its assets straight away. This possibility would fit FPM. However firms can also file for bankruptcy under Chapter 11 of the U.S. Bankruptcy Code and start a court-supervised liquidation process. The authors refer to existing literature (p. 390) to provide some evidence about the relevance of Chapter 11:

Upon default, the court grants the firm a period of observation during which the firm can renegotiate its claims. At the end of this period, the court decides whether the firm continues as a going concern or not.

Empirical studies show that most firms emerge from Chapter 11. Only a few firms (5%, according to Gilson, John, and Lang [1990] and Weiss [1990], and between 15% and 25%, according to Morse and Shaw [1988]) are eventually liquidated under Chapter 7 after filing Chapter 11. Why do some firms recover while others do not? It is generally acknowledged (see Wruck 1990 or White 1996) that there exist two types of defaulting firms. First,

firms that are economically sound promptly recover under Chapter 11. Default was only due to a temporary financial distress. Second, firms that are economically unsound keep on losing value under Chapter 11.²⁹

François and Morellec consider that, after a default event, i.e. after the asset value V_t goes below the lower threshold K , a firm is liquidated if and only if V_t remains below K consecutively during a period of time of a given length d (which in their numerical simulations they take to be two years). If a default event happens and the asset value remains under the lower threshold for a period lower than d , the liquidation process finishes and the firm continues in business as usual. The term consecutively in the definition of liquidation above means that the number of successfully managed past default events and liquidation periods³⁰ the firm has experienced does not affect the maximum length d of future liquidation periods.

The authors provide closed-form solutions for corporate debt and equity values and analyze the implications of the model for optimal leverage and credit spreads. Numerical simulations show that credit spreads are an increasing function of the length d .

Moraux (2004) extends the François and Morellec (2004) model including an additional cause of liquidation to François and Morellec's one (which they call liquidation procedure A). Under his proposed liquidation procedure, procedure B, liquidation happens when the total, that is, cumulative, time the firm's assets value stands under the default threshold exceeds d . The difference between procedures A and B lies in the words consecutively and cumulative, and Moraux (2004, p. 17) explains it clearly:

Under the procedure A, each time the firm value process passes through and above K , the liquidation procedure is closed and the hypothetical distress counter is set to zero. The next time a default event occurs, an identical procedure is run and an equal period of time d is granted. . . . Under the procedure B, the distress counter is never set to zero. Subsequent granted periods (and therefore tolerance) will be lower and lower as more default events and

long financial distress will be observed. In fact, the granted time is lowered (each time) by the duration just used.

Financial distress refers to the situation in which $V_t < K$. A firm can be liquidated by either one or the other liquidation procedures. Moraux (2004) shows that any liquidation procedure based on the time spent by the firm in financial distress is bounded by the procedures A and B in the sense that its implied liquidation date will be higher (lower) than the liquidation date implied by procedure B (A).

The author derives closed form solutions for different claims such as equity, different seniority debts, and convertible debt. In particular, the value of equity is derived as a down and out Parisian option written on the firm assets under liquidation procedure A and as a down and out cumulative call option under liquidation procedure B. Numerical simulations show that the value of equity is an increasing function of d , and that, unlike in François and Morellec (2004), credit spreads increase or decrease with d depending on the seniority of the debt.

Galai, Raviv, and Wiener (2005) represent a step forward in the refinement of LPM, proposing a model extending and including the two previous ones. They argue that in the two previous models, the only thing that matters for a firm to be liquidated is the amount of time it spends in financial distress (either successively or cumulatively), but they fail to (p. 5) “capture the following two common features of bankruptcy procedures: (i) Recent distress events may have a greater effect on the decision to liquidate a firm’s assets than old distress events. . . . (ii) Severe distress events may have greater effect on the decision to liquidate a firm than mild distress events.” To account for such two stylized facts, the authors propose a structural model in which a firm is liquidated when a state variable representing the cumulative weighted time period spent by the firm in distress exceeds d . At each time, the cumulative weighted time period is computed as a weighted average of the total time spent by

the firm in distress, weighted by (1) how far away in the past such distress occurred and (2) how severe was such a distress, where distress severity is measured as an increasing function of $\max\{0, K - V\}$.

Galai, Raviv, and Wiener’s model has as special cases models such as Merton (1974), Black and Cox (1976), Leland (1994), Fan and Sundaresan (2001), François and Morellec (2004), and Moraux (2004). As a consequence it represents a general LPM so far. They solve the model numerically using Monte Carlo simulation based on Parisian options and Parisian contracts techniques to value debt and equity. They provide a very intuitive comparison of the liquidation mechanics in their general model with François and Morellec’s and Moraux’s ones, showing that Moraux’s cumulative liquidation procedure (B) has too strong memory because far-away distress periods have the same impact on liquidation triggering as current ones.

Although theoretically very appealing, LPM have not, unlike FPM, been empirically tested, and remains a field for future research.

State Dependent Models

Another avenue for (so far) theoretical research within the structural approach consists of extending standard models with regime switching: Some of the model parameters are state-contingent. As we review below, states can represent the state of the business cycle or simply the firm’s external rating. Cash flows, bankruptcy costs, and funding costs might be state-dependent.

This branch of structural models is able to reduce the problems of predictability of defaults (and recovery) suffered by standard models because the firm is subject to exogenous changes of parameters, which affect its ability to generate cash flows or its funding costs, which are the main drivers of default probabilities.

Hackbarth, Miao, and Morellec (2004) and Elizalde (2005b) put forward two different

models illustrating the previous ideas. In both cases the authors provide closed form expressions for the value of equity and debt, whose solutions imply solving systems of ordinary differential equations.

In Hackbarth, Miao, and Morellec (2004) cash flows and recovery rates depend on the state of the business cycle. Cash flows x_t follow a geometric Brownian motion and are scaled by a business cycle scalar factor: They are higher in expansions $y_H x_t$ than in recessions $y_L x_t$, $y_H > y_L$. In the same way, bankruptcy costs are expressed as a state-dependent fraction $1 - \alpha$ of the firm's assets; again, the recovery rate in expansions α_H is higher than in recessions α_L , $\alpha_H > \alpha_L$. At each point in time, there is an exogenous probability of switching between recession and expansion. The default threshold is endogenously chosen by equityholders to maximize the value of equity, and it turns out to be higher in recessions: The firm defaults earlier in recessions than in expansions. Numerical examples illustrate the implications of the model for default thresholds, default clustering, optimal leverage (countercyclical), and credit spreads. As argued above the model is able to generate nontrivial short-term spreads.

Elizalde (2005b) develops a structural model which, although originally applied to banks, can be extended to any firm. In contrast with previous models, the firms' asset value is assumed to be unobserved by debtholders. Debtholders rely on the ratings published by rating agencies to set the debt's coupon as a function of those ratings. As a consequence, the firms' funding costs are contingent on their ratings. Rating agencies perform timely audits to firms, with a given frequency, to find out their risk and asset levels, which determine the rating. Switching from one rating to another implies changes in the cost of debt and, as a consequence, in the ability of the firm to repay it. As in Hackbarth, Miao, and Morellec (2004) the default threshold is chosen endogenously by equityholders and it is rating-dependent.

As described by Duffie (2005, p. 2772), "It has become increasingly common for bond issuers to link the size of the coupon rate on their debt with their credit rating, offering a higher coupon rate at lower ratings, perhaps in an attempt to appeal to investors based on some degree of hedging against a decline in credit quality." This embedded derivative is called a ratings-based step-up. The author illustrates an example of a ratings-based step-up bond issued by Deutsche Telecom in 2002 with coupon payments linked to the firm's rating. While Elizalde (2005b) derives the price of such a bond using a structural model, Duffie provides its pricing formula using an intensity model.³¹

Like LPM, state-dependent models have only been developed theoretically and their future success in credit risk modeling (if any) lies in their empirical applicability and their ability to replicate and predict credit spreads and default probabilities.

DEFAULT CORRELATION

To incorporate default dependencies between firms using structural models the literature has essentially relied on natural extensions of single firm's models, either Merton (1974) type models or FPM. We will start this section reviewing these extensions, under which the default dependences between firms are introduced through correlated asset processes.³² Giesecke (2004) and Giesecke and Goldberg (2004) suggest that the default correlation implied by the use of correlated firms' asset processes accounts for the dependence of the firms' credit quality on common macroeconomic factors, what they term cyclical default correlation, but it does not account for credit risk contagion across firms and periods of default clustering. In order to introduce the contagion correlation in the model, Giesecke (2004) and Giesecke and Goldberg (2004) propose a model in which the firms' default thresholds are dependent one to each other and are unknown to investors.

After reviewing Giesecke (2004) and Giesecke and Goldberg (2004) we present factor models, which express the value of the firms' assets as a function of several common factors, which generate the correlation, and an idiosyncratic factor.³³ Duan et al. (2002) and Hull and White (2001) present two alternative approaches to deal with default correlation in structural models.

Cyclical Default Correlation

The most natural way to introduce default dependencies between firms in structural models is by correlating the firms' asset processes.³⁴ Suppose we have $i = 1, \dots, I$ different firms with asset value processes given by

$$dV_{i,t} = rV_{i,t}dt + \sigma_{V_i} V_{i,t}dW_{i,t} \quad (23)$$

for $i = 1, \dots, I$, where $W_{1,t}, \dots, W_{I,t}$ are correlated Brownian motions. As in the single firm case, these models imply predictable defaults. One way of getting rid of the default predictability would be to introduce jump components in the firms' asset processes. Those jump components could be either correlated or uncorrelated across firms. Correlated jump components, besides making defaults unpredictable, would also account for credit risk contagion effects. The main problem lies in the calibration of the jump components.

Contagion Default Correlation

Cyclical default correlation does not account for all the credit risk dependence between firms. Giesecke (2004) and Giesecke and Goldberg (2004) extend structural models for default correlation to incorporate credit risk contagion effects. The default of one firm can trigger the default of related firms. Furthermore, default times tend to concentrate in some periods of time in which the probability of default of all firms is increased and which cannot be to-

tally, or even partially, explained by the firms' common dependence on some macroeconomic factors.

Contagion effects can arise in this setting by direct links between firms in terms of, for example, commercial or financial relationships. The news about the default of one firm has a big impact on the credit quality of other related firms, which is immediately reflected in their default probabilities.

In structural FPM we assume that investors have complete information about both asset processes and default thresholds, so they always know the nearness of default for each firm, that is, the distance between the actual level of the firm's assets and its default threshold.³⁵

Giesecke (2004) and Giesecke and Goldberg (2004) introduce contagion effects in the model by relaxing the assumption that investors have complete information about the default thresholds of the firms. In Giesecke (2004), bondholders do not have perfect information, neither about the thresholds nor about their joint distribution. However, they form a prior distribution, which is updated at any time one of such thresholds is revealed, which only happens when the corresponding firm defaults. In Giesecke (2004) investors have incomplete information about the firms' default thresholds but complete information about their asset processes. Giesecke and Goldberg (2005) extend that framework to one in which investors do not have information about the firms' asset values or about their default thresholds. In this case, default correlation is introduced through correlated asset processes and, again, investors receive information about the firms' asset and default barrier only when they default. Such information is used to update their priors about the distribution of the remaining firms' asset values and default thresholds.

The incomplete information about the level of the default thresholds and the fact that those levels are dependent among firms (through a copula function) generate the source of credit

risk contagion. Investors form a belief about the level of the firms' default thresholds. Each time one of the firms defaults, the true level of its default threshold is revealed, and investors use this new information to update their beliefs about the default thresholds of the rest of the firms. This sudden updating of the investors' perceptions about the default thresholds of the firm, and thus about the nearness of default for each firm, introduces the default contagion effects in the models.

This model allows for the introduction of default correlation both through dependencies between firms' asset values, cyclical default correlations, and through dependencies between firms' default barriers, contagion effects.

The major problem of this approach is to calibrate and estimate the default threshold copula. See Giesecke (2003) for some remarks on how to choose and calibrate that copula.

Factor Models

Factor models consider the firms' asset values as a function of a group of common factors, which introduce the default correlation in the model, plus a firm's specific factor:

$$V_{i,t} = \sum_{j=1}^J w_{i,j} Z_{j,t} + \epsilon_{i,t} \quad (24)$$

where Z_1, \dots, Z_J represent the common factors, $\epsilon_1, \dots, \epsilon_I$ the firms' specific factors (independent of Z_1, \dots, Z_J), and the correlation structure is given by the coefficient w . Once we know the realization of the common factors, the firms' asset value and thus the firms' default probabilities are independent.

The calibration of factor models is usually carried out by a logit or probit regression, depending on the assumptions about the distribution of the factors. Schönbucher (2000), Finger (1999), and Frey, McNeil, and Nyfeler (2001) present illustrations of these models.

KEY POINTS

- The structural approach for credit risk modeling considers the link between the credit quality of a firm and the firm's economic and financial conditions. As a consequence, defaults are endogenously generated within the models (instead of exogenously given as in reduced-form models). By relying on the firm's assets and liabilities to model default risk, structural models also provide a framework to analyze recovery rates.
- The structural literature on credit risk started with the Merton model, which used option pricing theory for valuing the debt of a firm. In the Merton model, the firm's capital structure is composed by equity and a zero-coupon. The firm is assumed to default at the bond maturity if the value of its assets is below the face value of the bond.
- The structural modeling approach has mainly developed by relaxing the strict assumptions of the Merton model, generating more realistic models, which take into account different characteristics of firms' capital structure, bankruptcy laws, macro variables, and so on.
- Structural models include first passage models, liquidation process models, and state-dependent models. In first passage models a default occurs the first time the firm's asset value goes below a certain lower threshold (related to the firm's level of debt). These models assume that the firm is liquidated immediately after the default event. Liquidation process models extend first passage models by taking into account the fact that firms that file for bankruptcy may avoid liquidation. Finally, state-dependent models assume that some of the parameters governing the firm's ability to generate cash flows or its funding costs depend on variables such as the business cycle (recession vs. expansion) or the firm's external rating.
- There are several ways to account for default correlation within the structural approach. Cyclical default correlation and factor

models consider the dependence of firms' credit quality on common macroeconomic factors. Contagion models include the dependence of firms' credit quality on other firms' credit quality.

NOTES

1. For a review of reduced form models, see Entry 22.
2. See, for example, Leland (1994) and Leland and Toft (1996).
3. See Black and Cox (1976), Geske (1977), Leland (1994), and Leland and Toft (1996).
4. See Ronn and Verma (1986), Kim, Ramaswamy, and Sundaresan (1993), Nielsen et al. (1993), Longstaff and Schwartz (1995), Briys and de Varenne (1997), and Hsu, Saá-Requejo, and Santa-Clara (2004).
5. We reproduce here an updated list of extensions and improvements within the literature of structural models provided by Eom, Helwege, and Huang (2003, p. 500): "See, for example, Black and Cox (1976), Bryis and De Varenne (1997), Goldstein, Ju, and Leland (2001), Ho and Singer (1982), Kim, Ramaswamy, and Sundaresan (1993), Leland (1994, 1998), Nielsen, Saá-Requejo, and Santa-Clara (1993), and Titman and Torous (1989). Anderson and Sundaresan (1996) and Mella-Barral and Perraudin (1997) incorporate strategic defaults into traditional structural models. See also Acharya and Carpenter (2002), Acharya et al. (2000), Anderson, Sundaresan, and Tychon (1996), Fan and Sundaresan (2000), and Huang (1997). Duffie and Lando (2001) take into account incomplete accounting information. Garbade (1999) examines managerial discretion. Huang and Huang (2002) and Zhou (2001) incorporate jumps."
6. See Zhou (2001b) and Giesecke (2004).
7. See Giesecke (2004) and Giesecke and Goldberg (2004).
8. See Schönbucher (2000), Finger (1999), and Frey, McNeil, and Nyfeler (2001).
9. For our purposes we shall use the class of equivalent probability measures \mathbf{P} where nondividend-paying asset processes discounted with the risk-free interest rate are \mathbf{P} -martingales.
10. Since we are working under the risk-neutral probability measure, the drift term of the asset value process is given by the risk-free instantaneous interest rate. Under the physical probability measure, r would be replaced by a parameter μ_V representing the mean rate of return on assets; and the firm's asset process would be given by

$$dV_t = \mu_V V_t dt + \sigma_V V_t d\tilde{W}_t$$
 where \tilde{W}_t is a Brownian motion under the physical probability measure $\tilde{\mathbf{P}}$.
11. Since shareholders finance each coupon issuing new equity, the dilution effect reduces the relative value of each share.
12. See Jones et al. (1984) and Franks and Torous (1989).
13. See Iori (2003) and Chapter 3.1 in Jeanblanc and Rutkowski (2000).
14. For an extensive review of FPM, see Chapter 3 in Bielecki and Rutkowski (2002) and references therein.
15. See, for example, Mello and Parsons (1992), Nielsen et al. (1993), Leland (1994), Anderson and Sundaresan (1996), Leland and Toft (1996), Mella-Barral and Perraudin (1997), and François and Morellec (2004).
16. For a discussion of strategic debt service, see Mella-Barral and Perraudin (1997), Fan and Sundaresan (2000), and references therein.
17. See Black and Cox (1976), Leland (1994), and Leland and Toft (1996) for models with constant interest rates, and see Kim, Ramaswamy, and Sundaresan (1993), Nielsen et al. (1993), Longstaff and Schwartz (1995), Briys and de Varenne (1997), Collin-Dufresne and Goldstein (2001), and Hsu,

- Saá-Requejo, and Santa-Clara (2004) for models with stochastic interest processes.
18. See Anderson and Sundaresan (2000), Eom, Helwege, and Huang (2003), and Ericsson and Reneby (2004).
 19. See Jones et al. (1984), Franks and Torous (1989), Sarig and Warga (1989), Fons (1994), Huang and Huang (2003), and Leland (2004).
 20. Elizalde (2005a) presents a review of structural models that appeared in the literature, which consider incomplete information assumptions and bridge the gap between the structural and the reduced approach.
 21. At any given point in time the firms' owners face a decision of whether to liquidate the firm, or to maintain the status quo by continuing operations under the current regime, also referred to as a going concern.
 22. See, for example, Jones et al. (1984), Ronn and Verma (1986), Eom et al. (2000), Delianedis and Geske (2003), and Ericsson and Reneby (2005).
 23. Crosbie and Bohn (2003) point out that this equation holds only instantaneously, and that in practice, market leverage, which would be represented here by $\frac{e^{-t(T-t)}D}{V_t}$, moves around far too much for that equation to provide reasonable results.
 24. Ronn and Verma (1986) extend the estimation to the cases of nonstationary σ_E and stochastic interest rates.
 25. For a complete description of this method see Duan (1994), Duan et al. (2002), and Ericsson and Reneby (2005), who also present a comparison with the traditional method.
 26. Eom et al. (2000) suggest another procedure to estimate σ_V , which they term the bond-implied volatility method.
 27. In fact, we called simply *default* to such an occurrence. In what follows we shall use the terms *default* and *liquidation* with the same meaning (different from default event!). A default event starts the process of liquidation. The process of liquidation has two possible endings: liquidation or default and reorganization (which happens when the firm manages to improve its financial health and avoid closure).
 28. See also Frank and Torous (1989) and Gilson (1997).
 29. See also Kahl (2001) and Morrison (2003).
 30. Successfully managed means that such liquidation periods did not last longer than d and, as a consequence, did not trigger liquidation.
 31. Manso, Strulovici, and Tchisty (2004) present an alternative derivation of ratings-based step-up bonds using structural models, and review the existing literature.
 32. See Zhou (2001b).
 33. See Schönbucher (2000), Finger (1999), and Frey, McNeil, and Nyfeler (2001).
 34. See Zhou (2001b).
 35. We are not considering here jump components in the dynamics of the assets processes.

REFERENCES

- Acharya, V., and Carpenter, J. (2002). Corporate bond valuation and hedging with stochastic interest rates and endogenous bankruptcy. *Review of Financial Studies* 15: 1355–1383.
- Acharya, V., Huang, J. Z., Subrahmanyam, M., and Sundaram, R. (2000). Costly financing, optimal payout policies and the valuation of corporate debt. Working Paper, NYU.
- Altman, E., and Eberhart, A. (1994). Do priority provisions protect a bond-holder's investment? *Journal of Portfolio Management* 20: 67–75.
- Anderson, R., and Sundaresan, S. (1996). Design and valuation of debt contracts. *Review of Financial Studies* 9: 37–68.
- Anderson, R., and Sundaresan, S. (2000). A comparative study of structural models of corporate bond yields. *Journal of Banking and Finance* 24: 255–269.
- Anderson, R., Sundaresan, S., and Tychon, P. (1996). Strategic analysis of contingent claims. *European Economic Review* 40: 871–881.
- Betker, B. (1995). An empirical examination of prepackaged bankruptcy. *Financial Management* 24: 3–18.

- Bielecki, T. R., and Rutkowski, M. (2002). *Credit Risk: Modeling, Valuation and Hedging*. New York: Springer Finance.
- Bingham, N. H., and Kiesel, R. (2002). Semi-parametric modelling in finance: Theoretical foundations. *Quantitative Finance* 2: 241–250.
- Black, F., and Cox, J. C. (1976). Valuing corporate securities: Some effects of bond indenture provisions. *Journal of Finance* 31: 351–367.
- Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.
- Briys, E., and de Varenne, F. (1997). Valuing risky fixed rate debt: An extension. *Journal of Financial and Quantitative Analysis* 31: 239–248.
- Bruche, M. (2005). Estimating structural bond pricing models via simulated maximum likelihood. London School of Economics, Financial Markets Group Discussion Paper 534.
- Collin-Dufresne, P., and Goldstein, R. (2001). Do credit spreads reflect stationary leverage ratios? *Journal of Finance* 56: 1929–1957.
- Couderc, F., and Renault, O. (2005). Times-to-default: Life cycle, global and industry cycle impacts. FAME Research Paper No. 142.
- Crosbie, P. J., and Bohn, J. R. (2002). Modelling default risk. Working Paper, KMV Corporation.
- Davydenko, S. (2005). When do firms default? University of Toronto Working Paper.
- Delianedis, G., and Geske, R. (2003). Credit risk and risk neutral default probabilities: Information about rating migrations and defaults. UCLA Working Paper.
- Delianedis, G., and Geske, R. (2001). The components of corporate credit spreads: Default, recovery, tax, jumps, liquidity, and market factors. Working Paper, UCLA.
- Duan, J. C. (1994). Maximum likelihood estimation using price data of the derivative contract. *Mathematical Finance* 4: 155–157.
- Duan, J. C., and Fulop, A. (2005). Estimating the structural credit risk model when equity prices are contaminated by trading noises. University of Toronto Working Paper.
- Duan, J. C., Gauthier, G., Simonato, J. G., and Zaanoun, S. (2002). Maximum likelihood estimation of structural credit spread models: Deterministic and stochastic interest rates. Working Paper.
- Duan, J. C., Gauthier, G., Simonato, J. G., and Zaanoun, S. (2004). Estimating Merton's model by maximum likelihood with survivorship consideration. EFA 2004 Maastricht Meetings Paper No. 4190.
- Duffie, D. (2005). Credit risk modelling with affine processes. *Journal of Banking and Finance* 29: 2751–2802.
- Duffie, D., and Lando, D. (2001). Term structure of credit spreads with incomplete accounting information. *Econometrica* 69: 633–664.
- Duffie, D., and Singleton, K. J. (2003). *Credit Risk: Pricing, Measurement and Management*. Princeton: Princeton Series in Finance.
- Elicizalde, A. (2005a). Reconciliation of structural-reduced form models. Available at www.abelelizalde.com.
- Elicizalde, A. (2005b). From Basel I to Basel II: An analysis of the three pillars. Available at www.abelelizalde.com.
- Eom, Y. H., Helwege, J., and Huang, J. Z. (2003). Structural models of corporate bond pricing: An empirical analysis. *Review of Financial Studies* 17: 499–544.
- Ericsson, J., and Reneby, J. (2005). Estimating structural bond pricing models. *Journal of Business* 78: 707–736.
- Fan, H., and Sundaresan, S. (2000). Debt valuation, renegotiation and optimal dividend policy. *Review of Financial Studies* 13: 1057–1099.
- Finger, C. (1999). Conditional approaches for credit-metrics portfolio distributions. *Credit Metrics Monitor* Spring: 14–26.
- Fons, J. S. (1994). Using default rates to model the term structure of credit risk. *Financial Analysts Journal* 50, 5, 25–33.
- Fouque, J., Sircar, R., and Solna, K. (2006). Stochastic volatility effects on defaultable bonds. *Applied Mathematical Finance* 13, 3, 215–244.
- François, P., and Morellec, E. (2004). Capital structure and asset prices: Some effects of bankruptcy procedures. *Journal of Business* 77: 387–411.
- Franks, J., and Torous, W. (1989). An empirical investigation of U.S. firms in reorganization. *Journal of Finance* 44: 747–769.
- Frey, R., McNeil, A. J., and Nyfeler, M. A. (2001). Modelling dependent defaults: Asset correlations are not enough! Working Paper, Department of Mathematics, ETHZ, Zurich.
- Galai, D., Raviv, A., and Wiener, Z. (2005). Liquidation triggers and the valuation of equity and debt. Working Paper, School of Business Administration, The Hebrew University of Jerusalem.
- Garbade, K. (1999). Managerial discretion and the contingent valuation of corporate securities. *Journal of Derivatives* 6: 65–76.

- Geske, R. (1977). The valuation of corporate liabilities as compound options. *Journal of Financial and Quantitative Analysis* 12: 541–552.
- Geske, R. (1979). The valuation of compound options. *Journal of Financial Economics* 7: 63–81.
- Gilson, S., John, K., and Lang, L. (1990). Troubled debt restructurings: An empirical study of private reorganization of firms in default. *Journal of Financial Economics* 27: 315–353.
- Gilson, S. (1997). Transaction costs and capital structure choice: Evidence from financially distressed firms. *Journal of Finance* 52: 161–196.
- Goldstein, R., Ju, N., and Leland, H. (2001). An EBIT-based model of dynamic capital structure. *Journal of Business* 74: 483–512.
- Giesecke, K. (2004). Correlated default with incomplete information. *Journal of Banking and Finance* 28: 1521–1545.
- Giesecke, K. (2005). Default and information. Working Paper, Cornell University.
- Giesecke, K., and Goldberg, L. R. (2004). Sequential defaults and incomplete information. *Journal of Risk* 7: 1–26.
- Hackbarth, D., Miao, J., and Morellec, E. (2004). Capital structure, credit risk, and macroeconomic conditions. FAME Research Paper No. 125.
- Helwege, J. (1999). How long do junk bonds spend in default? *Journal of Finance* 54: 341–357.
- Hilberink, B., and Rogers, L. C. G. (2002). Optimal capital structure and endogenous default. *Finance and Stochastics* 6: 237–263.
- Ho, T., and Singer, R. (1982). Bond indenture provisions and the risk of corporate debt. *Journal of Financial Economics* 10: 375–406.
- Hotchkiss, E. (1995). Postbankruptcy performance and management turnover. *Journal of Finance* 50: 3–21.
- Hsu, J., Saá-Requejo, J., and Santa-Clara, P. (2004). Bond pricing with default risk. UCLA Working Paper.
- Huang, J. Z. (1997). Default risk, renegotiation, and the valuation of corporate claims. Ph.D. dissertation, NYU.
- Huang, J. Z., and Huang, M. (2003). How much of the corporate-treasury yield spread is due to credit risk? Working Paper, Penn State and Stanford Universities.
- Hull, J., Nelken, I., and White, A. (2004). Merton's model, credit risk, and volatility skews. *Journal of Credit Risk* 1: 3–28.
- Hull, J., and White, A. (2001). Valuing credit default swaps II: Modelling default correlations. *Journal of Derivatives* 8: 12–22.
- Iori, G. (2003). Exotic options. Lecture Notes, MSc in Financial Mathematics, King's College London.
- Jarrow, R. A., and Protter, P. (2004). Structural versus reduced form models: A new information based perspective. *Journal of Investment Management* 2: 1–10.
- Jeanblanc, M., and Rutkowski, M. (2000). *Modelling of default risk: An overview*. In *Mathematical Finance: Theory and Practice*. Beijing: Higher Education Press, pp. 171–269.
- Jones, P., Mason, S., and Rosenfeld, E. (1984). Contingent claim analysis of corporate capital structures: An empirical investigation. *Journal of Finance* 39: 611–625.
- Kahl, M. (2001). Financial distress as a selection mechanism: Evidence from the United States. Anderson School, Finance Working Paper No. 16-01.
- Kim, I. J., Ramaswamy, K., and Sundaresan, S. M. (1993). Does default risk in coupons affect the valuation of corporate bonds? A contingent claims model. *Financial Management* 22: 117–131.
- Leland, H. E. (1994). Risky debt, bond covenants and optimal capital structure. *Journal of Finance* 49: 1213–1252.
- Leland, H. E. (1998). Agency costs, risk management, and capital structure. *Journal of Finance* 51: 987–1019.
- Leland, H. E., and Toft, K. B. (1996). Optimal capital structure, endogenous bankruptcy and the term structure of credit spreads. *Journal of Finance* 50: 789–819.
- Leland, H. E. (2004). Predictions of default probabilities in structural models of debt. *Journal of Investment Management* 2.
- Longstaff, F. A., and Schwartz, E. S. (1995). A simple approach to valuing risky fixed and floating rate debt. *Journal of Finance* 50: 789–819.
- Manso, G., Strulovici, B., and Tchisty, A. (2004). Performance-sensitive debt. Working Paper.
- Mella-Barral, P., and Perraudin, W. (1997). Strategic debt service. *Journal of Finance* 52: 531–566.
- Mello, A., and Parsons, J. (1992). Measuring the agency cost of debt. *Journal of Finance* 47: 1887–1904.
- Merton, R. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29: 449–470.
- Moroux, F. (2004). Valuing corporate liabilities when the default threshold is not an absorbing barrier. University of Rennes Working Paper.

- Morrison, E. (2003). Bankruptcy decision-making: An empirical study of small-business bankruptcies. Columbia Law and Economics Working Paper No. 239.
- Morse, D., and Shaw, W. (1988). Investing in bankruptcy firms. *Journal of Finance* 43: 1193–1206.
- Nielsen, L. T., Saá-Requejo, J., and Santa-Clara, P. (1993). Default risk and interest rate risk: The term structure of default spreads. Working Paper, INSEAD.
- Ronn, E. I., and Verma, A. K. (1986). Pricing risk-adjusted deposit insurance: An option based model. *Journal of Finance* 41: 871–895.
- Sarig, O., and Warga, A. (1989). Some empirical estimates of the risk structure of interest rates. *Journal of Finance* 44: 1351–1360.
- Schönbucher, P. J. (2000). Factor models for portfolio credit risk. Working Paper, Department of Statistics, Bonn University.
- Schönbucher, P. J. (2003). *Credit Derivatives Pricing Models*. Hoboken, NJ: Wiley Finance.
- Sundaram, R. K. (2001). The Merton/KMV approach to pricing credit risk. NYU Working Paper.
- Titman, S., and Torous, W. (1989). Valuing commercial mortgages: An empirical investigation of the contingent claim approach to pricing risky debt. *Journal of Finance* 44: 345–373.
- Weiss, L. (1990). Bankruptcy resolution: Direct costs and violation of priority claims. *Journal of Financial Economics* 27: 285–314.
- White, M. (1996). The costs of corporate bankruptcy: A US-European comparison. In *Corporate Bankruptcy: Economic and Legal Perspectives*. Cambridge: Cambridge University Press.
- Wruck, K. (1990). Financial distress, reorganization, and organizational efficiency. *Journal of Financial Economics* 27: 419–444.
- Zhou, C. (1997). A jump-diffusion approach to modelling credit risk and valuing defaultable securities. Washington, DC: Federal Reserve Board.
- Zhou, C. (2001a). The term structure of credit spreads with jump risk. *Journal of Banking and Finance* 25: 2015–2040.
- Zhou, C. (2001b). An analysis of default correlations and multiple defaults. *Review of Financial Studies* 14: 555–576.

Modeling Portfolio Credit Risk

SRICHANDER RAMASWAMY, PhD

Senior Economist, Bank for International Settlements, Basel, Switzerland

Abstract: Modeling credit risk is more challenging than modeling market risk. Some of these challenges relate to the differences in the conceptual approaches used for modeling credit risk and the data limitations associated with the estimation of key model parameters. Hence, there is invariably a subjective element to the modeling of credit risk. A better understanding of these subjective elements can help practitioners to exercise sound judgment and to raise the right questions when trying to interpret the statistical outputs provided by credit risk models.

This entry describes the building blocks to modeling credit risk. Key elements of the building blocks include probability of default of the issuer; recovery rate in the event of issuer default; and the probabilities of migrating to different credit rating states. Various techniques that can be employed to estimate the probability of issuer default, including their relative merits and limitations, are then discussed. Subsequently, the common approaches to quantifying credit risk are introduced. These include the default mode paradigm, which considers default and no default as two states of the world; and the migration mode paradigm, which includes migrations to other credit rating categories including the default state. The entry concludes with a numerical example to illustrate the various concepts presented.

ELEMENTS OF CREDIT RISK

Credit risk is the risk that a borrower will be unable to make payment of interest or principal in a timely manner. Under this definition, a delay in repayments, restructuring of borrower repayments, and bankruptcy, which constitute default events, will fall under credit risk. In addition to this, the mark-to-market loss of a bond resulting from a change in the market perception of the issuer to service the debt in future is also attributed to credit risk. This manifests itself in the form of a widening of the credit spread of the security in question against a risk-free asset, such as the Treasury bond, of similar maturity. The fluctuations in the credit spread between the two securities reflect views on the intrinsic creditworthiness of the issuer of the defaultable security.

The views expressed here are those of the author and not necessarily those of the Bank for International Settlements.

The key determinants of credit risk at the security level include *probability of default* (PD) of the issuer, that is, the probability that the issuer will default on its contractual obligations to repay its debt; *recovery rate* given that the issuer has defaulted; and *rating migration* probabilities, that is, the extent to which the credit quality of the issuer improves or deteriorates as expressed by a change in the probability of default of the issuer. The following sections discuss in greater detail these determinants of credit risk for corporate issuers, and wherever relevant, methods commonly employed to estimate them will be indicated.

Probability of Default

Assessments about an issuer's ability to service debt obligations play a fundamental role in establishing the level of credit risk embedded in a security. This is usually expressed through the default probability that quantifies the likelihood of the issuer not being able to service the debt obligations. Since probability of default is a function of the time horizon over which one measures the debt servicing ability, it is standard practice to assume a one-year time horizon to quantify this.

In general, the approaches used to determine default probabilities at the issuer level fall into two broad categories. The first is empirical in nature and requires the existence of a public credit-quality rating scheme. The second is based on Merton's options theory framework (Merton, 1974), and hence, is a structural approach. The empirical approach to estimating PD makes use of a historical database of corporate defaults to form a static pool of companies having a particular credit rating for a given year. Annual default rates are then calculated for each static pool, which are then aggregated to provide an estimate of the average historical default probability for a given credit rating. If one uses this approach, then the default probabilities for any two issuers having the same credit rating will be identical. The op-

tion pricing approach to estimate default probability makes use of the current estimates of the firm's assets, liabilities, and asset volatility, and hence, is related to the dynamics of the underlying structure of the firm. Each of these approaches is discussed below in greater detail.

Empirical Approach

The empirical approach to determining probability of default is taken by major rating agencies that include Moody's Investors Service, Standard & Poor's Corporation, and Fitch Ratings. The rating agencies assign credit ratings to different issuers on the basis of extensive analysis of both the quantitative and qualitative performance of a firm, which is intended to capture the level of credit risk. (How credit ratings are assigned is not discussed in this entry.) For purpose of illustrating the empirical approach used to determining default probabilities for different credit ratings, we will discuss Moody's methodology.

Moody's rating symbols (Aa, A, Baa, etc.) for issuer ratings are opinions of the ability of the issuer to honor senior unsecured financial obligations and contracts denominated in foreign and/or domestic currency. The rating gradations provide bondholders with a simple system to measure an issuer's ability to meet its senior financial obligations.

In addition to the generic rating categories, Moody's applies numerical modifiers 1, 2, and 3 for the rating categories from Aa to Caa. The modifier 1 indicates that the issuer is in the higher end of its letter-rating category; the modifier 2 indicates a mid-range ranking; the modifier 3 indicates that the issuer is in the lower end of the letter-ranking category. It is customary to refer to a rating change from grade Aa1 to Aa2 as a one-notch rating downgrade. Bonds issued by firms rated between Aaa to Baa are referred to as investment-grade bonds and the rest as non-investment-grade bonds.

It is important to emphasize here that Moody's ratings incorporate assessments of both the likelihood and the severity of default.

Considering that a particular issuer could have debt issues with different collateral and seniority, Moody's approach will lead to different debt issues of a particular issuer having different ratings. However, when an issuer is deemed to have defaulted on a particular debt issue, cross default clauses will require all outstanding debt of the issuer to be considered as having defaulted. This in turn brings us to the following question: What events signal the default of an issuer? Moody's definition of default considers three types of default events:

1. There is a missed or delayed disbursement of interest and/or principal including delayed payments made within a grace period.
2. An issuer files for bankruptcy or legal receivership occurs.
3. A distressed exchange occurs where (1) the issuer offers bondholders a new security or package of securities that amount to a diminished financial obligation, or (2) the exchange had the apparent purpose of helping the borrower to default.

One may note here that the above definitions of default are meant to capture events that change the relationship between the bondholder and bond issuer, which subjects the bondholder to an economic loss.

The empirical approach to determining probability of default relies on historical defaults of various rated issuers. This requires forming a static pool of issuers with a given rating every year and computing the ratio of defaulted issuers after a one-year period to the number of issuers that could have potentially defaulted for the given rating. If, during the year, ratings for certain issuers are withdrawn, then these issuers are subtracted from the potential number of issuers who could have defaulted in the static pool. Specifically, the one-year default rates for A-rated issuers during a given year represent the number of A-rated issuers that defaulted over the year divided by the number of A-rated issuers that could have defaulted over that year. Annual default rates calculated in this manner

for each rating grade are then aggregated to provide an estimate of the average historical default probability for a given rating grade.

We mentioned earlier in this entry that although different debt issues of a particular issuer could have different ratings assigned depending on the seniority of the issue, cross default clauses will require all outstanding debt of a particular issuer to default at the same time. This raises an important question when managing corporate bond portfolios, namely, whether the issuer rating or the rating of the bond issue is to be considered when implying the probability of default. The short answer to this question is that it depends on how credit risk will be quantified for the given bond. The approach taken here to quantify bond-level credit risk requires that the credit rating of the bond issuer is the one to be used. This will be evident when we discuss the quantification of credit risk at the bond level.

Merton's Approach

Merton's approach to estimating the probability of default of a firm builds on the limited liability rule that allows shareholders to default on their obligations while surrendering the firm's assets to the creditors. In this framework, the firm's liabilities are viewed as contingent claims on the assets of the firm, and default occurs at debt maturity when the firm's asset value falls below the debt value. Assuming that the firm is financed by means of equity S_t and a single zero-coupon debt maturing at time T with face value F and current market value B_t , the firm's assets at time t can be represented as

$$A_t = S_t + B_t \quad (1)$$

The probability of default in Merton's framework for the firm will be the probability that the firm's assets is less than the face value of the debt, which is given by,

$$PD = \text{prob}[A_T < F] \quad (2)$$

In order to determine PD in Merton's framework, we need to select a suitable model for the

process followed by A_t . Standard assumption is to postulate that A_t follows a log-normal process with growth rate μ and asset volatility σ_A which is given below:

$$A_t = A_0 \exp[(\mu - 0.5\sigma_A^2)t + \sigma_A\sqrt{t}z_t] \quad (3)$$

In equation (3) z_t is a normally distributed random variable with zero mean and unit variance. Using equation (3) in conjunction with equation (2) we can denote the PD as

$$PD = \text{prob}[\ln A_0 + (\mu - 0.5\sigma_A^2)T + \sigma_A\sqrt{T}z_T < \ln F] \quad (4)$$

In equation (4) we have taken logarithm on both sides of the inequality, since doing so does not change the probabilities. Rearranging the terms in equation (4), the probability of default for the firm can be represented as

$$PD = \text{prob} \left[z_T < -\frac{\ln \frac{A_0}{F} + (\mu - 0.5\sigma_A^2)T}{\sigma_A\sqrt{T}} \right] \quad (5)$$

Since z_T is a normally distributed random variable, PD can be represented as

$$PD = N(-D) \quad (6)$$

where

$$D = \frac{\ln \frac{A_0}{F} + (\mu - 0.5\sigma_A^2)T}{\sigma_A\sqrt{T}} \quad (7)$$

$$N(-D) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-D} \exp(-\frac{1}{2}x^2)dx \quad (8)$$

In equation (7), D represents the distance to default, which is the distance between the logarithm of the expected asset value at maturity and the logarithm of the default point normalized by the asset volatility.

Although Merton's framework for determining PD for issuers is rather simple, applying this directly in practice runs into difficulties. This is because firms seldom issue zero coupon bonds and usually have multiple liabilities. Furthermore, firms in distress may be able to draw on lines of credit to honor coupon and principal

payments, resulting in a maturity transformation of their liabilities.

To resolve these difficulties Moody's KMV suggests some modifications to Merton's framework to make the default probability estimate meaningful in a practical setting (see Crosbie and Bohn, 2002). (Moody's KMV refers to probability of default as expected default frequency or EDFTM). For instance, rather than using face value of the debt to denote the default point, Moody's KMV suggests using the sum of the short-term liabilities (coupon and principal payments due in less than one year) and one-half of the long-term liabilities. This choice is based on the empirical evidence that firms default when their asset value reaches a level that is somewhat between the value of total liabilities and the value of short-term liabilities. Further, since the asset returns of the firms may in practice deviate from a normal distribution, Moody's KMV maps the distance to the default variable to a historical default statistics database to estimate the probability of default. In the KMV framework, default probabilities for issuers can take values in the range between 0.02% and 20%.

To illustrate the KMV approach, let DPT denote the default point, which is equal to the sum of the short-term liabilities due in less than one year and one-half of the long-term liabilities, and $E(A_T)$ the expected value of the firm's assets one year from now. Then the distance to default is given by,

$$D = \frac{\ln \frac{E(A_T)}{DPT}}{\sigma_A} = \frac{\ln \frac{A_0}{DPT} + (\mu - 0.5\sigma_A^2)T}{\sigma_A\sqrt{T}} \quad (9)$$

In equation (9), the market value of the firm's assets is not observed since liabilities of the firm are not traded. What can be observed in the market is the equity value of the firm because it is traded. Since the value of the firm's equity at time T can be seen as the value of a call option on the assets of the firm with a strike price equal to the book value of the liabilities, we have the

following equation:

$$S_T = A_T \times N(d_1) - e^{-rT} \times DPT \times N(d_2) \quad (10)$$

In equation (10), $N(\cdot)$ is cumulative standard unit normal distribution, r is the risk-free interest rate, and the variables d_1 and d_2 are given by,

$$d_1 = \frac{\ln(A_T/DPT) + (r + \frac{1}{2}\sigma_A^2)T}{\sigma_A\sqrt{T}} \quad (11)$$

$$d_2 = d_1 - \sigma_A\sqrt{T} \quad (12)$$

It is possible to show that equity and asset volatility are related through the following relation:

$$\sigma_S = \frac{A_T}{S_T} \times N(d_1) \times \sigma_A \quad (13)$$

From this relation it is possible to solve for the asset value and asset volatility, given the equity value and equity volatility using an iterative procedure. Knowing the asset volatility and asset value, it is possible to compute the distance to default using equation (9) from which probability of default can be inferred.

Relative Merits

The empirical and structural approaches to determine the probability of default for issuers can result in significant differences in the estimates of PD. Both approaches have their relative advantages and disadvantages. For instance, the empirical approach has the implicit assumption that all issuers having the same credit rating will have identical PD. Furthermore, this default probability will be equal to the historical average rate of default. Use of the structural approach, on the other hand, will result in PD being more responsive to changes in economic conditions and business cycles as it incorporates current estimates of the asset value and asset volatility of the firm in deriving this information. One drawback, however, is that the historical database of defaulted firms is comprised mostly of industrial corporates. As a consequence, use of the industrial corpo-

rate default database to infer the PD of regulated financial firms could potentially result in biased PD estimates. Seen from a trading perspective, credit spreads for corporates tend to be influenced much more by agency ratings and credit rating downgrades rather than EDF values. This has the implication that bond market participants tend to attach greater significance to rating agency decisions for pricing. For the purpose of modeling portfolio credit risk and selecting an optimal corporate bond portfolio to replicate the benchmark risk characteristics, we will demonstrate the usefulness of both approaches in the entries to follow.

On Rating Outlooks

Rating agencies provide forward-looking assessment of the creditworthiness of issuers over the medium term. Such forward-looking credit assessments of issuers are referred to as rating outlooks. Outlooks assess the potential direction of an issuer's rating or creditworthiness over the next six months to three years. A positive outlook suggests an improvement in credit rating, a negative outlook indicates deterioration in credit rating, and a stable outlook suggests a rating change is less likely to happen. Bond prices tend to react to changes in rating outlook although no actual change in credit rating has occurred. In particular, the impact on prices is much more significant if the issuer is Baa since a rating downgrade can result in the issuer being rated non-investment grade. Furthermore, if a particular sector (such as Telecom) is having a negative rating outlook, a change in rating outlook from stable to negative for an issuer in this sector can also have a significant effect on bond prices.

The above observations raise the following important question: Should a negative or a positive rating outlook for a given issuer be incorporated in our assessment of PD through a downgrade or upgrade before it has actually happened? The short answer to this question is no, primarily because our estimate of credit risk will incorporate the probability that credit

rating of issuers can change over time. Forcing a rating change for the issuer before it has actually happened may tend to bias our estimate of credit risk.

Captive Finance Companies

Large companies in most industrial sectors have captive finance subsidiaries. The principal function of any financial subsidiary is to support the sales of the parent's products. This function can make the finance company a critical component of the parent's long-term business strategy. In light of this close relationship between the captive finance company and its parent, credit ratings for both are usually identical. However, if the legal clauses guarantee that the parent company's bankruptcy does not automatically trigger the bankruptcy of the financial subsidiary, rating differences may exist between the parent company and its financial subsidiary. For the purpose of quantifying credit risk, we will use the actual credit rating of the financial subsidiary in the calculations.

Estimating the probability of default of financial subsidiaries on the basis of Merton's structural model can lead to difficulties. This is because the equity of the financial subsidiary may not be traded. For example, Ford Motor is traded whereas the financial subsidiary Ford Credit is not traded. Considering that the financing arm of major industrial corporates is vital to the survival of both the parent and the subsidiary, one can argue that the equity market takes this relationship into account when valuing the parent company. Under this argument, one can assign the same probability of default to both companies where only one of them is traded in the market.

Recovery Rate In the event of default, bondholders will not receive all of the promised coupon and principal payments on the bond. Recovery rate for a bond, which is defined as the percentage of the face value that can be recovered in the event of default, will be of natural interest to investors. Considering that

credit market convention is to ask how much of promised debt is lost rather than how much of it is recovered, the term "loss given default" (LGD), which is defined as one minus recovery rate, is also commonly used in the credit risk literature.

In general, estimating the recovery value of the bond in the event of default is rather complex. This is because the payments made to bondholders could take the form of a combination of equity and derivative securities, new debt, or modifications to the terms of the surviving debt. Considering that there may be no market for some forms of payments, it may not be feasible to measure the recovery value. Moreover, the amount recovered could take several months or even years to materialize and could potentially also depend on the relative strength of the negotiating positions. As a result, estimating historical averages of amounts recovered from defaulted debt will require making some simplifying assumptions.

Moody's, for instance, proxy the recovery rate with the secondary market price of the defaulted instrument approximately one month after the time of default. The motivation for such a definition is that many investors may wish to trade out of defaulted bonds, and a separate investor clientele may acquire these and pursue the legal issues related to recovering money from defaulted debt instruments. In this context, Moody's recovery rate proxy can be interpreted as a transfer price between these two investor groups.

Empirical research on recovery rates suggests that industrial sector, seniority of the debt, state of the economy, and credit rating of the issuer one year prior to default are variables that have significant influence on potential recovery rates. For example, during periods of economic downturns, the recovery rate is usually lower relative to historical averages. This has the implication that there is also a time dimension to the potential recovery rates. Differences in recovery rates for defaulted debt across industry sectors arise because the recovery amount will

depend on the net worth of the firm’s tangible assets. For instance, firms belonging to industrial sectors with physical assets such as public utilities have higher recovery rates compared to the industry-wide average. Empirical results also tend to suggest that issuers that were rated investment grade one year prior to default tend to have higher recovery values compared to issuers that were rated non-investment grade.

In order to incorporate the variations in the observed recovery rates over time and between issuers when quantifying credit risk, the standard deviation of recovery rates, denoted σ_{RR} , is taken into account. Including the uncertainty in recovery rates will have the effect of increasing credit risk at the issuer level. Common practice is to use beta distribution to model the observed variations in recovery rates. The advantage of choosing beta distribution is that it has a simple functional form dependent on two parameters that allows for high recovery rate outliers observed in the empirical data to be modeled. The beta distribution has support on the interval 0 to 1, and its density function is given by,

$$f(x, \alpha, \beta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where $\alpha > 0$, $\beta > 0$, and $\Gamma(\cdot)$ is the gamma function. The mean and variance of the beta distribution are given by,

$$\mu = \frac{\alpha}{\alpha + \beta} \tag{15}$$

$$\sigma^2 = \frac{\alpha \cdot \beta}{(\alpha + \beta)^2 \cdot (\alpha + \beta + 1)} \tag{16}$$

Table 1 shows the empirical estimates of recovery rates on defaulted securities covering

the period 1978 to 2001 based on prices at time of default. One can notice that senior secured debt recovers on average 53% of the face value of the debt whereas senior unsecured debt recovers only around 42% of face value. The standard deviation of the recovery rates for all seniority classes is roughly around 25%.

The empirical estimates for average recovery rates tend to vary somewhat depending on the data set used and the recovery rate definition. For instance, the study by Moody’s using defaulted bond data covering the period 1982–2008 suggest that mean recovery rates for senior secured bonds is 53%, for senior unsecured bonds is 32.4%, and for subordinated bonds is 23.5%.

In the numerical examples to be presented in this entry, we have assumed that the bonds under consideration are senior unsecured debt. Furthermore, we have assumed that the standard deviation of the recovery rate is 25% and the average recovery rate is 35%, which is closer to Moody’s estimate incorporating more recent default data.

Rating Migrations The framework for assessing the issuer’s PD comprised of estimating the probability associated with the issuer defaulting on its promised debt payments. In this framework, the issuer is considered to be in one of two states: its current rating or the default state. In practice, default is just one of many states to which the issuer’s rating can transition. Actions of rating agencies can result in the issuer’s rating being downgraded or upgraded by one or several notches. One can associate the concept of a state with each rating grade

Table 1 Recovery Rate Statistics on Defaulted Securities (1978–2001)

Bond Seniority	Number of Issuers	Median	Mean	Standard Deviation
Senior secured	134	57.42%	52.97%	23.05%
Senior unsecured	475	42.27%	41.71%	26.62%
Senior subordinated	340	31.90%	29.68%	24.97%
Subordinated	247	31.96%	31.03%	22.53%

Source: Altman, Resti, and Sironi (2001).

second row correspond to the one-year migration probabilities of an issuer that is currently rated Aa1.

Considering that Table 2 is representative of a typical rating transition matrix that credit agencies publish, one can draw interesting conclusions from the relative frequency of rating downgrades and upgrades from this table. For example, the rating transition matrix suggests that higher ratings have generally been less likely to be revised over one year than lower ratings. Another observation is that large and sudden rating changes occur infrequently. As one moves down the rating scale, the likelihood of a mult notch rating change increases.

Quantifying Credit Risk

In the previous section we identified the important variables that influence credit risk at the security level. In this section we will focus our attention on quantifying credit risk at the security level. Without loss of generality, it will be assumed that the security is a *corporate bond*. Most of us are familiar with the concept of risk in connection with financial assets. In broad terms, risk is associated with potential financial loss that can arise from holding the asset, the exact magnitude of which is difficult to forecast. As a result, it is common to describe the potential loss in value using an appropriate probability distribution whose mean and standard deviation serve as useful measures for risk quantification.

The above practice is well known in the equities market where investors focus on market risk that model variations in stock return. This leads us to quantifying the market risk measures through expected return and standard deviation of return. Under the assumption that equity returns are normally distributed, the realized return will lie within one standard deviation of the expected return with two-thirds probability.

Quantifying credit risk for a corporate bond is similar in principle. Unlike the case for eq-

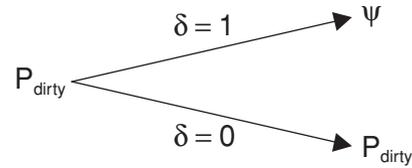


Figure 1 Typical Shape of the Credit Loss Distribution

uities, corporate bond investors focus on the distribution of potential losses that can result from the issuer-specific credit events. Borrowing the principle from equities market, it has become common practice to quantify credit risk at the security level through the mean and standard deviation of the loss distribution. However, there is an important difference between the two risk measures. This pertains to the distribution of credit loss, which unlike for market risk, is far from being a normal distribution. Hence, deviations from the expected loss by one standard deviation can occur more frequently than on one in three occasions. Credit market convention is to refer to the standard deviation of loss resulting from credit events as *unexpected loss* (UL) and the average loss as *expected loss* (EL). Figure 1 shows the typical shape of the distribution of credit losses.

In this section we will discuss how expected and unexpected loss used to quantify credit risk at the security or bond level can be determined. Depending on whether the loss distribution takes into account the changes in bond prices resulting from rating migrations, we can compute two sets of loss variables, one in the default mode and another in the migration mode. Quantification of credit risk in both these modes is discussed below.

Expected Loss Under Default Mode Expected loss under default mode of a bond is defined as the average loss the bondholder can expect to incur if the issuer goes bankrupt. Considering that default probability estimates are based on a one-year holding period, expected loss is also expressed over a one-year period. In

practice, the issuer could actually default at any time during the one-year horizon. Since a bond portfolio manager is usually interested in the worst-case loss scenario, which corresponds to the issuer defaulting in the immediate future, we will use the one-year PD to quantify the worst-case loss. This has the implication that we can quantify credit risk using the current trading price for the bond rather than its one-year forward price. Often, a portfolio manager's goal is to manage relative risk versus a benchmark. In this case, the use of one-year PD in conjunction with current trading prices will not bias the relative risk estimates. Moreover, this assumption leads to considerable simplification in quantifying credit risk since deriving forward yield curves for various credit ratings is quite tedious.

The estimate of expected loss for a security depends on three variables: probability of default of the issuer, the average recovery rate, and the nominal exposure (NE) to the security. One can think of the default process δ as being a Bernoulli random variable that takes the value 0 or 1. The value $\delta = 1$ signals a default and the value $\delta = 0$ signals no default. Conditional upon default, the recovery rate \Re is a random variable whose mean recovery rate is RR. Figure 2 pictorially depicts the default process and the recovery values. In this exhibit, P_{dirty} denotes the dirty price (clean price plus accrued interest) for a \$1 face value of the bond.

Figure 2 indicates that if the issuer defaults, the price of the bond will be equal to its recovery rate ψ , which is a random variable. If the issuer

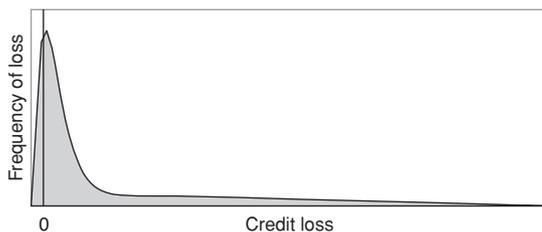


Figure 2 Bond Price Distribution Under Default Mode

does not default, the bond can be sold for a value equal to its current dirty price P_{dirty} . In this default mode framework, the price of the risky debt can be written as,

$$\tilde{P} = P_{dirty} \times I_{[\delta=0]} + \psi \times I_{[\delta=1]} \quad (17)$$

In equation (17), I is the indicator function of the default process. For the purpose of quantifying credit risk, the variable of interest to us is the credit loss resulting from holding the corporate bond. This is a random variable, which we denote $\tilde{\ell}$, and is given by,

$$\begin{aligned} \tilde{\ell} &= P_{dirty} - \tilde{P} = P_{dirty} - P_{dirty} \\ &\quad \times I_{[\delta=0]} - \psi \times I_{[\delta=1]} \end{aligned} \quad (18)$$

Taking expectations on both sides of equation (18) will allow us to compute the expected loss arising from credit risk. This is given by,

$$\begin{aligned} EL &= E(\tilde{\ell}) = P_{dirty} - P_{dirty} \\ &\quad \times (1 - PD) - E(\psi \times I_{[\delta=1]}) \end{aligned} \quad (19)$$

We note that computing expected loss requires taking the expectation of the product of two random variables, the recovery rate process and the default process. Knowledge of the joint distribution of these two random variables will be required to compute this expectation. Most credit risk models will make the simplifying assumption that these two random variables are independent. If we make this assumption, we get the equation for expected loss as given below:

$$\begin{aligned} EL &= P_{dirty} \times PD - RR \times PD \\ &= PD \times (P_{dirty} - RR) \end{aligned} \quad (20)$$

We remind the reader that P_{dirty} is the dirty price of the bond for \$1 nominal and RR in equation (20) is the mean recovery rate, which is expressed as a fraction of the face value of the debt. It is important to draw attention to the fact that the quantity $(P_{dirty} - RR)$ is different from LGD, which is defined as one minus the recovery rate. We therefore introduce the term "loss on default" (LD) to capture this new quantity as given below:

$$LD = P_{dirty} - RR \quad (21)$$

We note that loss on default will be identical to the quantity loss given default if the dirty price of the bond is equal to one. In all other circumstances these two quantities will not be the same.

Equation (20) has been derived under the assumption that the nominal exposure is one dollar. The expected loss from credit risk for a nominal exposure equal to NE is given by,

$$EL = NE \times PD \times LD \quad (22)$$

The use of the quantity LD rather than LGD in defining expected loss might raise some doubts in the minds of the reader. To clear these doubts, let us consider the following example that illustrates why LD is more appropriate than LGD in the context of bond portfolio management.

Let us consider the case of a bond portfolio manager who has the option to invest \$1 million either in a bond with dirty price \$100 (issuer A) or in a bond with dirty price \$80 (issuer B). In the latter case, the portfolio manager will buy \$1.25 million nominal value of issuer B's bond to fully invest the \$1 million. Let us assume that both issuers default within the next year and the recovery value is \$50 for \$100 face value of exposure. If the portfolio manager had invested in issuer A's bond, he would recover \$500,000 since the nominal exposure is \$1 million. On the other hand, if the portfolio manager invested in issuer B's bond, then the amount recovered would be \$625,000. This is because the portfolio manager has a nominal exposure of \$1.25 million of issuer B's bond. Clearly, from the portfolio manager's perspective the credit loss resulting from an investment in issuer A's bond is \$500,000, whereas the credit loss from an investment in issuer B's bond is only \$375,000, although both investments recovered 50% of the face value of debt. Use of the quantity LD correctly identifies the losses in both circumstances whereas the LGD definition will indicate that the losses are \$500,000 for issuer A's bond and \$625,000 for issuer B's bond. In practice, LGD is used in conjunction with the exposure amount of the transaction to identify the expected loss. How-

ever, this definition will also incorrectly identify the losses as being identical for both bonds in this example.

Unexpected Loss Under Default Mode We learned that the expected loss on the bond is the average loss that the investor can expect to incur over the course of a one-year period. However, the actual loss may well exceed this average loss over certain time periods. The potential deviation from the expected loss that the investor can expect to incur is quantified in terms of the standard deviation of the loss variable defined in equation (18). Credit market convention is to refer to the standard deviation of loss as unexpected loss. Hence, to derive the unexpected loss formula, we need to compute the standard deviation of the random variable $\tilde{\ell}$. To facilitate this computation, we will rewrite equation (18) as follows:

$$\begin{aligned} \tilde{\ell} &= P_{dirty} - P_{dirty} \times (1 - I_{[\delta=1]}) + \psi \times I_{[\delta=1]} \\ &= I_{[\delta=1]} \times (P_{dirty} - \psi) \end{aligned} \quad (23)$$

Recalling a standard result from probability theory, the variance of any random variable z can be written as the difference between the expected value of the random variable squared minus the square of its expected value. In equation form this is given by,

$$\sigma_z^2 = E(z^2) - [E(z)]^2 \quad (24)$$

We will again make the simplifying assumption that the default and recovery rate processes are independent in deriving the unexpected loss formula. Under this assumption the variance of the random variable $\tilde{\ell}$ can be written as,

$$\begin{aligned} Var(\tilde{\ell}) &= E(I_{[\delta=1]}^2) \times E[(P_{dirty} - \psi)^2] \\ &\quad - [E(I_{[\delta=1]})]^2 \times [E(P_{dirty} - \psi)]^2 \end{aligned} \quad (25)$$

Taking expected values and using the relation (24), equation (25) simplifies to,

$$Var(\tilde{\ell}) = [\sigma_{PD}^2 + PD^2] \times [\sigma_{RR}^2 + LD^2] - PD^2 \times LD^2 \quad (26)$$

In the above equation, σ_{PD}^2 is the variance of the Bernoulli random variable δ , which is given by

$$\sigma_{PD}^2 = PD \times (1 - PD) \quad (27)$$

Simplifying the terms in equation (26), it can be shown that unexpected loss, which is the standard deviation of the loss variable, is given by

$$UL = \sqrt{PD \times \sigma_{RR}^2 + LD^2 \times \sigma_{PD}^2} \quad (28)$$

The above formula for unexpected loss assumes that the nominal exposure is equal to one dollar. For a nominal exposure equal to NE, the unexpected loss at the security level will be given by

$$UL = NE \times \sqrt{PD \times \sigma_{RR}^2 + LD^2 \times \sigma_{PD}^2} \quad (29)$$

On the Independence Assumption

In deriving the expressions for expected and unexpected losses on a bond resulting from credit risk, we made the simplifying assumption that the default process and recovery rate process are independent. The question we should ask ourselves is whether this assumption is a reasonable one to make. Examining existing theoretical models on credit risk does not give us a definitive answer to this question. For instance, in Merton's framework the default process of a firm is driven by the value of the firm's assets. The risk of a firm's default is therefore explicitly linked to the variability in the firm's asset value. In this setup both the default process and the recovery rate are a function of the structural characteristics of the firm, and one can show that PD and RR are inversely related.

The reduced-form models, unlike structural models, do not condition default on the value of the firm. The default and recovery processes are modeled independently of the structural features of the firm and are further assumed to be independent of each other. This independence assumption between default and recovery processes, which is fundamental to reduced-form

models, is pervasive in all existing credit value at risk models.

Empirical results on the relationship between default and recovery values tend to suggest that these two variables are negatively correlated. The intuition behind this result is that both default rate and recovery rate may depend on certain structural factors. For instance, if a borrower defaults on the debt payments, the recovery rate will depend on the net worth of the firm's assets. This net worth, which is usually a function of prevailing economic conditions, will be lower during periods of recession. On the contrary, during recession the probability of default of issuers tends to increase. The combination of these two effects will result in a negative correlation between default and recovery rates.

Empirical research on the relationship between default and recovery rate processes suggests that a simple microeconomic interpretation based on supply and demand tends to drive aggregate recovery rate values. In particular, during high default years the supply of defaulted securities tends to exceed demand, which in turn drives secondary market prices down. Considering that RR values are based on bond prices shortly after default, the observed recovery rates are lower when there is an excess supply of defaulted securities.

In order to incorporate the empirical evidence that recovery values decrease when default rates are high, we will have to identify periods when PD is high relative to normal levels. If PD values are determined on the basis of historical average default rates as is done by rating agencies, it is difficult to distinguish between low and high default periods. On the other hand, if a structural approach is used to estimate PD values as is done by KMV Corporation, it is possible to signal periods when PD values are higher than historical average levels. This information can then be incorporated to determine the appropriate recovery rates to be used. Such an approach will amount to the use of a regime-switching model to determine the average recovery rates.

Expected Loss Under Migration Mode

To derive the formula for expected loss under default mode we took into consideration the credit event that results in the issuer defaulting on debt payments. In general, this is not the only credit event the bondholder will experience that influences the market price of the bond. More frequent are credit events that result in rating upgrades or downgrades of the bond issuer. These credit events correspond to a change in the opinion of the rating agencies concerning the creditworthiness of the issuer. Since rating changes are issuer-specific credit events, the associated bond price changes will fall under credit risk. Including price risk resulting from rating migrations in the calculation of potential credit losses is referred to as credit risk under migration mode.

In practice, the change in bond price can be both positive and negative depending on whether the rating change results in an upgrade or downgrade, respectively. However, we will use the term “credit loss” generically to refer to a change in bond price as a result of a credit event. Before proceeding to derive the formula that quantifies expected loss under migration mode, we will indicate how the price change resulting from a credit event can be estimated.

Estimating Price Changes Practitioners familiar with pricing of corporate bonds know that the issuer’s rating does not fully explain yield differentials between bonds of similar maturities. In an empirical study, Elton, Gruber, Agrawal and Mann (2002) find that pricing errors can vary from 34 cents per \$100 for Aa financials to over \$1.17 for Baa industrials. Their study suggests that the following factors have an important influence on observed price differentials between corporate bonds:

- The finer rating categories introduced by the major rating agencies when combined with the bond’s maturity
- Differences between Standard and Poor’s and Moody’s ratings for the issuer
- Differences in expected recovery rate for the bond
- The coupon on the bond due to different tax treatment
- Whether the bond is new and has traded for more than one year

These observations indicate that we cannot use generic yield curves for various rating grades to reprice bonds when the issuer’s rating changes. We will have to adopt a different technique to estimate the price risk resulting from rating changes. It is important to bear in mind that in the context of credit risk quantification, our objective is to estimate approximate price changes from rating migrations rather than to capture the correct trading price for the bond. To this end, rating migrations should result in a price change that is consistent with perceived change in the creditworthiness of the issuer.

The technique we will adopt here to estimate the change in bond price due to a rating change makes use of the current modified duration and convexity of the bond. To determine the change in yield associated with a rating change, we will assume that there exists a fixed yield spread between each rating grade that is a function of the debt issue’s seniority. These yield spreads will be taken relative to the government yield curve. If we denote modified duration of the bond by D and convexity by C , then the change in price of the bond due to a change Δy in the bond yield as a result of the rating change is given by,

$$\begin{aligned} \text{Price change} = & -P_{\text{dirty}} \times D \times \Delta y + 0.5 \\ & \times P_{\text{dirty}} \times C \times \Delta y^2 \end{aligned} \quad (30)$$

Considering that our interest is to estimate the loss resulting from the rating change to quantify credit risk, the following equation is the one that is relevant to us:

$$\Delta P = P_{\text{dirty}} \times D \times \Delta y - 0.5 \times P_{\text{dirty}} \times C \times \Delta y^2 \quad (31)$$

The advantage of such a technique is that it will retain price differentials observed in the market between bonds with similar maturity

Table 3 Example Yield Spreads for Different Rating Grades and Debt Seniority

Rating Grade	Rating Description	Senior Unsecured	Subordinated
1	Aaa / AAA	15 bp	20 bp
2	Aa1 / AA+	30 bp	40 bp
3	Aa2 / AA	45 bp	60 bp
4	Aa3 / AA-	60 bp	80 bp
5	A1 / A+	75 bp	100 bp
6	A2 / A	90 bp	120 bp
7	A3 / A-	105 bp	140 bp
8	Baa1 / BBB+	130 bp	180 bp
9	Baa2 / BBB	155 bp	220 bp
10	Baa3 / BBB-	180 bp	260 bp
11	Ba1 / BB+	230 bp	330 bp
12	Ba2 / BB	280 bp	410 bp
13	Ba3 / BB-	330 bp	480 bp
14	B1 / B+	430 bp	610 bp
15	B2 / B	530 bp	740 bp
16	B3 / B-	630 bp	870 bp
17	Caa-C / CCC	780 bp	1040 bp

and credit rating when the issuer migrates to a different rating grade. Table 3 shows the indicative yield spreads relative to government bonds for different rating grades as a function of the seniority of the debt issue. These yield spreads will be used to illustrate how the price change resulting from a rating migration can be estimated by using it in conjunction with the current duration and convexity of the bond.

Deriving Expected Loss

Unlike in the case of the default mode, the issuer can migrate to one of several rating grades under the migration mode during the course of the year. Associated with these rating migrations are discrete transition probabilities that comprise the rows of the rating transition matrix given in Table 2. In the rating migration framework, the transition probabilities represent historical averages and can be treated as deterministic variables. The random variables here are the credit losses that the bondholder incurs when the issuer rating changes. The expected value of the credit loss for a rating change from the i th grade to the k th grade is given by,

$$\Delta P_{ik} = P_{dirty} \times D \times \Delta y_{ik} - 0.5 \times P_{dirty} \times C \times \Delta y_{ik}^2 \quad (32)$$

In equation (32), Δy_{ik} denotes the yield change when the issuer rating changes from grade i to grade k . When the issuer migrates to the default state, the credit loss ΔP_{ik} will be equal to the loss on default LD. Considering that there are 18 rating grades including the default state, the expected loss under the rating migration mode for an issuer whose current credit rating is i is given by,

$$EL = \sum_{k=1}^{18} p_{ik} \times \Delta P_{ik} \quad (33)$$

In equation (33), p_{ik} denotes the one-year transition probability to migrate from rating grade i to rating grade k . The above equation quantifies the expected loss over a one-year horizon for a nominal exposure of one dollar. For a nominal exposure NE, the expected loss under migration mode is given by,

$$EL = NE \times \sum_{k=1}^{18} p_{ik} \times \Delta P_{ik} \quad (34)$$

Unexpected Loss Under Migration Mode

By definition, unexpected loss under migration mode is the standard deviation of the credit loss

variable. The loss variable under the migration mode is given by,

$$\tilde{\ell} = \sum_{k=1}^{18} p_{ik} \times \Delta \tilde{P}_{ik} \quad (35)$$

In equation (35), $\Delta \tilde{P}_{ik}$ denotes the credit loss when the credit rating changes from grade i to grade k , which is regarded as a random variable. The expected value of this random variable is ΔP_{ik} , and we shall denote its variance by σ_{ik}^2 . When k is equal to the default state, σ_{ik} will be equal to σ_{RR} , which is the standard deviation of the recovery rate. Recalling equation (24), we can write the variance of the loss variable as,

$$\begin{aligned} \text{Var}(\tilde{\ell}) &= E \left(\sum_{k=1}^{18} p_{ik} \times \Delta \tilde{P}_{ik}^2 \right) \\ &\quad - \left[E \left(\sum_{k=1}^{18} p_{ik} \times \Delta \tilde{P}_{ik} \right) \right]^2 \end{aligned} \quad (36)$$

Taking expectations and making use of the relation (24) once more, we get the following expression for the variance of the loss variable:

$$\begin{aligned} \text{Var}(\tilde{\ell}) &= \sum_{k=1}^{18} p_{ik} \times (\Delta P_{ik}^2 + \sigma_{ik}^2) \\ &\quad - \left[\sum_{k=1}^{18} p_{ik} \times \Delta P_{ik} \right]^2 \end{aligned} \quad (37)$$

If we assume that there is no uncertainty associated with the credit losses except in the default state, all σ_{ik}^2 terms in equation (37) will drop out other than σ_{RR}^2 . Making this assumption and noting that p_{ik} is equal to PD when k is the default state, the unexpected loss under migration mode for a nominal exposure NE is given by,

$$\begin{aligned} UL &= NE \\ &\times \sqrt{PD \times \sigma_{RR}^2 + \sum_{k=1}^{18} p_{ik} \times \Delta P_{ik}^2 - \left[\sum_{k=1}^{18} p_{ik} \times \Delta P_{ik} \right]^2} \end{aligned} \quad (38)$$

Numerical Example

In this section we will consider a numerical example to illustrate the computations of ex-

Table 4 Security Level Details of the Example Bond

Description	Value
Issuer rating grade	A3
Dirty price for \$1 nominal	1.0533
Nominal exposure	\$1,000,000
Modified duration	4.021
Convexity	19.75
Mean recovery rate	35%
Volatility of RR	25%

pected and unexpected losses under the default mode and migration mode. The security level details of the example we will consider are given in Table 4.

Since the mean recovery rate is assumed to be 35%, the loss on default for this security is equal to 0.7033 for one-dollar nominal exposure. The probability of default for this security is equal to 0.10%, which corresponds to the last column in row A3 of the transition matrix given in Table 3. The expected and unexpected losses in the default mode when PD = 0.001 are given below.

$$\begin{aligned} EL &= NE \times PD \times LD \\ &= 1,000,000 \times 0.001 \times 0.7033 = \$703.3 \\ UL &= NE \times \sqrt{PD \times \sigma_{RR}^2 + LD^2 \times \sigma_{PD}^2} \\ &= 1,000,000 \\ &\quad \times \sqrt{0.001 \times 0.25^2 + 0.7033^2 \times 0.001 \times (1 - 0.001)} \\ &= \$22,369.3 \end{aligned}$$

Under the migration mode, the breakdown of the calculations involved in estimating expected and unexpected losses are given in Table 5.

The expected loss under migration mode is given by,

$$\begin{aligned} EL &= NE \times \sum_{k=1}^{18} p_{ik} \times \Delta P_{ik} \\ &= 1,000,000 \times 0.003132 = \$3,132 \end{aligned}$$

The unexpected loss under migration mode is given by,

$$\begin{aligned} UL &= NE \\ &\times \sqrt{PD \times \sigma_{RR}^2 + \sum_{k=1}^{18} p_{ik} \times \Delta P_{ik}^2 - \left[\sum_{k=1}^{18} p_{ik} \times \Delta P_{ik} \right]^2} \\ &= 1,000,000 \\ &\quad \times \sqrt{0.001 \times 0.25^2 + 6.803 \times 10^{-4} - 0.003132^2} \\ &= \$27,073.8 \end{aligned}$$

Table 5 Calculation of EL and UL Under Migration Mode

Grade	P_{ik}	Δy_{ik}	ΔP_{ik}	$P_{ik} \times \Delta P_{ik}$	$P_{ik} \times \Delta P_{ik}^2$
1	0.05%	-0.90%	-0.0390	-0.000019	7.590E-07
2	0.11%	-0.75%	-0.0323	-0.000036	1.151E-06
3	0.05%	-0.60%	-0.0258	-0.000013	3.325E-07
4	0.24%	-0.45%	-0.0193	-0.000046	8.912E-07
5	1.55%	-0.30%	-0.0128	-0.000198	2.539E-06
6	8.68%	-0.15%	-0.0064	-0.000553	3.529E-06
7	75.40%	0.00%	0.0000	0.000000	0.000E+00
8	7.03%	0.25%	0.0105	0.000740	7.785E-06
9	3.83%	0.50%	0.0209	0.000801	1.676E-05
10	1.50%	0.75%	0.0312	0.000468	1.458E-05
11	0.57%	1.25%	0.0513	0.000293	1.501E-05
12	0.20%	1.75%	0.0709	0.000142	1.006E-05
13	0.23%	2.25%	0.0900	0.000207	1.864E-05
14	0.35%	3.25%	0.1267	0.000443	5.615E-05
15	0.05%	4.25%	0.1612	0.000081	1.299E-05
16	0.05%	5.25%	0.1937	0.000097	1.876E-05
17	0.01%	6.75%	0.2385	0.000024	5.688E-06
18	0.10%		0.7033	0.000703	4.946E-04
			Sum	0.003132	6.803E-04

It is useful to note here that under migration mode the expected loss is more than four times higher. The increase in the unexpected loss in migration mode is, however, only around 21% higher than the unexpected loss under default mode.

KEY POINTS

- Approaches used to determine default probabilities at the issuer level fall under two broad categories: the empirical approach that uses historical default data and public credit rating schemes; and the structural approach that uses options theory framework.
- Recovery rates on defaulted bonds vary over the business cycle and across industry sectors; and there is a negative relationship between recovery rates and probability of default.
- Credit risk for a corporate bond can be quantified in terms of the first

two moments of its loss distribution: expected loss and unexpected loss.

- Approaches to quantifying credit risk fall under two categories: those that are based on two states of the world, namely default or no default; and those that include migrations to other credit rating categories including the state of default.

REFERENCES

- Altman, E. I., Resti, A., and Sironi, A. (2001). Analyzing and explaining default recovery rates. A report submitted to The International Swaps & Derivatives Association.
- Crosbie, P. J., and Bohn, J. R. (2002). Modeling default risk. KMV.
- Elton, E. J., Gruber, M. J., Agrawal, D., and Mann, C. (2002). Factors affecting the valuation of corporate bonds. Working Paper, Stern Business School.
- Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29: 449–470.

Simulating the Credit Loss Distribution

SRICHANDER RAMASWAMY, PhD

Senior Economist, Bank for International Settlements, Basel, Switzerland

Abstract: Monte Carlo methods have become a valuable computational tool in modern finance as the increased availability of powerful computers has enhanced their efficiency. A particularly useful feature of Monte Carlo methods is that their computational complexity increases linearly with the number of variables. Moreover, they are flexible and easy to implement for a range of distributional assumptions for the underlying variables that influence the outcomes of interest. Monte Carlo methods are particularly effective for simulating credit loss distribution and for evaluating tail risk measures, and they are computationally less intensive than analytical methods.

The distribution of portfolio *credit risk* is highly skewed and has a long fat tail. Unlike the case for a normally distributed loss distribution, knowledge of the first two moments of the credit loss distribution provides little information about *tail risk*. To compute tail risk (large losses that occur with a low probability) one has to simulate the credit loss distribution using Monte Carlo techniques. In this entry we will provide a brief introduction to Monte Carlo methods and subsequently describe the computational process involved in performing a Monte Carlo simulation to generate the distribution of credit losses. Simulating the credit loss distribution is discussed under the assumption that the asset returns that drive credit events are either multivariate normal or multivariate *t* distributed. The discussion and the examples cited

in this entry assume that the credit risk arises from holding a portfolio of *corporate bonds*.

MONTE CARLO METHODS

Numerical methods known as *Monte Carlo methods* can be loosely described as statistical simulation methods that make use of sequences of random numbers to perform the simulation. The first documented account of Monte Carlo simulation dates back to the 18th century when a simulation technique was used to estimate the value π . However, it is only since the digital computer era that this technique has gained scientific acceptance for solving complex numerical problems in various disciplines. The name “Monte Carlo” was coined by Metropolis

The views expressed here are those of the author and not necessarily those of the Bank for International Settlements.

during the Manhattan Project of World War II because of the similarity of statistical simulation to games of chance symbolized by the capital of Monaco. Von Neumann laid much of the early foundations of Monte Carlo simulation that require generation of pseudo-random number generators and inverse cumulative distribution functions. The application of Monte Carlo simulation techniques to finance was pioneered by Phelim Boyle (1977) in connection with pricing of options.

It is tempting to think of Monte Carlo methods as a technique to simulate random processes that are described by a stochastic differential equation. This belief stems from the option pricing applications of Monte Carlo methods in finance where the underlying variable of interest is the evolution of stock prices that are described by a stochastic differential equation. However, this description is too restrictive because many Monte Carlo applications have no apparent stochastic content, such as the evaluation of a definite integral or inversion of a system of linear equations. In many applications of Monte Carlo methods, the only requirement is that the physical or mathematical quantity of interest to us can be described by a probability distribution function.

Monte Carlo methods have become a valuable computational tool in modern finance to price complex derivative securities and to perform value at risk calculations. An important advantage of Monte Carlo methods is that they are flexible and easy to implement. Further, the increased availability of powerful computers has enhanced the efficiency of these methods. Notwithstanding this, the method can still be slow and standard errors of estimates can be large when applied to high-dimensional problems or if the region of interest to us is not around the mean of the distribution. In such cases, we require a large number of simulation runs to estimate the variable of interest with reasonable accuracy. The standard errors on the estimated parameters can be reduced using conventional variance reduction procedures

such as control variate techniques or antithetic sampling approaches.

More recent techniques to speed up the convergence of Monte Carlo methods for high-dimensional problems make use of deterministic sequences rather than random sequences. These sequences are known by the name quasi-random sequences in contrast to the pseudo-random sequences commonly used in standard Monte Carlo methods. The advantage of using quasi-random sequences is that they generate sequences of n -tuples that fill n -dimensional space more uniformly than uncorrelated points generated by pseudo-random sequences. However, the computational advantage of quasi-random sequences diminishes as the number of variables increases beyond 30.

An important advantage of Monte Carlo methods is that the computational complexity increases linearly with the number of variables. In contrast, the computational complexity increases exponentially in the number of variables for discrete probability tree approaches for solving similar kinds of problems. This point is best illustrated by considering the problem of credit loss simulation. One approach to computing the loss distribution of a two-bond portfolio is to enumerate all possible combination of credit states this portfolio can be in one year's time. Assuming there are 18 possible credit states that each bond can be in, the two-bond portfolio could take one of 324 (18 times 18) credit states. Valuing the credit loss associated with each one of the 324 states will allow us to derive the credit loss distribution of the two-bond portfolio. If the number of bonds in the portfolio increases to 10, the total number of possible credit states will be equal to 18 to the power 10, which is equal to 3.57×10^{12} credit states. Clearly, even with such a small portfolio, it is practically impossible to enumerate all the states and compute the credit loss distribution.

If we use Monte Carlo simulation, on the other hand, the problem complexity remains the same irrespective of whether the portfolio is comprised of 2, 10, or more bonds. In each of these

cases we may wish to run several scenarios, each of which corresponds to a simulation run, and under each scenario compute the credit loss associated with the portfolio. Performing many simulation runs will allow us to compute the credit loss distribution of the bond portfolio. As the number of bonds in the portfolio increases, the computational effort involved increases linearly in the number of bonds in the portfolio.

The basic building blocks for performing Monte Carlo simulation will require a scheme to generate uniformly distributed random numbers and a suitable transformation algorithm if the probability distribution of the variable simulated is different from a uniform distribution. Most applications in finance require the generation of a normally distributed random variable. To simulate such a random variable, the standard transformation techniques used are either the Box-Muller method or the inverse cumulative normal method. If the simulated random variables are greater than one, we need methods to generate correlated random numbers that model the relationship between the variables.

Credit Loss Simulation

At the security level, credit loss arises from credit events that include rating migrations and outright default. As these credit events are associated with changes in perceptions about an obligor's ability to make the contractual debt payments, one needs to identify variables that influence the obligors' ability to pay. The variable that is often used in practice is the asset returns of the obligor. The motivation for using asset returns is that changes in asset values of a firm influence its solvency position. When asset values fall below outstanding liabilities, the firm is no longer considered solvent. But other thresholds based on rating transition probabilities can be derived and used to infer how changes in asset values will influence credit ratings. Simulating asset returns and checking their values against these thresholds

will allow us to signal credit events, which can then be used to estimate the credit loss for a particular simulation run.

Computing portfolio credit risk requires extending the above approach to model joint rating migrations, which in turn requires modeling the comovement of asset returns of different obligors. Considering that the marginal distribution of asset returns is assumed to be normal in Merton's option pricing framework (Merton, 1974), one can make a simplifying assumption that the joint distribution of asset returns is multivariate normal. The joint evolution of the asset returns of the obligors under the multivariate normal distribution will signal how the value of the portfolio evolves, or equivalently, what the credit loss on the portfolio will be. The distribution of obligor asset returns under the multivariate normal distribution can be generated using Monte Carlo simulation. This will allow us subsequently to compute the loss distribution of the bond portfolio resulting from credit events.

The description given above provides the basic intuition behind the use of Monte Carlo simulation for computing the credit loss distribution. In the context of its intended use here, the Monte Carlo simulation technique can be described as a computational scheme that utilizes sequences of random numbers generated from a given probability distribution function to derive the distribution of portfolio credit loss. The distribution of portfolio credit loss can be computed both under the default mode, which only considers whether the obligor is solvent or not, and under the migration mode that includes credit events arising from rating changes. Consequently, to compute the credit loss under the default mode, we only need to consider the loss resulting from obligor default; whereas under the migration mode, we have to compute the credit loss associated with rating migrations in addition to the credit loss resulting from obligor default.

To generate the credit loss for one run of the Monte Carlo simulation, we need to go

through three computational steps described below.

1. Simulate correlated random numbers that model the joint distribution of asset returns of the obligors in the portfolio.
2. Infer the implied credit rating of each obligor based on simulated asset returns.
3. Compute the potential loss in value based on the implied credit rating, and in those cases where the asset return value signals an obligor default, compute a random loss on default value by sampling from a beta distribution function.

Repeating the above simulation run many times and computing the credit loss under each simulation run will allow us to generate the distribution of portfolio credit loss under the migration mode. If we are only interested in the credit loss distribution under the default mode, we can compute this by setting credit loss associated with rating migrations to zero in the simulation run. In the following sections we will briefly describe the computational steps that are required to generate the credit loss distribution.

Generating Correlated Asset Returns

We briefly described earlier the steps involved in simulating the credit loss distribution for a bond portfolio. As the first step, we mentioned that correlated random numbers that model the joint distribution of asset returns have to be simulated. An immediate question that will arise in our minds is whether the obligor-specific means and standard deviations of asset returns have to be taken into account in the simulations. The simple answer to this question is no. This is because the simulated asset returns will be compared against the rating migration thresholds, which are computed under the assumption that asset returns are standardized normal random variables. As a result, the obligor-specific mean and standard deviation of asset returns are not required for

simulating the loss distribution. Hence, we will assume that obligor asset returns are standard normal random variables (having mean zero and standard deviation equal to one). Under this assumption, the Monte Carlo simulation method will require generating a sequence of random vectors that are sampled from a standardized multivariate normal distribution.

Many standard numerical packages provide routines to generate sequences of random vectors sampled from a multivariate normal distribution. Although the details of the implementation are not discussed here, we will briefly outline the numerical procedure commonly used to generate sequences of multivariate normal random vectors. Let us assume that the multivariate normal random vector has a mean vector \vec{a} and covariance matrix C . Covariance matrices have the property that they are symmetric and positive definite (meaning all its eigenvalues are greater than zero). Given such a matrix, it is possible to find a unique lower triangular matrix L such that,

$$L L^T = C \quad (1)$$

The matrix L is referred to as the Cholesky factor corresponding to the positive definite matrix C . Once the Cholesky factor is determined, generating a sequence of random vectors with the desired multivariate distribution only requires generating a sequence of independent standard normal random variables. If \vec{x} denotes the vector of independent standard normal random variables, the vector \vec{r} with the desired multivariate normal distribution can be constructed as follows:

$$\vec{r} = \vec{a} + L \vec{x} \quad (2)$$

The sequence of random vectors \vec{r} that are generated will have the property that their joint distribution is multinormal with mean vector \vec{a} and covariance matrix C .

It is useful to note here that by setting the mean vector \vec{a} to zero and the covariance matrix equal to the correlation matrix, we can generate a sequence of random vectors whose

Table 1 Rating Transition Probabilities and z-Thresholds

Transition to Rating	Transition Probabilities		z-Threshold (Gaussian)		z-Threshold (Student's <i>t</i>)	
	A2-rated	A3-rated	A2-rated	A3-rated	A2-rated	A3-rated
Aaa	0.05	0.05	3.28	3.28	5.04	5.04
Aa1	0.06	0.11	3.05	2.95	4.43	4.15
Aa2	0.30	0.05	2.64	2.86	3.49	3.96
Aa3	0.80	0.24	2.25	2.61	2.77	3.43
A1	5.57	1.55	1.49	2.05	1.66	2.45
A2	80.75	8.68	-1.15	1.24	-1.24	1.35
A3	7.48	75.40	-1.65	-1.08	-1.86	-1.16
Baa1	2.99	7.03	-2.05	-1.48	-2.45	-1.65
Baa2	0.83	3.83	-2.27	-1.87	-2.79	-2.18
Baa3	0.41	1.50	-2.43	-2.15	-3.08	-2.61
Ba1	0.29	0.57	-2.60	-2.33	-3.40	-2.90
Ba2	0.11	0.20	-2.69	-2.41	-3.58	-3.05
Ba3	0.12	0.23	-2.81	-2.54	-3.86	-3.28
B1	0.03	0.35	-2.86	-2.85	-3.96	-3.96
B2	0.07	0.05	-2.98	-2.94	-4.25	-4.15
B2	0.03	0.05	-3.06	-3.06	-4.43	-4.43
Caa-C	0.03	0.01	-3.16	-3.09	-4.67	-4.50
Default	0.08	0.10	-1000	-1000	-1000	-1000

joint distribution is standardized multivariate normal. Since the joint distribution of obligor asset returns was assumed to be standardized multivariate normal, this sequence of random vectors will be the one of interest to us.

Inferring Implied Credit Rating

The next step in the credit loss simulation process is to infer the credit rating of the various obligors in the portfolio as implied by the simulated asset return vector. In order to do this, we need to determine the thresholds against which the asset returns will be compared to identify rating changes or obligor default. To illustrate how these thresholds can be determined, let us consider an obligor that has a current credit rating of A1. (Moody's rating categories are used here to denote the credit rating of an obligor.) Let $p_{A1, Aaa}$ denote the probability of transitioning to the credit rating Aaa. Under the assumption that the asset returns of the obligor are normally distributed, the credit event that signals the obligor rating migration from A1 to Aaa will occur when the standardized asset returns of the obligor exceed the threshold $z_{A1, Aaa}$. This

threshold can be determined by solving the following integral equation:

$$p_{A1, Aaa} = \frac{1}{\sqrt{2\pi}} \int_{z_{A1, Aaa}}^{\infty} \exp\left(-\frac{1}{2} x^2\right) dx \quad (3)$$

A rating transition of this obligor from A1 to Aa1 will occur if the asset return falls between the thresholds $z_{A1, Aaa}$ and $z_{A1, Aa1}$. The threshold $z_{A1, Aa1}$ can be determined by solving the following integral equation:

$$p_{A1, Aa1} = \frac{1}{\sqrt{2\pi}} \int_{z_{A1, Aa1}}^{z_{A1, Aaa}} \exp\left(-\frac{1}{2} x^2\right) dx \quad (4)$$

One can extend this sequential rule to determine the thresholds for migrating to other rating grades. We note here that these z-thresholds are a function of the current credit rating of the obligor. Table 1 shows the rating transition probabilities and the corresponding z-thresholds for two different obligor credit ratings when the asset returns are assumed to be Gaussian (normal distribution).

Let us consider the two-bond portfolio given in Table 2 to illustrate the specific steps involved

Table 2 Security Level Details for the Two-Bond Portfolio

Description	Bond 1	Bond 2
Issuer rating grade	A3	A2
Dirty price for \$1 nominal	1.0533	1.0029
Nominal exposure	\$1,000,000	\$1,000,000
Modified duration	4.021	3.747
Convexity	19.75	16.45
Mean recovery rate	35%	35%
Volatility of recovery rate	25%	25%

in computing the credit loss from one simulation run for this portfolio. Suppose during one draw from a bivariate normal distribution the random asset returns are, respectively, 2.5 for bond 1 and -3.5 for bond 2. Given the initial issuer rating of A3 for bond 1, one can infer from the z -threshold values for A3-rated issuers in Table 1 that an asset return value of 2.5 implies a credit rating change of the issuer to an A1 rating. Similarly, one can infer from Table 1 that an asset return value of -3.5 for an A2-rated issuer will imply that the issuer defaults on the outstanding debt. Proceeding in this manner, the implied credit rating of the debt issuers in the two-bond portfolio for every simulation run can be derived on the basis of the z -threshold values in Table 1.

For a general n -bond portfolio, the implied credit rating of the debt issuers for each simulation run can be similarly determined. It is important to note here that the number of obligors in an n -bond portfolio will be less than or equal to n . In the case where there are fewer than n obligors, credit rating changes should be identical for all bonds issued by the same obligor in any simulation run. This has the implication that the dimension of the simulated asset return vector should be equal to the number of obligors or debt issuers in the bond portfolio.

Computing Credit Loss

Once the implied rating changes for the obligors are determined for the simulated asset re-

turn vector, the corresponding credit loss associated with such implied rating changes could be determined. It is important to note here that we generically refer to the price change resulting from the rating change as a loss although a credit improvement of the obligor will result in a price appreciation for the bond. The price change of a bond as a result of a rating change for the bond issuer will be a function of the change in the yield spreads and the maturity of the bond. Assuming that our interest is to estimate the credit loss due to a change in the bond's mark to market value as a result of the rating change, we would want to know at what time horizon the bond's price has to be marked to market. If we were to compute the worst-case loss scenario, it would correspond to a rating change of the obligor during the next trading day. In this case, the current trading price of the bond and its risk parameters, duration, and convexity serve to characterize the credit loss. The credit loss resulting from a rating change from the i th grade to the k th grade will be a function of the change in the bond yield and is given by,

$$\Delta P_{ik} = P_{dirty} \times D \times \Delta y_{ik} - 0.5 \times P_{dirty} \times C \times \Delta y_{ik}^2 \quad (5)$$

In equation (5), P_{dirty} is the dirty price of the bond (accrued interest plus traded price), Δy_{ik} is the yield change when issuer rating changes from grade i to grade k , D is the modified duration of the bond, and C the convexity. When the issuer migrates to the default state, the credit loss will be equal to the dirty price P_{dirty} minus the recovery rate.

To illustrate the credit loss computation, let us again focus on the two-bond portfolio example. In this example, the asset return value signaled an upgrade to an A1 rating from the current rating of A3 for bond 1. Suppose the change in the yield spread associated with this rating change is -30 basis points. Then, substituting the various parameter values into equation (5), the credit loss for \$1 million notional amount

held of bond 1 is given by,

$$\begin{aligned} \text{Credit loss} &= 1,000,000 \times [1.0533 \times 4.021 \\ &\quad \times (-0.003) - 0.5 \times 1.0533 \times 19.75 \\ &\quad \times (-0.003)^2] = -\$12,799.6 \end{aligned}$$

We note here that the negative sign associated with the credit loss is suggesting that this rating change results in a profit rather than a loss.

For bond 2, the simulated asset return value of -3.5 implies default of the obligor. In this case, we must find a random loss on default, which will be a function of the assumed recovery rate distribution. Many credit risk models assume the recovery rate process to have a beta distribution with mean μ and standard deviation σ . Given the values for μ and σ , the parameters α and β that define the beta distribution with the desired mean and standard deviation can be computed as given below:

$$\alpha = \frac{\mu^2(1-\mu)}{\sigma^2} - \mu \quad (6)$$

$$\beta = \frac{\alpha}{\mu} - \alpha \quad (7)$$

For the bond in question, let us assume the mean recovery rate to be $\mu = 35\%$ and the standard deviation of the recovery rate to be $\sigma = 25\%$. Corresponding to these recovery rate values, the parameters of the beta distribution function are $\alpha = 0.924$ and $\beta = 1.716$.

The random recovery rate for bond 2 for the simulation run is determined by drawing a random number from a beta distribution with α and β parameter values as above. Let us assume that the simulated recovery value is 40% for bond 2. The implied loss on default for the bond that trades at a dirty price of \$1.0533 is then equal to 0.6533 (bond dirty price minus the recovery value). The credit loss arising from bond 2 for this simulation run will be equal to the nominal exposure times the loss on default, which is equal to \$653,300.

For the two-bond portfolio, the total credit loss for this simulation run is the sum of the two losses. If this simulation run corresponds to the i th run, the portfolio credit loss under the

i th simulation run, denoted ℓ_i , is given by,

$$\ell_i = -\$12,799.6 + \$653,300 = \$640,500.4$$

It is important to emphasize here that for a general n -bond portfolio, all bonds of a particular issuer should have the same recovery value for any one simulation run if they have the same seniority. This information must be taken into account when simulating the credit loss distribution of a general n -bond portfolio.

Computing Expected and Unexpected Loss

The above procedure outlined how the portfolio credit loss can be computed for one simulation run. By repeating the simulation run N times where N is sufficiently large, the distribution of the credit losses can be generated. Given the simulated loss distribution, one can compute various risk measures of interest. For instance, the expected and unexpected credit loss (the first two moments of the loss distribution) using the simulated loss data can be computed as follows:

$$EL_P = \frac{1}{N} \sum_{i=1}^N \ell_i \quad (8)$$

$$UL_P = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (\ell_i - EL_P)^2} \quad (9)$$

To reduce the standard error of the estimated portfolio expected loss, it is common practice to perform antithetic sampling when performing the Monte Carlo simulation. The idea behind antithetic sampling technique is that when random samples are drawn from a symmetric distribution, sampling errors can be avoided if the antithetic or symmetric part of the random sample is also drawn. This will ensure that the empirical mean of the random samples is equal to the mean of the distribution function from which the samples are drawn. Including the antithetic part of the samples will double the total number of simulation runs.

Importance Sampling

The Monte Carlo simulation technique described so far is based on random sampling. In such a sampling process, the probability of any value being generated is proportional to the probability density at that point. This property will have the effect of generating asset return values in the simulations that tend to cluster around the mean of the normal distribution function. Rating migrations and obligor defaults, however, are events that are driven by asset return values that deviate significantly from the mean of the normal distribution. The implication is that a significant proportion of the simulation runs will not trigger any credit events. If our intention is to compute the expected and unexpected loss of the portfolio from the simulations, random sampling will be the appropriate method to use. If, on the other hand, we expect to compute risk measures associated with tail events from the simulated data, random sampling will be inefficient.

If our primary intention of performing Monte Carlo simulations is to compute tail risk measures (to be discussed in the next section), we can improve the simulation efficiency through importance sampling (see Glasserman and Li, 2005). Simulation efficiency in our context refers to the number of simulation runs required to compute the risk measure of interest for a specified standard error of the estimate. Importance sampling artificially inflates the probability of choosing random samples from those regions of the distribution that are of most interest to us. This would mean that our sampling process is biased in such a manner that a large number of credit events are simulated relative to what would occur in practice. In the Monte Carlo simulation terminology, the adjustment made to the probability of a particular point being sampled is referred to as its importance weight. To estimate the true probability distribution of the simulated losses when performing importance sampling, we have to restore the actual probability of each sample by multiplying it by the inverse of its importance weight. In

practice, when the number of obligors in the portfolio is large (this is usually true for the benchmark portfolio), performing importance sampling will lead to improved computational efficiency.

Tail Risk Measures

The discussions so far focused on how the mean (expected loss) and standard deviation (unexpected loss) of the credit loss distribution for a corporate bond portfolio can be computed from the simulations. If the distribution of credit losses is normally distributed, standard deviation can be interpreted as the maximum deviation around the mean that will not be exceeded with a 66% level of confidence. Since the credit loss distribution is not normal, a similar interpretation to the standard deviation of credit loss does not hold. In most cases, computing the probability of incurring a large credit loss on a corporate bond portfolio using unexpected loss information is usually not possible.

In general, a major preoccupation of most corporate bond portfolio managers is to structure the portfolio so as to minimize the probability of large losses. To do this an estimate of the potential downside risk of the portfolio becomes a key requirement. Computing any downside risk measure requires an estimate of the probability mass associated with the tail of the loss distribution. If the simulated credit loss distribution is available, it is quite easy to derive appropriate tail risk measures of interest. For a corporate bond portfolio, the tail risk measures of interest are credit value at risk and expected shortfall risk. Both these risk measures are discussed below, and the method to compute these measures using the simulated loss distribution is also indicated.

Credit Value at Risk

Credit value at risk (CVaR) is a tail risk measure that quantifies the extreme losses arising from credit events that can occur at a prespecified

level of confidence over a given time horizon. In practical terms, CVaR provides an estimate of the maximum credit loss on a portfolio, which could be exceeded by a probability p . Without loss of generality, it will be assumed that this probability is expressed in percentage. If the probability p is chosen to be sufficiently small, one can expect that the credit loss will not exceed the CVaR amount at a high confidence level given by $(100 - p)\%$. Stated differently, CVaR at a confidence level of $(100 - p)\%$ refers to the maximum dollar value of loss that will only be exceeded $p\%$ of the time over the given time horizon. Since losses from credit risk are measured over a one-year horizon, the CVaR measure we will compute also relates to a one-year time horizon.

In order to compute CVaR to quantify the tail risk of the credit loss distribution in a corporate bond portfolio, we need to specify the confidence level at which it should be determined. Within the framework of economic capital allocation, CVaR is usually measured at a confidence level that reflects the solvency standard of the institution in question. For instance, the solvency standard of an AA-rated institution is typically 99.97%, and hence, CVaR will be computed at this confidence level. From a portfolio management perspective, however, the confidence level of interest for CVaR estimate would typically be much lower. The motivation for this is that portfolio managers have to provide monthly performance reports to clients, and return deviations over this period need to be explained. In this case, estimating CVaR at a confidence level of 91.6% would imply that the underperformance relative to the benchmark will exceed the monthly CVaR estimate once during the year on average if monthly performance reporting is used. In this case, the CVaR estimate provides useful information to the portfolio manager and the client in terms of both the return surprises one could expect and also to actually observe it happen.

Motivated by the above observation, we will choose the confidence level for the CVaR esti-

mate to be 90%. At this level of confidence, the portfolio manager can expect the credit losses to exceed the monthly CVaR estimate for one reporting period during the year. Once the confidence level for CVaR is specified, estimating CVaR from the simulated loss distribution is quite simple. If, for instance, the number of simulation runs is equal to 10,000, then the 90% CVaR will be equal to the 1,000th worst-case credit loss. Assuming that the simulated credit losses are sorted in an ascending order of magnitude, the credit loss corresponding to the 9,000th row in the sorted data will be the CVaR at 90% confidence level for 10,000 simulation runs.

Considering that standard practice in portfolio management is to report risk measures relative to the current market value of the portfolio, we will introduce the term “percentage credit value at risk.” If M_p denotes the current market value of the portfolio, the percentage CVaR at 90% confidence level is defined as,

$$\%CVaR_{90\%} = \frac{CVaR_{90\%}}{M_p} \quad (10)$$

Expected Shortfall

Although CVaR is a useful tail risk measure, it fails to reflect the severity of loss in the worst-case scenarios in which the loss exceeds CVaR. In other words, CVaR fails to provide insight as to how far the tail of the loss distribution extends. This information is critical if the portfolio manager is interested in restricting the severity of the losses in the worst-case scenarios under which losses exceed CVaR. In order to better motivate this point, Figure 1 shows the credit loss distribution for two portfolios that have identical CVaR at the 90% level of confidence.

Examining Figure 1 it is clear that although both portfolios have identical CVaR at the 90% confidence level, the severity of the worst-case losses that exceed the 90% confidence level are lower for portfolio 1 compared to portfolio 2. This example suggests that in order to

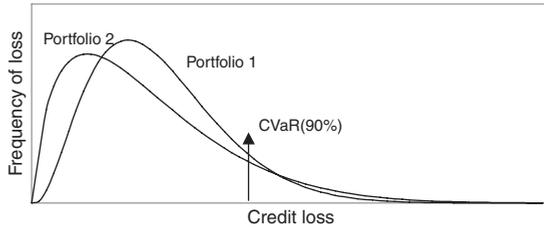


Figure 1 Credit Loss Distribution for Two Portfolios

investigate whether portfolio credit risk is well diversified, it is not sufficient if we only examine the tail probability beyond some confidence level. Examining the loss exceedence beyond the desired confidence level at which CVaR is estimated is important to gauge the loss severity in the tail part of the loss distribution.

One such risk measure that provides an estimate of the loss severity in the tail part of the loss distribution is the *expected shortfall* (ES), which is sometimes also referred to as conditional VaR. Similar to CVaR, expected shortfall requires specifying a confidence level and a time horizon. Considering that ES is usually used in conjunction with CVaR, the confidence level should be chosen as 90% and the time horizon one year. A simple interpretation of ES is that it measures the average loss in the worst $p\%$ scenarios where $(100 - p)\%$ denotes the confidence level at which CVaR is estimated. In mathematical terms, expected shortfall can be defined as the conditional expectation of that part of the credit loss that exceeds the CVaR limit. The interpretation of ES as conditional VaR follows from this definition. If $\tilde{\ell}$ denotes the loss variable, ES can be defined as given below:

$$ES = E[\tilde{\ell} \mid \tilde{\ell} > CVaR] \quad (11)$$

Given the simulated loss distribution of the portfolio, computing expected shortfall risk is quite simple. Let ℓ_i denote the simulated credit loss distribution for the i th simulation run, and let us assume that the losses are sorted in ascending order. If the number of simulation runs is equal to N , the relevant equation to compute

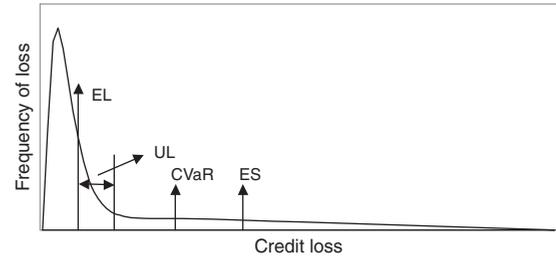


Figure 2 Various Risk Measures for Portfolio Credit Risk

ES at the 90% confidence level from the simulations is given below:

$$ES_{90\%} = \frac{1}{(1 - 0.9)N} \times \sum_{i=0.9N+1}^N \ell_i \quad (12)$$

The percentage ES at 90% confidence level is defined as,

$$\%ES_{90\%} = \frac{ES_{90\%}}{M_p} \quad (13)$$

Figure 2 shows the various credit risk measures presented here that can be computed from the simulated loss data.

Relaxing the Normal Distribution Assumption

A growing body of empirical studies conducted on financial time series data suggests that returns on traded financial instruments exhibit volatility clustering and extreme movements that are not representative of a normally distributed random variable. Another commonly observed property of financial time series is that during times of large market moves, there is greater degree of comovement of returns across many firms compared to those observed during normal market conditions. This property, usually referred to as tail dependence, captures the extent to which the dependence (or correlation) between random variables arises from extreme observations. Stated differently, for a given level of correlation between the random variables a multivariate distribution with tail dependence has a much greater tendency to generate

simultaneous extreme values for the random variables in contrast to those distributions that do not have this property.

A multivariate normal distribution does not exhibit tail dependence. The dependence or correlation structure exhibited between the random variables in a multivariate normal distribution arises primarily from comovements of the variables around the center of the distribution. As a consequence, contagion or herding behavior commonly observed in financial markets is difficult to model within the framework of multivariate normal distributions. In order to capture contagion and herding behavior in financial markets, distributions that exhibit tail dependence should be used to model financial variables of interest. In the context of credit risk modeling, contagion effects would result in greater comovement of asset returns across firms during periods of recession leading to higher probability of joint defaults. If we model the joint distribution of asset returns to be multivariate normal, we will fail to capture the effects of contagion in the aggregate portfolio credit risk measures we compute. In the next section we relax the assumption that the distribution of asset returns is multivariate normal.

Student's t Distribution

Among the class of distribution functions that exhibit tail dependence, the family of multivariate normal mixture distributions, which include *Student's t distribution* and generalized hyperbolic distribution, is an interesting alternative. This is because normal mixture distributions inherit the correlation matrix of the multivariate normal distribution. Hence, correlation matrices for normal mixture distributions are easy to calibrate.

Formally, a member of the m -dimensional family of variance mixtures of normal distributions is equal in distribution to the product of a scalar random variable s and a normal random

vector \vec{u} having zero mean and covariance matrix Σ . The scalar random variable s is assumed to be positive with finite second moment and independent of \vec{u} . If \vec{x} denotes a random vector having a multivariate normal mixture distribution, our definition leads to the following equation:

$$\vec{x} = s \cdot \vec{u} \quad (14)$$

Since normal mixture distributions inherit the correlation matrix of the multivariate normal distribution, we have the following relationship:

$$\text{Corr}(x_i, x_k) = \text{Corr}(u_i, u_k) \quad (15)$$

The random vector \vec{x} will have multivariate t distribution with ν degrees of freedom if the scalar random variable s is defined as below:

$$s = \sqrt{\frac{\nu}{\omega}} \quad (16)$$

In equation (16), ω is a chi-square distributed random variable with ν degrees of freedom. For $\nu > 2$, the resulting Student's t distribution will have zero mean vector and covariance matrix $\frac{\nu}{\nu-2}\Sigma$. The Student's t distribution has the property that as ν increases, the distribution approaches a normal distribution. In fact, for values of ν greater than 25, it is difficult to distinguish between a normal distribution and t distribution. In a multivariate setting, as ν decreases, the degree of tail dependence between the random variables will increase. For financial time series, ν is typically around 4 (Platen and Sidorowicz, 2007).

An important distinction between the t distribution and the normal distribution is that uncorrelated random variables are mutually independent, whereas the components of multivariate t are in general dependent even if they are uncorrelated. In modeling credit risk, this property makes it possible to capture comovements of asset returns between firms in extreme market situations even if the asset returns exhibit little or no correlation under normal market conditions.

In the univariate case, the probability density function of the Student's t distribution with ν degrees of freedom has the following functional form:

$$f_\nu(x) = \frac{\Gamma((\nu + 1)/2)}{\sqrt{\nu\pi} \times \Gamma(\nu/2)} \times \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2} \quad (17)$$

In equation (17) $\Gamma(\cdot)$ is the gamma function, which is given by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \quad (18)$$

Loss Simulation Under Multivariate t Distribution

The steps involved in simulating the credit loss distribution when asset returns are multivariate t follow the same procedure as for the multivariate normal case. Instead of generating the sequence of correlated asset returns from a multivariate normal distribution, we now have to generate this sequence from a multivariate t distribution. The next step will involve inferring the credit rating change of the various obligors in the portfolio as implied by the simulated asset return vector. To do this, we need to determine the thresholds against which the asset returns will be compared to identify rating changes or obligor default. These z -thresholds have to be calibrated to correspond to the Student's t distribution. Specifically, the integrand for computing the z -thresholds will be the Student's t density function.

For purpose of illustration, let us consider an obligor that has a current credit rating of A1. Let $p_{A1,Aaa}$ denote the probability of transitioning to the credit rating Aaa. Under the assumption that the asset returns of the obligor are t -distributed, the credit event that signals the obligor rating migration from A1 to Aaa will occur when the asset returns of the obligor exceed the threshold $z_{A1,Aaa}$. This threshold can be determined by solving the following integral

equation:

$$p_{A1,Aaa} = \frac{\Gamma((\nu + 1)/2)}{\sqrt{\nu\pi} \times \Gamma(\nu/2)} \int_{z_{A1,Aaa}}^\infty \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2} dx \quad (19)$$

A rating transition of this obligor from A1 to Aa1 will occur if the asset return falls between the thresholds $z_{A1,Aaa}$ and $z_{A1,Aa1}$. The threshold $z_{A1,Aa1}$ can be determined by solving the following integral equation:

$$p_{A1,Aa1} = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi} \times \Gamma(\nu/2)} \int_{z_{A1,Aa1}}^{z_{A1,Aaa}} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2} dx \quad (20)$$

One can extend this sequential rule to determine the thresholds for migrating to other rating grades. Table 1 shows the z -threshold values computed using the rating transition probabilities for A2- and A3-rated obligors when the asset returns are t -distributed.

The rest of this section discusses the procedure to generate a sequence of random vectors from a multivariate t distribution. Following the discussion earlier in this entry, a random vector with multivariate t distribution having ν degrees of freedom can be derived by combining a chi-square random variable with ν degrees of freedom and a random vector that is normally distributed and independent of the chi-square random variable. This procedure will allow us to generate a sequence of multivariate t -distributed random vectors with ν degrees of freedom.

To generate a sequence of chi-square distributed random variables, the standard procedure is to make use of the relationship between chi-square distribution and gamma distribution. A random variable x is said to have gamma distribution if its density function is defined as below:

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, & x > 0 \\ 0, & x \leq 0 \end{cases} \quad (21)$$

In equation (21) $\alpha > 0$ and $\beta > 0$ are the parameters of the gamma distribution, and $\Gamma(\alpha)$

is the gamma function given by equation (18). The chi-square distribution with ν degrees of freedom is a special case of the gamma distribution with parameter values $\alpha = \nu/2$ and $\beta = 2$.

Given the above relationship between gamma and chi-square distribution, a sequence of random variables having chi-square distribution with ν degrees of freedom can be generated by sampling from a gamma distribution with parameter values $\alpha = \nu/2$ and $\beta = 2$. Most standard software packages provide routines to generate random sequences from a gamma distribution. Hence, we will not discuss the details concerned with generating such a sequence of random variables.

To summarize, the following are the steps involved in generating an n -dimensional sequence of multivariate t distributed random variables with ν degrees of freedom.

Step 1: Compute the Cholesky factor L of the matrix C where C is the $n \times n$ asset return correlation matrix.

Step 2: Simulate n independent standard normal random variates z_1, z_2, \dots, z_n and set $\tilde{u} = L\tilde{z}$.

Step 3: Simulate a random variate ω from chi-square distribution with ν degrees of freedom that is independent of the normal random variates and set $s = \frac{\sqrt{\nu}}{\sqrt{\omega}}$.

Step 4: Set $\tilde{x} = s \cdot \tilde{u}$ which represents the desired n -dimensional t variate with ν degrees of freedom and correlation matrix C .

Repeating the steps 2 to 4 will allow us to generate the sequence of multivariate t -distributed random variables.

Computing the credit loss for the two-bond portfolio in Table 2 will require comparing the asset return values under each simulation run against the z -thresholds given in Table 1 to trigger rating migrations and defaults for the obligors in the two-bond portfolio. On the basis of the implied rating changes assigned to the obligors

using simulated asset returns, the credit loss for each simulation run can be calculated. The rest of the steps involved in computing the credit risk measures of interest from the simulated loss distribution are identical to the ones for the normal distribution case.

KEY POINTS

- Monte Carlo methods provide a flexible tool to simulate credit loss distribution and are relatively simple to implement.
- To simulate the credit loss distribution under the rating migration mode, rating transition probabilities have to be transformed into corresponding z -thresholds for the assumed distribution function for the asset returns.
- Simulating multivariate t random vectors requires appropriately scaling the sequence of multivariate normal vectors by another sequence of chi-square random variables that are uncorrelated with the normal random vectors.
- From the simulated loss distribution, various tail risk measures of interest can be computed.
- Using techniques such as importance sampling can significantly reduce the standard errors of tail risk measures for a given number of simulation runs.

REFERENCES

- Boyle, P. (1977). Options: A Monte Carlo approach. *Journal of Financial Economics* 4, 4: 323–338.
- Glasserman, P., and Li, J. (2005). Importance sampling for portfolio credit risk. *Management Science* 51, 11: 1643–1656.
- Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29: 449–470.
- Platen, E., and Sidorowicz, R. (2007). Empirical evidence on Student-t log-returns of diversified world stock indices. Research Paper 194, Quantitative Finance Research Centre, University of Technology, Sydney.

Managing Credit Spread Risk Using Duration Times Spread (DTS)

ARIK BEN DOR, PhD

Managing Director, Barclays

LEV DYNKIN, PhD

Managing Director, Barclays

JAY HYMAN, PhD

Managing Director, Barclays, Tel Aviv

Abstract: Extensive empirical research has shown that the spread volatility of credit securities is linearly proportional to their level of spread. This finding holds true across corporate and sovereign issuers, for both cash and credit default swaps. A superior measure of spread risk for credit securities is the product of spread duration and spread, a measure referred to as duration times spread (DTS). DTS measures the sensitivity of the price of a bond to relative changes in spread, which are much more stable through time and cross-sectionally than absolute spread volatilities. DTS allows for better risk projection, hedging, replication, and portfolio construction.

The traditional presentation of the asset allocation in a portfolio or a benchmark is in terms of percentage of market value. For fixed-income portfolios, it is widely recognized that this is not sufficient, as differences in durations can cause two portfolios with the same market weight allocations to have very different exposures to macro-level risks. Market practices have evolved to address this issue. A common approach to structuring a fixed-income portfolio or comparing it to a benchmark is to partition it into homogeneous market cells comprised of securities with similar characteristics. Many fixed-income portfolio managers have become

accustomed to expressing their cell allocations in terms of contributions to duration—the product of the percentage of portfolio market value in a given market cell and the average duration of securities comprising that cell. This represents the sensitivity of the portfolio to a parallel shift in yields across all securities within this market cell. For credit portfolios, the corresponding measure would be contributions to spread duration, measuring the sensitivity to a parallel shift in spreads. Determining the set of active spread duration bets to different market cells and issuers is one of the primary decisions made by credit portfolio managers.

Yet all spread durations were not created equal. Just as one could create a portfolio that matches the benchmark exactly by market weights, but clearly takes more *credit risk* (e.g., by investing in the longest duration credits within each cell), one could match the benchmark exactly by spread duration contributions and still take more credit risk—by choosing the securities with the widest spreads within each cell. These bonds presumably trade wider than their peer groups for a reason—that is, the market consensus has determined that they are more risky—and they are often referred to as high beta, because their spreads tend to react more strongly than the rest of the market to a systematic shock. We found strong empirical evidence that this relation takes on a nearly perfect linear form: Spread changes are linearly proportional to spread levels at the start of the period.

Based on the linear relation between spread level and the volatility of spread changes, we have advocated since 2005 a new measure of risk sensitivity that utilizes spreads as a fundamental part of the credit portfolio management process. To reflect the view that higher spread credits represent greater exposures to sector-specific risks, we represent sector exposures by contributions to *duration times spread (DTS)*, computed as the product of market weight, spread duration, and spread. For example, an overweight of 5% to a market cell implemented by purchasing bonds with a spread of 80 basis points (bps) and spread duration of three years would be equivalent to an overweight of 3% using bonds with an average spread of 50 bps and spread duration of eight years.

The shift from spread duration exposures to DTS exposures as the measure of market risk sensitivity embraces a different paradigm for credit spread movement—in the form of relative spread changes rather than parallel shifts in spread. The introduction of the DTS paradigm was motivated by an extensive empirical study using over 560,000 monthly observations of individual corporate bonds spreads, spanning the

period of September 1989 to January 2005.¹ The analysis showed that changes in spreads are not parallel, but rather depend on the level of spread. Specifically, spread change volatility (both systematic and idiosyncratic) was shown to be linearly proportional to spread level for both investment-grade and high-yield credit securities, irrespective of the sector, duration, or time period. Subsequent studies indicated that the results were not confined to the realm of U.S. corporate bonds, but also extend to other spread asset classes with a significant default risk such as credit default swaps, European corporate and sovereign bonds, and emerging market sovereign debt denominated in U.S. dollars.² Furthermore, even from a theoretical standpoint structural credit risk models such as Merton (1974) imply a near-linear relation between spread level and volatility.³

The DTS concept has many implications for portfolio managers, both in terms of the way they manage exposures to industry and credit quality factors (systematic risk) and in terms of their approach to issuer exposures (nonsystematic risk). After a short review of the DTS concept and the empirical evidence supporting it, we discuss how it can help investors improve projected risk estimates, hedging, replication, and portfolio construction.

THE DTS CONCEPT

To understand the intuition behind DTS, consider the return, R_{spread} , due strictly to change in spread. Let D denote the spread duration of a bond and s its spread; the spread change return is then:

$$R_{\text{spread}} = -D \cdot \Delta s \quad (1)$$

Or, equivalently,

$$R_{\text{spread}} = -D \cdot s \cdot \frac{\Delta s}{s} \quad (2)$$

That is, just as spread duration is the sensitivity to an absolute change in spread (e.g., spreads widen by 5 bps), DTS ($D \cdot s$) is the sensitivity to a relative change in spread (e.g., spreads widen

by 5%). Note that this notion of relative spread change provides for a formal expression of the idea mentioned earlier—that credits with wider spreads are riskier since they tend to experience greater spread changes.

In the absolute spread change approach shown in equation (1), we can see that the volatility of excess returns can be approximated by

$$\sigma_{\text{return}} \cong D \cdot \sigma_{\text{spread}}^{\text{absolute}} \quad (3)$$

while in the relative spread change approach of equation (2), excess return volatility follows

$$\sigma_{\text{return}} \cong D \cdot s \cdot \sigma_{\text{spread}}^{\text{relative}} \quad (4)$$

Given that the two representations above are equivalent, why should one of them be preferable to another? The key advantage of modeling changes in spreads in relative terms is the resulting stability. The above equations, for simplicity, present returns and volatilities as idealized concepts. We have not added subscripts to specify whether we are referring to specific securities or sectors, or over what time period. Yet the way spread changes of different securities relate to each other, or the way volatilities in one time period relate to those in another, can be of critical importance in measuring and controlling portfolio risk.

For example, to determine a portfolio's exposure to a systematic widening of spreads, one needs to know how spread changes are likely to be realized across a sector. If one is concerned that spreads might move in parallel, then exposures should be measured as the overall contribution to spread duration as per equation (1). However, if spreads tend to change proportionally, then the contribution to DTS provides the correct exposure to such an event.

Similarly, volatility can be measured or projected in many different ways. Historically realized volatilities can be measured using observed spread changes at a specified frequency over a given sample period. Projections of forward-looking volatilities are the key building blocks of risk management systems. The accuracy with which historically re-

alized volatilities can project future volatilities is therefore of fundamental importance. If relative spread volatilities can be predicted with greater accuracy than absolute spread volatilities, then equation (4) should be preferred over (3). We found this to be the case, based on extensive empirical evidence from credit markets.

DTS AS BETA-ADJUSTED SPREAD DURATION

What are the dynamics of credit spread changes? Do spreads tend to widen in parallel, or do wider spreads widen by more? Figure 1 shows a specific example in which spread changes show a clear dependence on spread. The figure shows the changes in spreads experienced by key issuers in the Communications sector of the Barclays Capital Corporate Index in January 2001, during a temporary rally in the midst of the dot-com crisis. It is clear that this sector-wide rally was not characterized by a purely parallel shift; rather, issuers with wider spreads tightened by more.

Certainly, not all spread changes follow such a clear pattern. In many months, there are no large industry-wide spread changes, and spread changes are mostly idiosyncratic in nature. Occasionally, an industry will experience a systematic spread change that does seem to take the form of a parallel shift. However, an extensive set of regressions using individual bond spread changes across eight distinct market sectors and 185 months indicated that systematic factors expressed in terms of relative spread changes across an industry were able to capture nearly twice as much of the overall spread variance as factors based on parallel shifts in industry spreads. Furthermore, Ben Dor et al. (2007) found clear evidence that whenever a systematic widening or tightening of spreads across an industry occurred, credits with the highest spreads at the beginning of the month were most likely to experience the largest change in spreads.

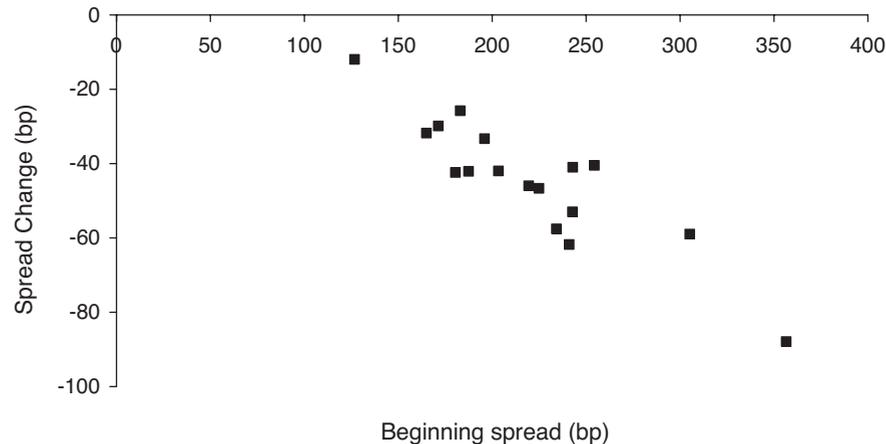


Figure 1 Average Spreads and Spread Changes for Key Issuers in the Communications Sector of the Barclays Capital Corporate Index (January 2001)
Source: Barclays Capital.

This idea may strike investors as reminiscent of the idea of “market beta” that is familiar from the capital asset pricing model (CAPM), in which the beta of a given security represents the extent to which it would be expected to participate in a market-wide rally or decline. Some credit market investors, in fact, have used models of *beta-adjusted spread duration* to measure *systematic risk exposures*. The difficulty with this approach lies in estimating the betas. Empirical betas can be backed out of historical data, for example, by regressing the spread changes realized by a given bond against the average spread changes of the sector. However, it is not clear how much historical data to use for this purpose—a short sample may not give a good statistical estimate, but a long sample may include observations from a time when the security had very different characteristics. From this viewpoint, we can offer another interpretation of DTS. Essentially, DTS can be viewed as an implementation of beta-adjusted spread duration, in which the betas are provided by the market in the form of spreads. The ratio of a given issuer’s spread to the average spread for the industry gives its beta, or sensitivity, to a relative spread change across the industry.

To demonstrate this, we carried out head-to-head tests of DTS versus empirical betas using weekly spread change data from the *credit default swap (CDS)* market.⁴ In the first test, we measured the empirical betas of each issuer’s CDS with respect to its industry peer group. We then tested two different predictors for this beta—either the empirical beta from the prior period, or the ratio of issuer DTS to the industry average DTS as of the beginning of the period. In the second test, we set up long-short CDS trades between two issuers from within the same industry and investigated different approaches to setting up the hedge ratios so as to minimize the systematic risk exposures of the trades. The DTS approach was found to be superior to empirical betas for both tasks.

THE RELATION BETWEEN SPREAD VOLATILITY AND SPREAD LEVEL

We now turn our attention to the dependence of spread change behavior on spread level. Figure 2 plots the relation between systematic spread volatility and spread level using over 15 years of monthly spread change data from

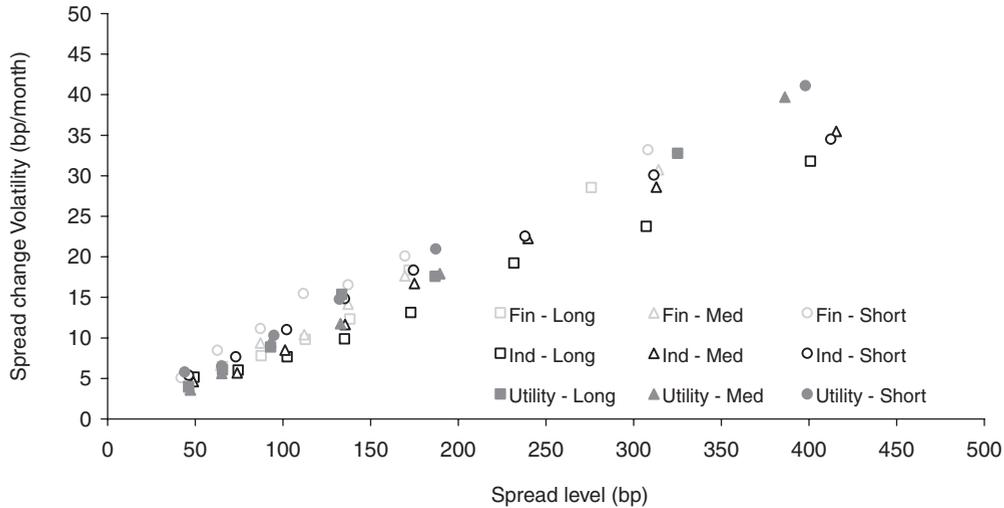


Figure 2 Systematic Spread Volatility versus Spread Level

Note: Based on monthly observations of bonds rated Aaa to B in the Barclays Capital Corporate and High-Yield Indexes, September 1989–January 2005.

Source: Barclays Capital.

U.S. credit markets, spanning investment-grade and high-yield rated bonds. The bonds in the Barclays Capital indexes for these markets were partitioned each month by sector, quality, and spread level. The average spread level for each market cell is plotted against the time-series volatility of the average absolute spread changes in each month. The results suggest that spread volatility can be closely approximated by a simple linear model of the form

$$\sigma_{\text{spread}}^{\text{absolute}}(s) \cong \theta \cdot s \tag{5}$$

This simple model provides an excellent fit to the data shown in Figure 2, with θ equal to 9.4% irrespective of sector or maturity. Hence, the results suggest that the historical volatility of systematic spread movements can be expressed quite compactly, in terms of a relative spread change volatility of about 9% per month. That is, spread volatility for a market segment trading at 50 bps should be about 4.5 bps/month, while that of a market segment at 200 bps should be about 18 bps/month. Ben Dor et al. (2007) documented a similar pattern for idiosyncratic volatility: The cross-sectional volatility of credit spread changes across a sec-

tor also exhibits a linear dependence on spread with about the same slope.

The results in Figure 2 suggest that measuring spread volatility in relative terms should be much more stable than absolute spread volatilities, and therefore forms the basis for more accurate projections of forward-looking risk. The advantage of using relative spread volatility should be particularly strong in the event of a market crisis. If we plot the absolute spread volatilities of various assets in the postcrisis period against their precrisis volatilities, we will find a marked increase across the board. Essentially, market data from the earlier period becomes useless for estimating risk in the postcrisis world. However, if we work with relative spread volatilities, we may find that they have not changed that much. The absolute spread volatility increases proportionally with the spread level, and the relative spread volatility remains stable. This relationship is illustrated in Figure 3 using data from U.S. credit markets in the period before and after the Russian crisis of 1998.

Two clear phenomena can be observed in Figure 3. First, as discussed above, most of the

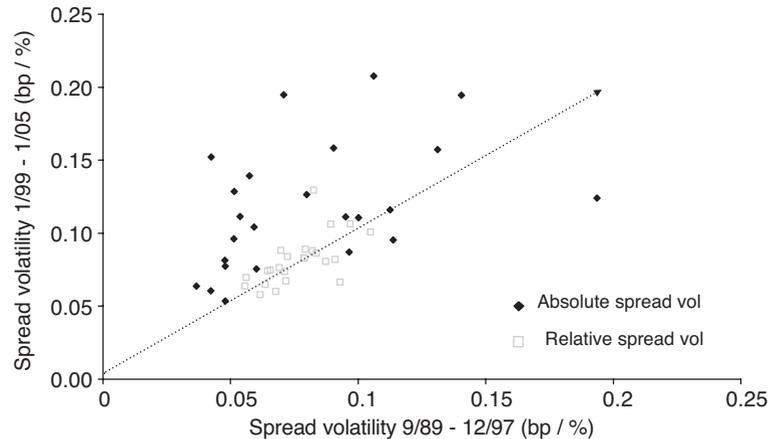


Figure 3 Absolute and Relative Spread Change Volatility before and after 1998

Note: Based on a partition of the Barclays Capital U.S. Corporate Index, 8 sectors \times 3 credit ratings. To enable the two to be shown on the same set of axes, both absolute and relative spread volatility are expressed in units with similar magnitudes. However, the interpretation is different: An absolute spread change of 0.1 represents a 10 bps parallel shift across a sector, while a relative spread change of 0.1 means that all spreads in the sector move by 10% of their current values (e.g., from 50 to 55, from 200 to 220). *Source:* Barclays Capital.

observations representing absolute spread volatilities are located far above the diagonal, pointing to an increase in volatility in the second period of the sample. In contrast, relative spread volatilities are quite stable, with almost all observations located on the 45-degree line or very close to it. This is because the pickup in volatility in the second period was accompanied by a similar increase in spreads. Second, the relative spread volatilities of various sectors are quite tightly clustered, ranging from 5%/month to a bit over 10%/month, whereas the range of absolute volatilities is much wider, ranging from 5 bps/month to more than 20 bps/month.

The results in Figure 3 exhibit the sharp discontinuity in credit market volatility that was experienced in 1998 due to the Russian crisis and the LTCM hedge fund failure. Since the introduction of DTS, global markets have provided us with ample opportunity to test the model with data from new out-of-sample crises. In both the credit crisis of 2007–2009 and the ensuing sovereign crisis that began in 2009, we have found that the DTS model has performed

admirably. In each case, a plot of precrisis vs. postcrisis volatility reveals results similar to Figure 3, showing the stability advantage of relative spread volatilities. The incorporation of spread into the projection of risk was shown to keep risk projections much more accurate than traditional absolute volatility risk measures.⁵

These results clearly indicate that absolute spread volatility is highly unstable and tends to rise with increasing spread. Computing volatilities based on relative spread change generates a more stable time series. These findings have important implications for the appropriate way of measuring credit exposures and projecting excess return volatility, which we discuss next.

DTS AND EXCESS RETURN VOLATILITY

If the volatility of both systematic and idiosyncratic spread changes is proportional to the level of spread, then equation (4) suggests two assertions regarding excess returns. First, excess return volatility should increase linearly with

DTS, where the slope represents the volatility of relative spread changes. Second, the magnitude of excess return volatility should be approximately equal across portfolios with similar DTS values, irrespective of their spread and spread duration characteristics. The results of Ben Dor et al. (2007) strongly supported both empirical predictions.

Another implication of the linear relation between spread level and spread volatility is that projecting volatility based on the current level of spread and the DTS slope from Figure 2 should be superior to using historical realizations of absolute spread changes. Specifically, using the product of DTS and the historical volatility of relative spread changes should generate better risk estimates than the product of spread duration and volatility of past absolute spread changes.

Our results confirmed that the DTS-based estimator was superior. A further indication that the DTS-based risk projection was more accurate is that it resulted in a smaller number of extreme realizations (above or below two standard deviations) than either of two estimators based on absolute spread volatility, using trailing windows of two different lengths.

Our understanding of these results is that the approach based on relative spread change volatility is able to give a more timely risk projection since it can react almost instantaneously to a change in market conditions reflected in the spread of the security. This should help the model react more quickly both to increase risk estimates at the onset of a crisis and to relax them once the turbulence subsides. Any significant widening or tightening of spreads will immediately flow through the DTS into the projection of excess return volatility.

IMPLICATIONS OF DTS FOR PORTFOLIO MANAGERS

We have highlighted above the key points that emerge from the empirical evidence supporting

the DTS paradigm. Spread changes are proportional to the level of spread. Systematic changes in spread across a sector tend to follow a pattern of relative spread change, in which bonds trading at wider spreads experience larger spread changes. The systematic spread volatility of a given sector (if viewed in terms of absolute spread changes) is proportional to the average spread in the sector; the nonsystematic spread volatility of a particular bond or issuer is proportional to its spread as well. Those findings hold irrespective of sector, duration, or time period.

There are several implications for a portfolio manager who wishes to act on these results. First, the best measure of exposure to a systematic change in spread within a given sector or industry is not the contribution to spread duration, but the contribution to DTS. At many asset management firms, the targeted active exposures for a portfolio relative to its benchmark are expressed as contribution-to-duration overweights and underweights within a sector by quality grid. Reports on the actual portfolio follow the same format. In the relative spread change paradigm, managers would express their targeted overweights and underweights in terms of contributions to DTS instead.

Second, our finding that the volatility of non-systematic return is proportional to DTS offers a simple mechanism for defining an issuer limit policy that enforces smaller positions in more risky credits. Many investors specify ad hoc weight caps by credit quality to control issuer risk. Alternatively, we can set a limit on the overall contribution to DTS for any single issuer. For example, say the product of Market value percentage \times Spread \times Duration must be 5 or less. Then, a position in issuer A, with a spread of 100 bps and a duration of five years, could be up to 1% of portfolio market value; while a position in issuer B, with a spread of 150 and an average duration of 10 years, would be limited to 0.33%. Issuer limits based on DTS and those based on market weight each have

their advantages and disadvantages; investors might want to consider some combination of the two.⁶

Third, DTS can help improve the hedging of security-vs.-security or security-vs.-market credit trades. Say a hedge fund manager has a view on the relative performance of two issuers within the same industry and would like to capitalize on this view by going long issuer A and short issuer B in a market-neutral manner. How do we define market neutrality? A typical approach might be to match the dollar durations of the two bonds, or to go long and short CDS of the same maturities with the same notional amounts. However, if issuer A trades at a wider spread than issuer B, our results would indicate that a better hedge against market-wide spread changes would be obtained by using more of issuer B, so as to match the contributions to DTS on the two sides of the trade.

Fourth, portfolio management tools such as risk and performance attribution models should represent sector exposures in terms of DTS contributions and sector spread changes in relative terms. A risk model for any asset class is essentially a set of factors that characterize the main risks that securities in that asset class are exposed to. The risk of an individual security or portfolio is computed based on its sensitivities to the various risk factors and the factor volatilities and correlations estimated from their past realizations. For credit-risky securities, traditional risk factors typically measure absolute spread changes based on a sector by quality partition that spans the universe of bonds. A risk factor specification based instead on relative spread changes has two important benefits. First, such factors would exhibit more stability over time and allow better forward-looking risk forecasts. Second, the partition by quality would no longer be necessary to control risk, and each sector can be represented by a single risk factor. This would allow managers to express more focused views, essentially trading off the elimination of the quality-based factors

with a more finely grained partition by industry. Similarly, a key goal for performance attribution models is to match the allocation process as closely as possible. If and when a manager starts to state allocation decisions in terms of DTS exposures, performance attribution should follow suit.

One practical difficulty that may arise in the implementation of DTS-based models is an increased vulnerability to pricing noise. Any small discrepancies in asset pricing should cause only small discrepancies in market values, but may potentially result in much larger variations in spreads. Consequently, managers who rely heavily on contribution-to-DTS exposures will need to implement strict quality controls on pricing.

Perhaps one of the most useful applications of DTS is in the management of core-plus portfolios that combine both investment-grade and high-yield assets. Traditionally, investment-grade credit portfolios are managed based on contributions to duration, while high-yield portfolios are managed based on market value weights. Using contributions to DTS across both markets could help bring consistency to this portfolio construction process. Skeptics may point out that in high-yield markets, especially when moving toward the distressed segment, neither durations nor spreads are particularly meaningful, and the market tends to trade on price, based on an estimated recovery value. A useful property of DTS in that context is that in the case of distressed issuers, where shorter duration securities tend to have artificially high spreads, DTS is fairly constant across the maturity spectrum, so that managing issuer contributions to DTS becomes roughly equivalent to managing issuer market weights.

The introduction of the DTS paradigm has had wide-ranging effects. It changed portfolio management practices across the industry and has been incorporated into some of the leading portfolio management analytics systems. We view it as a fundamental insight into the behavior of credit markets.

KEY POINTS

- Changes in credit spreads tend to be proportional to spread levels.
- Volatility of relative spread changes is more stable than volatility of absolute spread changes. This applies to all credit securities with a default component including corporate and sovereign issuers in developed and emerging market countries for both cash and derivatives.
- Whereas spread duration measures sensitivity to a parallel shift in spreads, DTS measures sensitivity to a relative change in spreads.
- The risk associated with credit spread exposures can therefore be managed more effectively using contributions to DTS than contributions to spread duration. This is true at the level of asset classes, industries, and individual issuers.
- Including spread in the estimation of risk can reduce the need to rely on credit ratings, allowing risk models to provide greater industry detail.

NOTES

1. See Ben Dor, Dynkin et al. (2007).
2. For example, see Ben Dor, Polbennikov, and Rosten (2007) and Ben Dor, Desclée, Hyman, and Polbennikov (2010).

3. See Chapter 4 in Ben Dor et al. (2012).
4. For details, see Chapter 8 in Ben Dor et al. (2012).
5. See Ben Dor et al. (2012) for details. The application of DTS to the modeling of European sovereign risk is discussed in Chapter 3 of Ben Dor et al. (2012), and Chapter 10 reviews the performance of the model through the 2007–2009 credit crisis.
6. For a detailed discussion of different approaches to issuer limits, see Chapter 11 in Ben Dor et al. (2012).

REFERENCES

- Ben Dor, A., Desclée, A., Hyman, J., and Polbennikov, S. (2010). Managing European sovereign spread risk. Barclays Capital, August.
- Ben Dor, A., Dynkin, L., Hyman, J., and Phelps, B. (2012). *Quantitative Credit Portfolio Management*. Fabozzi Series. Hoboken, NJ: John Wiley & Sons, 2012.
- Ben Dor, A., Dynkin, L., Hyman, J., Houweling, P., van Leeuwen, E., and Penninga, O. (2007). DTS (duration times spread). *Journal of Portfolio Management* 33(2): 77–100.
- Ben Dor, A., Polbennikov, S., and Rosten, J. (2007). DTS (duration times spread) for CDS: A new measure of spread sensitivity. *Journal of Fixed Income* 16(4): 32–44.
- Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29(2): 449–470.

Credit Spread Decomposition

SIDDHARTHA G. DASTIDAR, PhD, CFA
Vice President, Barclays

BRUCE D. PHELPS, PhD, CFA
Managing Director, Barclays

Abstract: Credit spread decomposition refers to breaking down a bond's option-adjusted spread to Treasuries into market-wide risk premium, expected default loss, and expected liquidity cost components. Credit spread decomposition is implemented empirically by regressing a bond's option-adjusted spread on a measure of its expected default losses (credit default swap spread) and expected liquidity cost. Credit spread decomposition can help investors determine the extent to which credit spreads reflect expected default losses, liquidity costs, or a market-wide risk premium. Investors can also apply spread decomposition analysis to construct targeted hedging strategies and to identify relative value opportunities. Regulators can use spread decomposition to monitor separately the liquidity and credit risk of the institutions they supervise, and to help determine capital adequacy.

At issuance, a credit bond has a positive yield spread (i.e., a credit spread) over comparable-maturity Treasury bonds to compensate investors for the chance that the bond may default with a recovery value less than par. However, studies have documented that credit spreads are generally much larger than justified by their subsequent default and recovery experience.¹

Beyond expected default losses, a portion of the credit spread may reflect the expected *liquidity cost* to execute a roundtrip trade. This cost is typically greater for a credit bond than for a comparable-maturity Treasury bond. Investors who anticipate selling a credit bond at some point demand compensation for this cost at the time of purchase in the form of a wider spread. Another portion of the credit spread may re-

flect a market-wide risk premium demanded by risk-averse investors due to the general uncertainty associated with the timing, magnitude, and recovery of defaults and the magnitude of liquidity costs. The greater the degree of this uncertainty, or the more risk-averse the marginal investor, the more the credit spread will exceed the expected *default cost*. *Credit spread decomposition* refers to the econometric exercise of breaking down a bond's option-adjusted spread (OAS) to Treasuries into its risk premium, expected default loss, and expected liquidity cost components.

Credit spread decomposition can serve many purposes. For example, suppose an insurance company, typically a buy-and-hold investor, is considering investing in credit bonds trading

at wide spreads. The company's decision will likely depend on whether the wide spreads are due to large expected default losses, high liquidity costs, or a high market risk premium. Presumably, the company can ride out periods of high liquidity cost and risk aversion. However, if the wide spreads reflect high expected default losses, the company may decide not to invest.

This entry begins with an example highlighting the ability of credit spread decomposition to reveal additional information hidden in a bond's OAS. Next, the entry outlines the specification of the spread decomposition model and shows how it can be implemented. Following a discussion on how to interpret the model results, the entry illustrates how they can be used in portfolio management applications. The entry concludes with a discussion of some alternative specifications of the spread decomposition model.

REVEALING THE DRIVERS OF CREDIT SPREADS

To illustrate the informational value of spread decomposition, consider the historical spread

behavior of a typical investment-grade bond. As shown in Figure 1 the bond's OAS varied over time. The figure also shows the level of the issuer's credit default swap (CDS) spread—a measure of expected default losses. While movements in the bond's OAS loosely tracked changes in the issuer's CDS, there was a wide and variable gap between the two spreads, reflecting movements in risk premium and expected liquidity costs.

Figure 1 also plots the bond's expected liquidity cost over the same period. To measure a bond's liquidity cost investors can use a bond's bid-ask spread (in price terms) expressed as a percentage of the bond's bid price. This cost is labeled as the bond's liquidity cost score (LCS) by Dastidar and Phelps (2009). Much of the variability in the OAS-CDS spread gap (the *CDS-cash basis*) mirrored movements in the bond's LCS. The initial rise in the issuer's OAS was driven by both default and liquidity concerns (all three lines moved up), whereas the larger subsequent spike was mainly a liquidity event (the line plotting the LCS moved up sharply while the CDS line was little changed). This example illustrates that investors need to measure the components of OAS separately to more fully understand OAS movements.

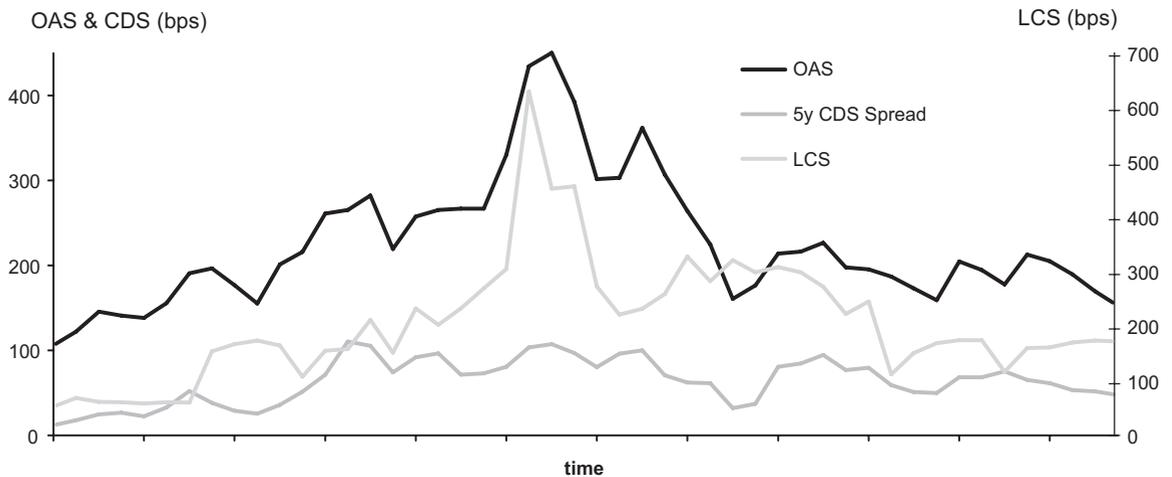


Figure 1 OAS, CDS, and LCS of a Typical Bond over Time

CREDIT SPREAD DECOMPOSITION: MODEL SPECIFICATION AND IMPLEMENTATION

To decompose credit spreads, a bond's OAS is regressed against three variables: a variable reflecting expected default cost, a variable reflecting expected liquidity cost, and a market-wide variable unrelated to the bond's attributes representing the *market-wide risk premium* demanded by investors. Conceptually, for every time t , the cross-sectional OLS regression model is:

$$OAS_{it} = \alpha_t + \beta_t \text{ExpectedDefaultCost}_{it} + \gamma_t \text{ExpectedLiquidityCost}_{it} + \eta_{it}$$

The risk premium variable (the intercept term, α) represents a market-level risk premium, not a risk premium specific to each bond. The value of the intercept is likely, but not necessarily, to be positive, reflecting that equilibrium credit spreads are typically determined at the margin by risk-averse investors.

Any bond-level risk premium is likely to be highly correlated with the bond's default cost or liquidity cost. In other words, an investor will demand a higher spread premium for a bond with a high liquidity cost as compensation for liquidity cost uncertainty. This makes it difficult to decompose a bond's spread into separate expected liquidity cost and liquidity risk premium components. The same applies to default cost and default risk premium. If default risk or liquidity risk premiums are highly correlated with default or liquidity costs, then the regression coefficients (β and γ) will be larger and/or more significant. Any part of the risk premiums that is unrelated to bond-level default and liquidity cost—in other words, a market-level risk premium—will show up in the intercept.

Credit spread decomposition is implemented empirically by running the following regression across a set of bonds (denoted by i) at a given

time t :

$$OAS_{it} = \alpha_t + \beta_t CDS_{it} + \gamma_t LCS_{it} + \eta_{it} \quad (1)$$

The LCS is used to measure bond-level expected liquidity cost. An issuer's CDS (with a similar spread duration as the bond) is used to measure its expected default cost (i.e., default probability and loss given default). If the CDS itself is illiquid, it will contain some illiquidity premium, thereby distorting results. So, only liquid CDS should be chosen. While an issuer's CDS can be used to measure the expected default cost of its bonds, other measures of expected default cost could be used in lieu of CDS. For example, some investors may use firm-specific fundamental information, equity prices, and macroeconomic data to estimate an issuer's default probability and recovery rate.

To get a sense of the value of incorporating a bond-level liquidity variable to explain the cross-sectional distribution of spreads, an investor can first estimate the model without LCS as an explanatory variable. The model can then be re-estimated adding LCS to see if the regression's fit improves and does not detract from the explanatory power of CDS. If LCS is a useful explanatory variable, adding LCS as a regressor should produce an improvement in the adjusted R^2 and a significant (and positive) LCS coefficient, with little disturbance to the significance and magnitude of the CDS coefficient.

INTERPRETING THE RESULTS OF THE CREDIT SPREAD DECOMPOSITION MODEL

The estimated regression coefficients can be used to break down the average OAS into the three spread components in terms of basis points. For example, suppose the average OAS, CDS, and LCS are 2.09%, 1.14%, and 0.73%, respectively. In addition, suppose the estimated coefficients of CDS and LCS are 0.67 and 1.41,

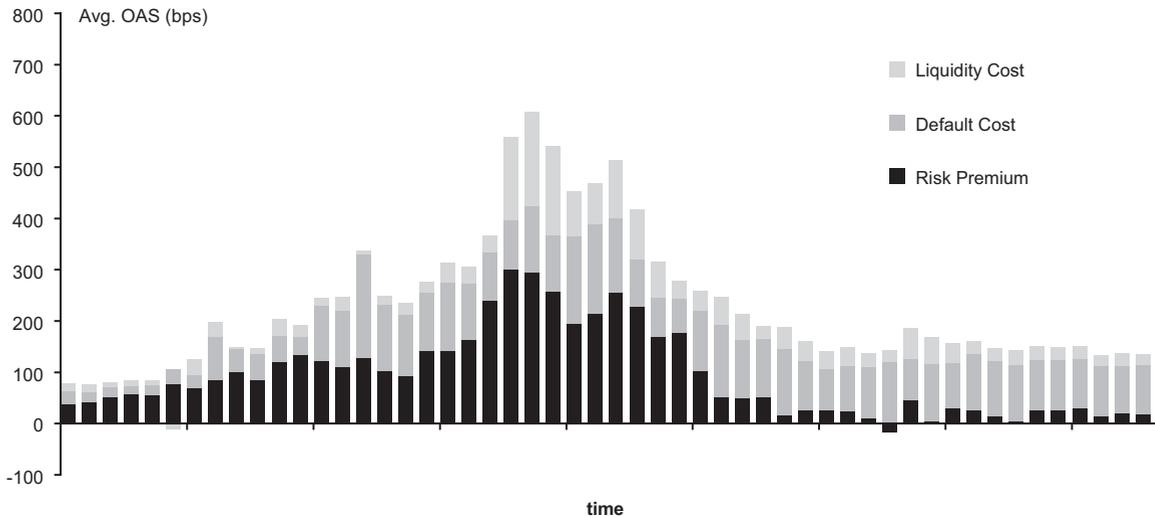


Figure 2 Relative Contribution of Default Cost, Liquidity Cost, and Risk Premium over Time

respectively. A variable's contribution to the average OAS can be determined by multiplying the average value of the variable by its estimated regression coefficient (e.g., $1.41 \times 0.73\%$ is the contribution of LCS to the average OAS). Repeating spread decomposition at different time periods can show fluctuations in the relative contributions to OAS of the three components over time as shown in Figure 2.

When liquidity is abundant, LCS might not play an important role in explaining spread differences across bonds. In fact, adding LCS to the regression may not meaningfully improve the R^2 . In contrast, when liquidity conditions deteriorate, adding LCS to the regression will likely improve the R^2 .

As discussed, the regression intercept captures the portion of (average) spread that is independent of CDS and LCS. The market risk premium is likely to be, at times, an important contributor to the level of OAS. The time series can be used as an indicator of the variation of the market risk premium—or risk aversion—in the credit market. When the intercept explains a relatively high proportion of OAS, this suggests that systematic market factors, rather than bond-specific factors, are driving spreads. This may occur because of very high levels of aggre-

gate risk aversion or because the market is pricing bonds with little concern for issuer-specific information. When the intercept explains a relatively low proportion of OAS, this suggests that bond-specific factors are driving spreads.

The regression coefficients for both CDS and LCS are expected to be positive. While the relationship of CDS with OAS is naturally tight, it may not be as close as one might think. Since default risk for high-grade bonds has been very low over long periods of time, a relatively large proportion of the OAS is likely liquidity-related.

APPLICATIONS OF CREDIT SPREAD DECOMPOSITION

The parameter estimates from the spread decomposition model can be used in a variety of portfolio management applications. Active portfolio managers can use spread decomposition to take positions in specific bonds with large liquidity or default components, depending on their views about how these components are likely to evolve. Regulators can use spread decomposition to monitor separately the liquidity and credit risk embedded in the credit

portfolios of the institutions they supervise, which can help determine capital adequacy.

Presented below is a discussion of two important applications of credit spread decomposition: identifying bonds that may be trading “rich” or “cheap,” and allowing the manager to construct hedges that target specific drivers of OAS fluctuations.

Identifying Relative Value

So far, credit spread decomposition analysis has been described using contemporaneous data to attribute OAS levels to default and liquidity cost components at a given time. However, investors can apply spread decomposition analysis to *ex ante* investment decisions as well.

In principle, spread decomposition should help identify *relative value* opportunities. A bond’s OAS can be compared with its estimated OAS using the parameters from the spread decomposition model. If the actual OAS is wider than the estimated OAS, it suggests that the bond is trading too wide, and vice versa. This may be a signal that the bond’s OAS may change to correct this “mispricing.”

To examine whether the realized residuals $\hat{\eta}_{i,t}$, from (1) can help predict future OAS changes, one can examine whether the bond’s future OAS changes are of the opposite sign to the sign of the residual by running the following regression and testing to see if the θ ’s are negative.

$$\Delta OAS_{it,t+j} = \alpha_t + \theta_t \hat{\eta}_{it} + \delta_t \text{MonthDummy}_t + e_{it} \quad (2)$$

Hedging a Credit Bond Portfolio

One method to determine a hedge for a credit is to use regression to examine the historical relationship between the bond’s OAS and potential hedge variables. The issuer’s CDS may be an effective hedge targeted against changes in expected default losses. Since movements in the volatility index (VIX) are closely related to changes in LCS,² VIX futures can potentially

be used as a *credit hedging* instrument to target spread changes related to changes in liquidity.

If an investor seeks to hedge the default or liquidity components separately, then the contribution to OAS in basis points from the credit spread decomposition model (in differences—discussed below) determines the appropriate hedge ratio for each component. Of course, the success of such a hedge depends on the goodness of fit and whether the historical relationship will hold in the future.

ALTERNATIVE CREDIT SPREAD DECOMPOSITION MODELS

There are alternative formulations of the credit spread decomposition model. As discussed earlier, the analysis has ignored explicit bond-level risk premium variables. Instead, it assumes that any bond-level risk premium is highly related to either the expected liquidity cost or the expected default cost. An alternative model can include a term representing a bond-level liquidity risk premium. This additional term reflects compensation demanded by investors for the risk that the actual cost at liquidation may be different from the expected liquidity cost as measured by the current LCS. A bond’s LCS volatility over the prior 12 months can be considered a measure of liquidity risk. For example, two bonds may have the same LCS today, but bond A may have a much more volatile LCS history than bond B. An investor may view bond A as having a riskier liquidity cost and demand an OAS premium versus bond B, all else equal.

The equation below shows the spread decomposition model incorporating a bond-level liquidity risk factor, $LCSVol_{i,t}$. Generally, the results may show that $LCSVol_{i,t}$ is highly significant, but absorbs part of the effect of LCS, thereby not improving the regression’s adjusted R^2 substantially.

$$OAS_{it} = \alpha_t + \beta_t CDS_{it} + \gamma_t LCS_{it} + \phi_t LCSVol_{it} + \delta_t \text{MonthDummy}_t + \eta_{it} \quad (3)$$

The credit spread decomposition model can also be estimated in differences to check if changes in the liquidity and default proxies affect changes in OAS (i.e., contemporaneous returns). The regression model below details the specification, where ΔOAS_{it} , ΔCDS_{it} , and ΔLCS_{it} refer to changes in a bond's characteristics in consecutive periods. As described above, this model of spread decomposition can be used for designing targeted hedges.

$$\Delta OAS_{it} = \alpha_t + \beta_t \Delta CDS_{it} + \gamma_t \Delta LCS_{it} + \delta_t \text{MonthDummy}_t + \eta_{it} \quad (4)$$

Finally, the spread decomposition model may be susceptible to outliers, especially since default and liquidity are arguably more important considerations for higher spread bonds. To check this, one can run log regressions (e.g., the dependent variable is $\ln(OAS)$ instead of OAS , similarly for the independent variables), as shown below. If the conclusions from the log model are unchanged, this would indicate that outliers are not driving the results.

$$\ln(OAS_{it}) = \alpha_t + \beta_t \ln(CDS_{it}) + \gamma_t \ln(LCS_{it}) + \eta_{it} \quad (5)$$

KEY POINTS

- Credit spread decomposition refers to breaking down a bond's option-adjusted spread (OAS) to Treasuries into market risk premium, expected default loss, and expected liquidity cost components.
- To decompose credit spreads, a bond's OAS is regressed on a measure of its expected default cost (CDS) and expected liquidity cost (LCS).
- Credit spread decomposition can help credit investors determine the extent to which spreads reflect expected default losses, high liquidity costs, or a high market-wide risk premium, and make portfolio decisions accordingly.
- Investors can also apply spread decomposition analysis for determining targeted hedging strategies and to help identify relative value opportunities.

NOTES

1. See, for example, Ng and Phelps (2011) and Elton et al. (2001).
2. Dastidar and Phelps (2009).

REFERENCES

- Dastidar, S., and Phelps, B. (2009). Introducing LCS: Liquidity cost scores for US credit bonds. Barclays Capital, New York.
- Dastidar, S., and Phelps, B. (2011). Credit spread decomposition: Decomposing bond-level credit OAS into default and liquidity components. *Journal of Portfolio Management* 37, 3: 70–84.
- Elton, E. J., Gruber, M. J., Agrawal, D., and Mann, C. (2001). Explaining the rate spread on corporate bonds. *Journal of Finance* 56: 247–277.
- Ng, K. Y., and Phelps, B. (2011). Capturing credit spread premium. *Financial Analysts Journal* 67, 3: 63–75.

Credit Derivatives and Hedging Credit Risk

DONALD R. VAN DEVENTER, PhD

Chairman and Chief Executive Officer, Kamakura Corporation

Abstract: The credit crisis of 2007–2009 in the United States and Europe and the collapse of the *Japanese bubble* in the 1990–2002 period show that, without hedging credit risk, the largest financial institutions in the world are very likely to fail. Many trillions of dollars of taxpayer bailouts have put the credit quality of the United States and Japan at risk. The solution to this financial institutions' risk management problem and the related sovereign risk problem is hedging with respect to macro factor movements. Hedging interest rate movements has a 40-year history, but now the focus has turned to a longer list of macro factors like home prices, commercial real estate prices, oil prices, commodity prices, foreign exchange rates, and stock indices. This hedging capability is now widely available in best practice enterprise risk management software. Stress testing with respect to macro factors is now a mandatory requirement of the European Central Bank and U.S. bank regulators.

In this entry, we examine practical tools for hedging *credit risk* at both the transaction level and the portfolio level, focusing on the interaction between the credit modeling technologies and traded instruments that would allow one to mitigate credit risk. We start with a discussion linking credit modeling and credit portfolio management in a practical way. We then turn to the credit default swap market as a potential hedging tool. Finally, the state of the art is discussed: hedging transaction level and portfolio credit risk using hedges that involve macroeconomic factors that are traded in the marketplace.

CREDIT PORTFOLIO MODELING: WHAT'S THE HEDGE?

One of the reasons that the popular value-at-risk (VaR) concept has been regarded as an incomplete risk management tool is that it provides little or no guidance on how to hedge if the VaR indicator of risk levels is regarded as too high. In a more subtle way, the same criticisms apply to many of the key modeling technologies that are popular in financial markets, like the *copula approach* to the simulation of credit portfolios. In this entry we summarize the



Figure 1 Cyclical Rise and Fall in 5-Year Reduced Form Default Probabilities: Citigroup and Ford Motor Company, 2006–2011

virtues and the vices from a hedging perspective of both various credit modeling techniques and credit derivative instruments traded in the marketplace. One of the key issues that requires a lot of attention in credit portfolio modeling is the impact of the business cycle on default probabilities. Default probabilities rise and fall when the economy weakens and strengthens. This is both obvious and so subtle that almost all commercially available modeling technologies ignore it. It's easy to talk about it and hard to do.

Figures 1 and 2 show the cyclical rise and fall in 5-year reduced-form default probabilities for Citigroup and Ford Motor Company for the periods 1990–2005 and 2006–2011.¹ The figures show the obvious correlation in default probabilities for both companies as they rise or fall in the 1990–1991 recession and in the recession spanning 1999–2003, depending on the sector, but the greatest correlation is in the *credit crisis* period of 2007–2009. Over the full 1990–2011 period, their respective 5-year default probabilities have a simple correlation of 45.2%.

With this common knowledge as background, we begin with the hedging implications of the *Merton model* at the individual transaction and portfolio level (see Merton, 1974).

THE MERTON MODEL AND ITS VARIANTS: TRANSACTION-LEVEL HEDGING

As of this writing, every publicized commercial implementation of the Merton model or its variants has one principal assumption in common: The only random factor in the model is the “value of company assets.” Regardless of the variety of Merton model used, all models of this type have the following attributes in common when the value of company assets rises:

- Stock prices rise.
- Debt prices rise.
- Credit spread falls.
- Default probability falls.

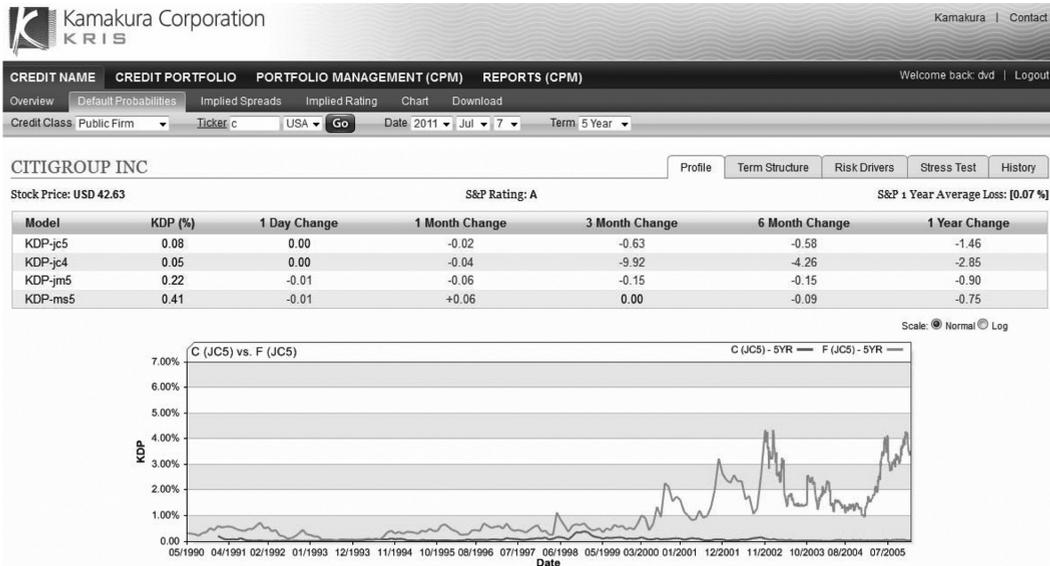


Figure 2 Cyclical Rise and Fall in 5-Year Reduced Form Default Probabilities: Citigroup and Ford Motor Company, 1990–2005

From a theoretical point of view, there are three obvious ways to think about hedging in the Merton context:

- Hedge a long position in the debt of the firm with a short position in the assets of the company.
- Hedge a long position in the debt of the firm with a short position in the common stock of the company.
- Hedge a long position in the debt of the firm with a short position in another debt instrument of the company.

The first hedging strategy is consistent with the assumptions of the Merton model and all of its commercial variants, because assets of the firm are assumed to be traded in perfectly liquid efficient markets with no transactions costs. Unfortunately, for most industrial companies, this is a very unrealistic assumption. Investors in Ford Motor Company cannot go long or short auto plants in any proportion. The third hedging strategy is also not a strategy that one can use in practice, although the credit derivative instruments we discuss in the next section provide a variation on this theme.

From a practical point of view, shorting the common stock is the most direct hedging route and the one that combines a practical hedge and one consistent with the model theory. Unfortunately, however, even this hedging strategy has severe constraints that restrict its practical use. Specifically, even if the Merton model or its variant is true, mathematically, the first derivative of the common stock price with respect to the value of company assets approaches zero as the company becomes more and more distressed. When the value of company assets is well below the amount of debt due, the common stock will be trading just barely above zero. One would have to short more and more equity to offset further falls in debt prices, and at some point a hedging strategy that shorts even 100% of the company's equity becomes too small to fully offset the risk still embedded in debt prices. In short, even if the Merton model is literally true, the model fails the hedging test ("What's the hedge?") for deeply distressed situations.

What about companies that are not yet severely distressed? Jarrow and van Deventer (1998, 1999), analyzed a 9-year weekly data

series of new issue fixed rate bond spreads collected by First Interstate Bancorp, which at the time was the seventh largest bank holding company in the United States. Over the sample period used by Jarrow and van Deventer, First Interstate's debt ratings varied from AA to BBB. They analyzed the debt and equity hedge ratios produced by the Merton model (and its variants) and tested for biases that would reduce hedging errors. The results of that analysis showed that a common stock hedge in the opposite direction of that indicated by the Merton model (and its variants) would have improved results. That is, one should have gone long the equity even if one is long the debt, not short the equity. Jarrow and van Deventer are careful to point out that this strategy is certainly not recommended. The reason for this finding was simple: During the 9-year weekly data series beginning in 1984, credit spreads and stock price changes move in the direction predicted by the Merton model less than 45% of the time. Van Deventer and Imai (2003) obtain similar results over a much larger sample.

Jarrow and van Deventer make the point that the Merton model is clearly missing key variables that would allow credit spreads and equity prices to move in either the same direction or the opposite direction as these input variables change. None of the Merton models in commercial use have this flexibility and therefore any hedge ratios they imply are quite suspect.

What about companies that are not investment grade but do not yet fall in the "severely distressed" category? It is in this sector that individual transaction hedging using Merton-type intuition is potentially the most useful. Most of the research that has been done in this regard has been done on a proprietary basis on Wall Street. Even if the Merton model hedging is useful for companies in the BB and B ratings grade, how effective can it be in protecting the owner of a bond that once was rated AA but sinks to a distressed CCC? Whether or not hedging errors in the AA to BBB and CCC ratings

ranges more than offset hedging benefits in the BB and B range is an important question. Modern corporate governance requires that users of the Merton model have evidence that it works in this situation, rather than relying on a belief that it works. On September 12, 2005, the *Wall Street Journal* reported on the hundreds of millions of dollars that were lost by arbitrageurs using Merton-type hedges on Ford and General Motors when both firms were downgraded by the major rating agencies.²

There are a few more points that one needs to make about the Merton model and all of its commercial variants when it comes to transaction level hedging:

- The Merton model default probability is not an input in this hedging calculation for the same reason that the return on the common stock is not an input in the Black-Scholes options model. The Merton model and all of its commercial variants incorporate all possible probabilities of default that stem from every possible variation in the value of company assets.
- Loss given default is also not an input in this hedging calculation because all possible loss given defaults (one for each possible ending level of company asset value) are analyzed by the Merton model and in turn have an impact on the calculated hedge ratio.

These insights are not widely recognized by analysts who consider hedging using the Merton technology. Given the value of company assets, we can calculate the Merton hedge ratio directly with no need for a default probability estimate or a loss given default estimate. If instead we are given the Merton (or its variants) default probability, we do not know the hedge ratio without full disclosure of how the default probability was derived. Any failure to make this disclosure is a probable violation of the Basel II capital accords from the Basel Committee on Banking Supervision.

THE MERTON MODEL AND ITS VARIANTS: PORTFOLIO-LEVEL HEDGING

One of the attractive things about the Merton model, in spite of the limitations mentioned above, is its simple intuition. We know that the basic businesses of Ford and General Motors are highly correlated, so it is a small logical step to think about how the assets of the two companies must be closely correlated. One has to make a very substantial set of additional assumptions if one wants to link the macroeconomic factors that drive correlated defaults to the value of company assets in the Merton framework or any of its one-factor commercial variants. Let's assume away those complexities and assume that we know the returns on the assets of Ford have a 0.25 correlation with the returns on the assets of General Motors. Note that the 0.25 correlation does not refer to

- The correlation in the default probabilities themselves.
- The correlation in the events of default, defined as the vector of 0s and 1s at each time step where 0 denotes no default and 1 denotes default.

These are different and mathematically distinct definitions of correlation. Jarrow and van Deventer (1998, 1999) show some of the mathematical links between these different definitions of correlation. Jarrow and van Deventer (2005) formalize these results.

Once we have the correlation in the returns on the value of company assets, we can simulate correlated default as follows:

- We generate N random paths for the values of company assets of GM and Ford that show the assumed degree of correlation.
- We next calculate the default probability that would prevail, given that level of company assets, at that point in time in the given scenario.
- We then simulate default/no default.

For any commercial variant of the Merton model, an increase in this "asset correlation" results in a greater degree of bunching of defaults from a time perspective. This approach was a common first step for analysts evaluating first-to-default swaps and collateralized debt obligations because they can be done in common spreadsheet software packages with a minimum of difficulty.

There are some common pitfalls to beware of in using this kind of analysis that are directly related to the issues raised about the Merton framework and its commercial variants:

- If one is using the original Merton model of risky debt, default can happen at only one point in time: the maturity date of the debt. This assumption has to be relaxed to allow more realistic modeling.
- If one is using the "down and out option" variation of the Merton model, which dates from 1976, one has to specify the level of the barrier that triggers default at each point in time during the modeling period.

Unless one specifically links the value of company assets to macroeconomic factors, the portfolio simulation has the same limitations from a hedging point of view as a single transaction. As explained earlier, the hedge using a short position in the common stock would not work for deeply troubled companies from a theoretical point of view and it does not work for higher rated credits (BBB and above) from an empirical point of view.

If one does link the value of company assets to macroeconomic factors, there is still another critical and difficult task one has to undertake to answer the key question: "What's the hedge?" One needs to convert the single-period, constant interest rates Merton model or Merton variant to a full valuation framework for multiperiod fixed-income instruments, many of which contain a multitude of embedded options (like a callable bond or a line of credit). One of the many lessons of the *Wall Street Journal* article cited above and subsequent

experience in 2007–2009 is easy to summarize: This approach to hedging and simulating credit risk (called the “copula approach” as well as the Merton approach) simply did not work. Salmon (2009) called the Merton/copula approach the “formula that killed Wall Street” via the \$945 billion in credit losses that resulted from the credit crisis.³

What if we want to use the Merton/copula approach in spite of its role in recent losses? As Lando (2004) discusses, this is a large set of non-trivial analytical issues to deal with. Most importantly, moving to a multiperiod framework with random interest rates leads one immediately to the *reduced form model* approach, where it is much easier for the default probability models to be completely consistent within the valuation framework. We turn to that task now.

Reduced-Form Models: Transaction-Level Hedging

One of the many virtues of the reduced form modeling approach is that it explicitly links factors driving default probabilities, like interest rates and other macroeconomic factors, to the default probabilities themselves. Just as important, the reduced form framework is a multiperiod, no-arbitrage valuation framework in a random interest rate context. Once we know the default probabilities and the factors driving them, credit spreads follow immediately, as does valuation. Valuation, even when there are embedded options, often comes in the form of analytical closed-form solutions. More complex options require numerical methods that are commonly used on Wall Street. The ability to stress test portfolio values and portfolio losses with respect to *macro factor* movements is now required by the European Central Bank and by U.S. bank regulators via two programs: the Comprehensive Capital Assessment Review and the Supervisory Capital Assessment Program. The later program, required of the top 19 U.S. financial institutions in 2009, mandated stress tests with respect to

changes in *home prices*, real gross domestic product, and the unemployment rate.

Suffice it to say that for any simulated value of the risk factors driving default, there are two valuations that can be produced in the reduced form framework. The first valuation is the value of the security in the event that the issuer has not defaulted. This value can be stress tested with respect to the risk factors driving default to get hedge ratios with respect to the nondiversifiable risk factors. The second value that is produced is the value of the security given that default has occurred. In the reduced form framework of Duffie and Singleton (1999) and Jarrow (2001), this loss given default can be random and is expressed as a fraction of the defaultable instrument one instant prior to default.

These default-related jumps in value have two components. The first part is the systematic (if any) dependence of the loss given default or recovery rate on macroeconomic factors. The second part is the issuer-specific default event, since (conditional on the current values of the risk factors driving default for all companies) the events of default are independent. At the individual transaction level, this idiosyncratic company-specific component can only be hedged by shorting a defaultable instrument of the same issuer or a credit default swap of that issuer.

At the portfolio level, this is not necessary. We explain why next.

Reduced-Form Models: Portfolio-Level Hedging

One of the key conclusions of a properly specified reduced form model is that the default probabilities of each of N companies at a given point in time are independent, conditional on the values of the macroeconomic factors driving correlated defaults. That is, as long as none of the factors causing correlated default have been left out of the model, then by definition, given the value of these factors, default is

independent. This is an insight of Jarrow, Lando, and Yu (2005).

This powerful result means that individual corporate credit risk can be diversified away, leaving only the systematic risk driven by the identified macroeconomic variables. This means that we can hedge the portfolio with respect to changes in these macroeconomic variables just as we do in every hedging exercise: We mark to market the portfolio on a credit-adjusted basis and then stress test with respect to one macroeconomic risk factor. We calculate the change in value that results from the macroeconomic risk factor shift and this gives us the “delta.” We then can calculate the equivalent hedging position to offset this risk. This is a capability that is present in modern enterprise-wide risk management software.⁴

This exercise needs to be done for a wide range of potential risk factor shifts, recognizing that some of the macroeconomic risk factors are in fact correlated themselves. Van Deventer, Imai, and Mesler (2004) outline procedures for doing this in great detail.

We turn now to commonly used credit-related derivative instruments and discuss what role they can play in a hedging program.

CREDIT DEFAULT SWAPS AND HEDGING

Credit default swaps in their purest form provide specific credit protection on a single issuer. They are particularly attractive when the small size of a portfolio (in terms of issuer names) or extreme concentrations in a portfolio rule out diversification as a vehicle for controlling the idiosyncratic risk associated with one portfolio name.

Generally speaking, credit default swaps should only be used when diversification does not work. As we discuss in a later section, dealing directly in the macroeconomic factors that are driving correlated default is much more efficient both in terms of execution costs and in

terms of minimizing counterparty credit risk. An event that causes a large number of corporate defaults over a short time period would also obviously increase the default risk of the financial institutions that both lend to them and act as intermediaries in the credit default swap market. This insight was not widely appreciated as recently as 2006, but it is now. The bankruptcy of Lehman Brothers on September 15, 2008, the March 2008 rescue of Bear Stearns, and the September 2008 rescues of Merrill Lynch (by Bank of America with U.S. government support) and Morgan Stanley (by the Federal Reserve) have convinced any doubters of the importance of counterparty credit risk in the credit default swap market. As of this writing, only 14 dealers are registered to clear credit default swaps with the Depository Trust and Clearing Corporation.⁵

Many researchers have begun to find that credit spreads and credit default swap quotations are consistently higher than actual credit losses would lead one to expect.⁶ How can such a “liquidity premium” persist in an efficient market? From the perspective of the insurance provider on the credit default swap, in the words of one market participant, “Why would we even think about providing credit insurance unless the return on that insurance was a lot greater than the average losses we expect to come about?” That preference is simple enough to understand, but why doesn’t the buyer of the credit insurance refuse to buy insurance that is “overpriced”?

One potential explanation is related to the lack of diversification that individual market participants face even if their employers are fully diversified. An individual fund manager may have only 10–20 fixed income exposures and a bonus pool that strictly depends on his or her ability to outperform a specific benchmark index over a specific period of time. One default may devastate the bonus, even if the fund manager in 1 billion repeated trials may in fact outperform the benchmark. The individual has more reason to buy single-name credit

insurance than the employer does because (1) his or her work-related portfolio is much less diversified than the entire portfolio of the employer, (2) the potential loss of the bonus makes him or her much more risk averse than the employer, and (3) the employer is much less likely to be aware that the credit insurance is (on average) overpriced than the individual market participant. Jarrow, Li, Mesler, and van Deventer (2007) have quantified the magnitude of this premium and shown that factors as diverse as company size (bigger firms get smaller spreads) and location (Japanese firms get smaller spreads) affect the premium of CDS spreads over default risk. These premiums are available daily via the Thomson Reuters "Credit Views" page.⁷

A more important concern with credit default swap hedging is the very thin trading volume in the CDS market in the aftermath of the credit crisis. A study⁸ found the following:

- Only 241 corporate reference names averaged more than 5 trades per day.
- Only 63 reference names averaged more than 10 trades per day.
- Only 14 reference names averaged more than 15 trades per day.
- No reference names averaged more than 23 trades per day.

Given these low volumes, there is a serious risk of market manipulation that should give any potential hedger great concern.

PORTFOLIO- AND TRANSACTION-LEVEL HEDGING USING TRADED MACROECONOMIC INDICES

The instantaneous probability of default can be specified as a linear function of one or more macroeconomic factors. An example is the case where the default intensity is a linear function of the random short-term rate of interest

r and a macroeconomic factor with normally distributed return Z :

$$\lambda(t) = \lambda_0 + \lambda_1 r(t) + \lambda_2 Z(t)$$

The constant term in this expression is an idiosyncratic term that is unique to the company. Random movements in the short rate r and the macroeconomic factor Z will cause correlated movements in the default intensities for all companies whose risk is driven by common factors. The default intensity has a term structure like the term structure of interest rates, and this entire term structure moves up and down with the business cycle as captured by the macroeconomic factors. The parameters of this reduced form model can be derived by observable histories of bond prices of each counterparty or from observable histories of credit derivatives prices using enterprise-wide risk management software.

Alternatively, a historical default database can be used to parameterize the term structure of default probabilities using discrete instead of continuous default probabilities, just as discrete interest rates are used in practice based on yield curve movements in continuous time. The most common approach to historical default probability estimation uses logistic regression. For each company, monthly observations are denoted 0 if the company is not bankrupt in the following month and 1 if the company does go bankrupt in the next month. Explanatory variables X_i are selected and the parameters α and β , which produce the best fitting predictions of the default probability using the following logistic regression formula:

$$P[t] = 1/[1 + \exp(-\alpha - \sum_{i=1}^n \beta_i X_i)]$$

By fitting this logistic regression for each default probability on the default probability term structure, one can build the entire cumulative and annualized default probability term structures for a large universe of corporations. Figure 3 shows the cumulative term structure of

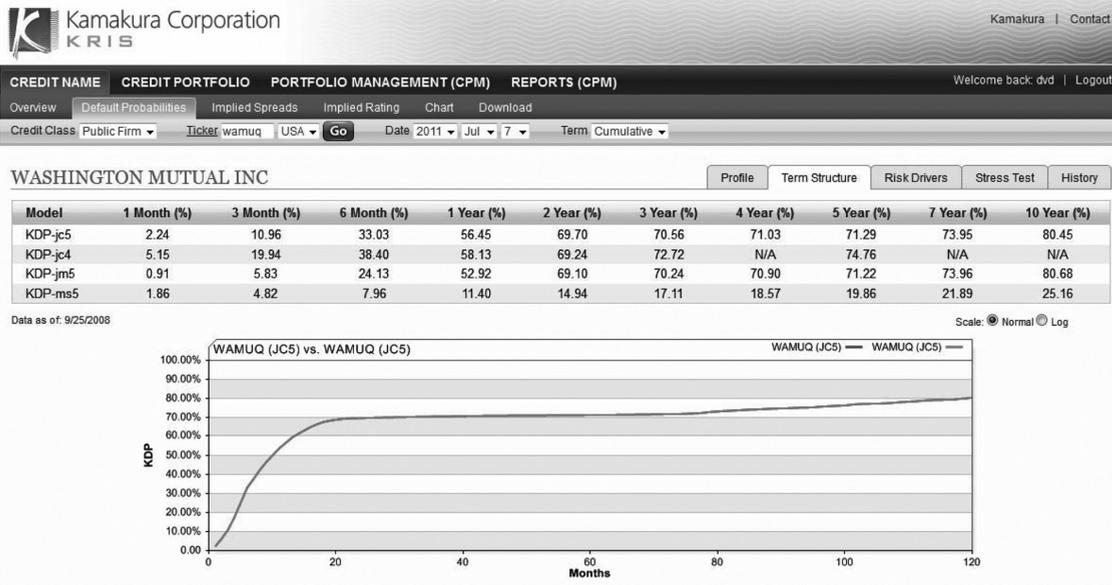


Figure 3 Cumulative Term Structure of Default Probabilities for Washington Mutual: September 2008, One Day Prior to Default

default probabilities for Washington Mutual, just prior to its failure in September 2008.

Alternatively, one can annualize the entire term structure of default probabilities for easy comparison with credit spreads and credit default swap quotations. The resulting curve is downward sloping for high-risk credits like Washington Mutual (see Figure 4).

The key advantage of the reduced-form approach is that critical macroeconomic factors can be linked explicitly to default probabilities as explanatory variables. The result is a specific mathematical link like the linear function of the pure Jarrow reduced form model or the logistic regression formula used for historical database fitting. The logistic regression formula is very powerful for simulating forward since it always produces default probability values between zero and 100%. These values can then be converted to the linear Jarrow form for closed-form mark-to-market values for every transaction in a portfolio.

Van Deventer, Imai, and Mesler (2004) then summarize how to calculate the macroeco-

nomie risk factor exposure as follows. The Jarrow model is much better suited to hedging credit risk on a portfolio level than the Merton model because the link between the (N) macro factor(s) M and the default intensity is explicitly incorporated in the model. Take the example of Washington Mutual, whose probability of default is driven by interest rates and home prices, among other things. If $M(t)$ is the macro factor defined as the one-year change in home prices, it can be shown that the size of the hedge that needs to be bought or sold to hedge one dollar of risky debt zero coupon debt with market value v under the Jarrow model is given by

$$\begin{aligned} \partial v_l(t, T : i) / \partial M(t) = & -[\partial \gamma_l(t, T) / \partial M(t) \\ & + \lambda_2(1 - \delta_i)(T - t) / \\ & \sigma_m M(t)] v_l(t, T : i) \end{aligned}$$

The variable v is the value of risky zero-coupon debt and γ is the liquidity discount function representing the illiquidities often observed in the debt market. There are similar formulas in the Jarrow model for hedging

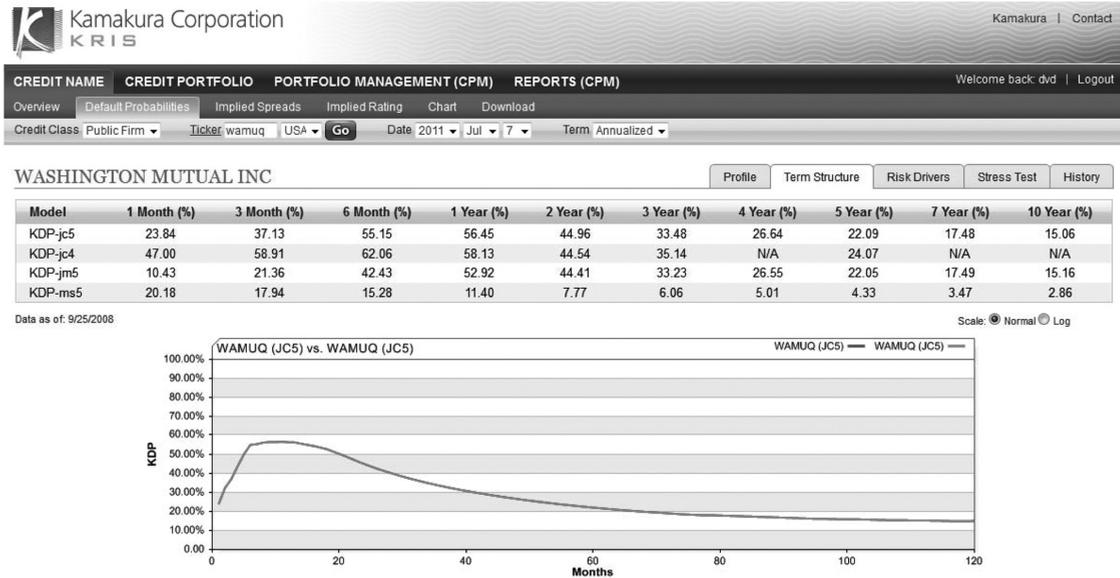


Figure 4 Annualized Term Structure of Default Probabilities for Washington Mutual: September 2008, One Day Prior to Default

coupon-bearing bonds, defaultable caps, floors, credit derivatives, and so on.

In practice, these hedge ratios are derived from a sophisticated simulation on “best practice” enterprise-wide risk management software. Van Deventer and Imai (2003) show that the steps in hedging the macro factor risk for any portfolio are identical to the steps that a trader of options has been taking for 30 years (hedging the net position with a long or short position in the common stock underlying the options):

- Calculate the change in the value (including the impact of interest rates on default) of all retail credits with respect to interest rates.
- Calculate the change in the value (including the impact of interest rates on default) of all small business credits with respect to interest rates.
- Calculate the change in the value (including the impact of interest rates on default) of all major corporate credits with respect to interest rates.

- Calculate the change in the value (including the impact of interest rates on default) of all bonds, derivatives, and other instruments.
- Add these “delta” amounts together.
- The result is the global portfolio “delta,” on a default-adjusted basis, of interest rates for the entire portfolio.
- Choose the position in interest rate derivatives with the opposite delta.
- This eliminates interest rate risk from the portfolio on a default-adjusted basis.

We can replicate this process for any macroeconomic factor that impacts default, such as home prices, exchange rates, stock price indices, oil prices, the value of class A office buildings in the central business district of key cities, and so on.

Most importantly,

- We can measure the default-adjusted transaction level and portfolio risk exposure with respect to each macroeconomic factor.
- We can set exposure limits on the default-adjusted transaction level and portfolio risk

exposure with respect to each macroeconomic factor.

- We know how much of a hedge would eliminate some or all of this risk.

The reason this analysis is so critical to success in credit risk portfolio management is the all-pervasiveness of correlated risk. Let us put aside the 2007–2009 credit crisis and look at other recent history. Take the Japan scenario. At the end of December 1989, the Nikkei stock price index had reached almost 39,000. Over the course of the next 14 years, it traded as low as 7,000. Commercial real estate prices fell by more than 60%. Single-family home prices fell in many regions for more than 15 consecutive years. More than 135,000 small businesses failed. Six of the 21 largest banks in Japan were nationalized in a span of two years. How would this approach have worked in Japan?

First of all, fitting a logistic regression for small businesses in Japan over this period shows that the properly specified inputs for the Nikkei and the yen/U.S. dollar exchange rates have *t*-score equivalents of more than 45 standard deviations from zero in a logistic regression. By *stress testing* a small business loan portfolio with this knowledge, we would have known how many put options on the Nikkei and put options on the yen were necessary to fully or partially offset credit-adjusted mark-to-market loan losses, just as the Federal Deposit Insurance Corporation announced it was doing in its 2003 Loss Distribution Model.⁹

This same approach works with

- Retail loan portfolios
- Small business loan portfolios
- Large corporate loan, bond, derivative, and other portfolios
- Sovereign and other government exposures

If common factors are found to drive each class of loans, then we have enterprise-wide correlations in defaults. An identical approach in the U.S. market would have spared many financial institutions tens of billions in losses that

resulted from an inability to do the stress tests described above and that are now mandated by the U.S. government and the European Central Bank.

The key to success in this analysis is a risk management software package that can handle it.¹⁰ What is also important in doing the modeling is to recognize that macroeconomic factors that are exchange traded (such as the S&P 500, home price futures, etc.) are much preferred to similar indicators that are not traded (such as the Conference Board index of leading indicators or the unemployment rate).

If one takes this approach, total balance sheet credit hedging is very practical

- Without using credit derivatives
- Without using first-to-default swaps
- Without using Wall Street as a counterparty from a credit risk point of view

All of these benefits are critical to answer the key question of “What’s the hedge?”

KEY POINTS

- It is not enough to know only the default risk of a counterparty. Over the full portfolio, a financial institution needs to know the answer to the question “What is the hedge?” if the measured credit risk is uncomfortably large.
- The major U.S. (2007–2009) and Japanese (1990–2002) financial institutions required government bailouts in the trillions of dollars because of their inability to measure and hedge macro factor risks like those of home price movements and commercial real estate price movements.
- The Merton model is a logical place to start thinking about how to hedge because of its simple structure and focus on the value of company assets.
- Unfortunately, for theoretical reasons alone, hedging in the Merton framework does not work for a company that is highly distressed. A perfect hedge could easily require a short

- position of more than 100% of the shares of outstanding common stock.
- The only practical and accurate approach to hedging credit risk is the reduced form modeling approach.
 - Hedging with credit default swaps is not practical because of the high degree of counterparty credit risk that is now obvious in the wake of the 2007–2009 credit crisis and the effective failures of investment banking firms like Bear Stearns, Lehman Brothers, Morgan Stanley, and Merrill Lynch. Moreover, trading volume in the credit default swap market is now so thin that large trades cannot be efficiently executed, and the risk of market manipulation is very high.
 - The reduced form approach explicitly links macro factors to both observable bond and CDS prices and to a historical default database. A similar approach links macro factors to credit spreads and recovery rates. The recovery on a mortgage that is in default is an example. Obviously, it depends on the value of the house that is collateral.
 - *Delta hedging* of aggregate portfolio exposure to these macro factors that drive credit risk is done in best practice enterprise-wide risk management software.
 - This modern application of stress testing, applied to a longer list of macro factors than interest rates alone, is not just theory. It is now mandated by the European Central Bank and U.S. regulatory authorities.
5. See www.dtcc.com for a list of dealers and related CDS trading volume.
 6. See Chapter 18 in van Deventer, Imai, and Mesler (2004) for a summary of the research in this area.
 7. The credit views page compares credit default swap spreads reported by Markit Partners with default probabilities from Kamakura Risk Information Services.
 8. van Deventer (2010).
 9. See press release dated December 10, 2003, on www.fdic.gov.
 10. See, for example, the Kamakura Risk Manager risk management software system.

REFERENCES

- Duffie, D., and Singleton, K. (1999). Modeling term structures of defaultable bonds. *Review of Financial Studies* 12: 197–226.
- Jarrow, R. A. (2001). Default parameter estimation using market prices. *Financial Analysts Journal* 57: 75–92.
- Jarrow, R. A., Lando, D., and Yu, F. (2005). Default risk and diversification: Theory and empirical applications. *Mathematical Finance* 15: 1–26.
- Jarrow, R. A., Li, L., Mesler, M., and van Deventer, D. R. (2007). The determination of corporate credit spreads. *Risk Magazine* (September).
- Jarrow, R. A., and van Deventer, D. R. (1998). Integrating interest rate risk and credit risk in asset and liability management. In *Asset and Liability Management: The Synthesis of New Methodologies*. London, UK: Risk Publications.
- Jarrow, R. A., and van Deventer, D. R. (1999). Practical usage of credit risk models in loan portfolio and counterparty exposure management: An update. In D. Shimko (ed.), *Credit Risk Models and Management*. London, UK: Risk Publications.
- Jarrow, R. A., and van Deventer, D. R. (2005). Estimating default correlations using a reduced form model. *Risk Magazine* (January).
- Lando, D. (2004). *Credit Risk Modeling*. Princeton, NJ: Princeton University Press.
- Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29: 449–470.

NOTES

1. Default probabilities presented in this chapter are supplied by Kamakura Corporation.
2. See Whitehouse (2005).
3. International Monetary Fund, Global Stability Report, as reported by the *Financial Times*, April 8, 2008.
4. See “Kamakura Risk Manager In Depth,” July 2011, available on www.kamakuraco.com for an example.

- Salmon, F. (2009). Recipe for disaster: The formula that killed Wall Street. *Wired Magazine*, February 23.
- van Deventer, D. R. (2010). Kamakura Blog: Daily trading volumes in credit default swaps. www.riskcenter.com.
- van Deventer, D. R., and Imai, K. (2003). *Credit Risk Models and the Basel Accords*. Hoboken, NJ: John Wiley & Sons.
- van Deventer, D. R., Imai, K., and Mesler, M. (2004). *Advanced Financial Risk Management: Tools and Techniques for Integrated Credit Risk and Interest Rate Risk Management*. Hoboken, NJ: John Wiley & Sons.
- Whitehouse, M. (2005). Slices of risk: How a formula that ignited market burned some big investors. *Wall Street Journal*, September 12, p. 1.

Derivatives Valuation

No-Arbitrage Price Relations for Forwards, Futures, and Swaps

ROBERT E. WHALEY, PhD

Valere Blair Potter Professor of Management and Co-Director of the Financial Markets Research Center, Owen Graduate School of Management, Vanderbilt University

Abstract: The three key factors that drive the valuation of a financial asset are risk, return, and timing of cash flows. A fundamental assumption in valuation is that in the absence of costless arbitrage opportunities, if two investments whose risk, return, and timing of cash flow properties are exactly the same are identified, they must have the same price in the marketplace. Otherwise, market participants can make free money by simultaneously selling the more expensive one and buying the cheaper one. This principle allows for the development of no-arbitrage price relations for forwards, futures, and swaps. The price of a futures contract is identical to the price of a forward contract in an environment in which short-term interest rates are known. In addition, a swap contract is nothing more than a portfolio of forward contracts. Hence, if a forward contract can be valued, a swap can be valued. The forward price and the underlying spot price are inextricably linked by the net cost of carry relation.

Exchange-traded and over-the-counter (OTC) derivatives contracts are traded worldwide. Of these, the lion's share is plain-vanilla forwards, futures, and swaps. The purpose of this entry is to develop no-arbitrage price relations for forwards, futures, and swap contracts. In doing so, we rely only on the assumption that two perfect substitutes must have the same price. The two substitutes, in this case, are a forward/futures contract and a levered position in the underlying asset. The key to understanding the forward/futures valuation lies in identifying the net cost of carrying (i.e., "buying and holding") an asset. We begin therefore with a discussion of carry costs/benefits. We then proceed by developing a number of important *no-arbitrage rela-*

tions governing forward and futures prices. Finally, we show that, since a swap contract is an exchange of future payments at a price agreed upon today, it can be valued as a portfolio of forward contracts.

UNDERSTANDING CARRY COSTS/BENEFITS

Derivative contracts are written on four types of assets—stocks, bonds, foreign currencies, and commodities. The derivatives literature contains seemingly independent developments of derivative valuation principles for each type

of asset. Generally speaking, however, the valuation principles are not asset-specific. The only distinction among assets is how carry costs/benefits are modeled.

The *net cost of carry* refers to the difference between the costs and the benefits of holding an asset. Suppose a breakfast cereal producer needs 5,000 bushels of wheat for processing in two months. To lock in the price of the wheat today, he can buy it and carry it for two months. One carry cost common to all assets is the opportunity cost of funds. To come up with the purchase price, he must either borrow money or liquidate existing interest-bearing assets. In either case, an interest cost is incurred. We assume this cost is incurred at the risk-free rate of interest. Beyond interest cost, however, carry costs vary depending upon the nature of the asset. For a physical asset or commodity such as wheat, we incur storage costs (e.g., rent and insurance). At the same time, certain benefits may accrue. By storing wheat we may avoid some costs of possibly running out of our regular inventory before two months are up and having to pay extra for emergency deliveries. This is called convenience yield. Thus, the net cost of carry for a commodity equals interest cost plus storage costs less convenience yield, that is,

$$\begin{aligned} \text{Net carry cost} &= \text{Cost of funds} + \text{Storage cost} \\ &\quad - \text{Convenience yield} \end{aligned}$$

For a financial asset or security such as a stock or a bond, the carry costs/benefits are different. While borrowing costs remain, securities do not require storage costs and do not have convenience yields. What they do have, however, is income (yield) that accrues in the form of quarterly cash dividends or semiannual coupon payments. Thus, the net cost of carry for a security is

$$\text{Net carry cost} = \text{Cost of funds} - \text{Income}$$

Carry costs and benefits are modeled either as *continuous rates* or as discrete flows. Some costs/benefits such as the cost of funds (i.e., the risk-free interest rate) are best modeled as continuous rates. The dividend yield on a broadly based stock portfolio, the interest income on a foreign currency deposit, and the lease rate on gold also fall into this category. Other costs/benefits such as warehouse rent payments for holding an inventory of grain, quarterly cash dividends on individual common stocks, and semiannual coupon receipts on a bond are best modeled as discrete cash flows. Below we provide the continuous rate and discrete flow cost of carry assumptions. For ease of exposition, we first introduce some notation. The current price of the asset is denoted S . Its price at future time T is \tilde{S}_T , where the tilde denotes the future asset price is uncertain. The opportunity cost of funds (i.e., the risk-free rate of interest) is assumed to be a constant, continuous rate and is denoted r . If we borrow to buy the asset today, we will owe Se^{rT} at time T .

Continuous Rates

The types of assets whose carry costs are typically modeled as constant, continuous rates include broadly based stock index portfolios, foreign currencies, and gold. Assume that we borrow at the risk-free rate of interest to buy a stock index portfolio that pays cash dividends at a constant continuous rate i . If we buy one unit of the index today and reinvest all dividends immediately as they are received in more shares of the index portfolio, the number of units of the index portfolio will grow to exactly e^{iT} units at time T . Alternatively, if we want exactly one unit of the index on hand at time T , we buy only e^{-iT} units today at a cost of Se^{-iT} . The terminal value of our investment in the index portfolio at time T will be \tilde{S}_T . The loan value has accrued from Se^{-iT} to $Se^{-iT}e^{rT} = Se^{(r-i)T}$. After repaying the loan, the terminal portfolio value will be $\tilde{S}_T - Se^{(r-i)T}$. Within this continuous

rate framework, the net cost of carry rate of an index portfolio equals the difference between the risk-free rate of interest r and the dividend yield rate i . The situation for a foreign currency is identical. If we borrow at the domestic risk-free rate, buy a foreign currency, and then invest the currency at the prevailing foreign risk-free rate, the net cost of carry rate equals the difference between the domestic interest rate r and the foreign interest rate i . Similarly, if we borrow at the risk-free rate, buy gold, and then lend it in the marketplace, the net cost of carry rate equals the difference between the interest rate r and the lease rate on gold i . Within this framework, the total cost of carry paid at time T is

$$\text{Net carry cost}_T = S[e^{(r-i)T}] - 1 \quad (1)$$

To illustrate, assume that the S&P 500 index is currently at a level of 1,100 and pays dividends at the continuous rate of 3% annually. Assume also that “shares” of the S&P 500 index can be purchased and sold at the index level (i.e., one share currently costs \$1,100). Suppose that an investor wants exactly 3,000 shares of the S&P 500 index on hand in five days. How many shares of the S&P 500 index must the investor buy today if all dividends paid are reinvested in more shares of the index portfolio?

If the investor wants 3,000 shares of the index on hand in five days, the investor needs to buy $3,000e^{-0.03(5/365)} = 2,998.77$ shares today. Over the first day, the number of shares will grow by a factor $e^{0.03(1/365)}$ due to the reinvestment of dividends, bringing the number of shares to $2,998.77e^{0.03(1/365)} = 2,999.01$. Over the second day, the number of shares will again grow by a factor $e^{0.03(1/365)}$ due to the reinvestment of dividends, bringing the number of shares to 2,999.26. Since the dividends are being paid at a constant, continuous rate, we know the original number of shares purchased will grow to exactly 3,000 shares by the end of day 5 (i.e., $3,000e^{0.03(5/365)}e^{-0.03(5/365)} = 3,000$), as is shown in the following table.

Day	Index Level	Units of Index	Value of Index Position
0	1,100.00	2,998.77	3,298,644
1	1,160.00	2,999.01	3,478,856
2	1,154.00	2,999.26	3,461,146
3	1,145.00	2,999.51	3,434,435
4	1,170.00	2,999.75	3,509,712
5	1,175.00	3,000.00	3,525,000

Discrete Flows

For most other types of assets including stocks with quarterly cash dividends and bonds with semiannual coupon payments, noninterest carry costs/benefits are best modeled as discrete flows. Suppose a stock promises to pay n known cash dividends in the amount I_i at time t_i , $i = 1, \dots, n$ between now and future time T . If we borrow S to cover the purchase price of the stock and reinvest all cash dividends as they are received at the risk-free rate of interest, the terminal value of our position will be

$$\tilde{S}_T + \sum_{i=1}^n I_i e^{r(T-t_i)} - S e^{rT}$$

In this instance, the net cost of carry at time T is

$$\text{Net carry cost}_T = S(e^{rT} - 1) - \sum_{i=1}^n I_i e^{r(T-t_i)}$$

For coupon-bearing bonds, the expressions are the same; however, S denotes the bond price and I_i at time t_i , $i = 1, \dots, n$ denote coupon payments.

To illustrate, an investor buys 10,000 shares of ABC Corporation and carries that position for 90 days. ABC's current share price is \$50, and the stock promises to pay a \$4 dividend in exactly 30 days. What will be the value of the portfolio when the investor unwinds in 90 days, assuming that the risk-free rate of interest is 5%? As Table 1 shows, the initial investment in 10,000 shares of ABC costs \$500,000. The investor financed the entire purchase price with risk-free borrowings, hence the initial investment is \$0. In 90 days, the investor has three components to the portfolio. First, the investor owns 10,000 shares valued at \tilde{S}_T a share.

Table 1 Future Value of Asset That Pays Discrete Cash Flows

Trade	Initial Investment	Value on Day T
Buy stock	$-50(10,000)$	$10,000\tilde{S}_T$
Borrow funds	500,000	$-500,000e^{0.05(90/365)} = -506,202.54$
Receive cash dividends on day t , and reinvest at risk-free rate until day T		$40,000e^{0.05(60/365)} = 40,330.12$
Value of position	0	$10,000\tilde{S}_T - 506,202.54 + 40,330.12$

Next, the investor must repay the \$500,000 in risk-free borrowings plus interest at a cost of \$506,202.54. Finally, the investor received cash dividends of \$4 a share or \$40,000 on day 30, which the investor invested immediately in risk-free discount bonds. Dividends plus accrued interest amount to \$40,330.12 on day T . Thus, the total value of the portfolio in 90 days is $10,000\tilde{S}_T - 506,202.54 + 40,330.12$.

Summary and Some Guidelines

Carry costs/benefits are the known costs/benefits associated with holding an asset over a fixed period of time. In general, they consist of two components—(1) interest and (2) income (in the case of a financial asset) or storage (in the case of a physical asset). The interest component is always expressed as a rate. If we buy an asset today with borrowed funds, we will owe e^{iT} per unit of the asset on day T . Income and noninterest costs are expressed either as a continuous proportion of the asset price or as discrete cash flows, depending upon the nature of the underlying asset. Firms potentially have four different sources of price risk—equity risk, interest rate risk, foreign exchange risk, and commodity price risk. Table 2 presents terminal values of leveraged asset positions using the net cost of carry assumption appropriate to each asset category.

VALUING FORWARDS

With the concept of net cost of carry in hand, we now turn to valuing forward contracts. A forward is a contract that requires its seller to

deliver the underlying asset on future day T at a price agreed upon today. We denote today's forward price as f . Its price on day T is denoted \tilde{f}_T . A forward with no time remaining to expiration must have the same price as the underlying asset, that is, $\tilde{f}_T = \tilde{S}_T$ as shown in Figure 1. Otherwise, a costless arbitrage profit is possible by buying the asset and selling the forward, or vice versa. The purpose of this section is to derive the value of a forward contract relative to its underlying asset price prior to time T under the continuous and discrete net carry cost assumptions.

Continuous Rates

To establish the price of a forward today, consider a U.S. corporation that needs to make a EUR 1,000,000 payment in T days and wants to lock in the U.S. dollar value of this payment today. The firm can accomplish this goal in two ways.

First, it can borrow U.S. dollars and buy euros today at the spot exchange rate S , and then carry the position for T days. To have one euro on hand in T days, they need to buy e^{-iT} units today where i is the risk-free interest rate in Europe. To finance the entire purchase today, they need to borrow Se^{-iT} . The repayment of the loan will occur in T days, and the principal plus interest will amount to $Se^{-iT}e^{iT}$ per euro where r is the U.S. risk-free interest rate.

Second, it can negotiate the price of a T -day forward contract with its bank. Under the terms of the forward contract, the firm will buy 1,000,000 euros in T days at a cost of f per euro. No money changes hands today. In making its

Table 2 Future value at time T of a leveraged asset position using continuous rate/discrete flow net cost of carry assumptions. All assets are assumed to incur interest cost at a constant continuous rate r .

Asset Type	Recommended Model	Terminal Value	Explanation of Noninterest Carry Costs/Benefits
Equity Individual stock or narrowly based stock portfolio	Discrete flow	$\tilde{S}_T - (Se^{rT} - FVI)$	For individual common stocks or narrowly based stock indexes, income accrues in the form of discrete cash dividends I_i paid at time t_i , and the dividends are carried forward until time T at the risk-free interest rate r , that is, $FVI = \sum_{i=1}^n I_i e^{r(T-t_i)}$
Stock portfolio	Continuous rate	$\tilde{S}_T - Se^{(r-i)T}$	For broadly based stock portfolios, income accrues at a constant, continuous dividend yield rate i .
Bonds	Discrete flow	$\tilde{S}_T - (Se^{rT} - FVI)$	For coupon-bearing bonds, income accrues in the form of discrete coupon payments I_i paid at time t_i , and the coupons are carried forward until time T at the risk-free interest rate r , that is, $FVI = \sum_{i=1}^n I_i e^{r(T-t_i)}$
Currency	Continuous rate	$\tilde{S}_T - Se^{(r-i)T}$	For foreign currency deposits, income accrues at a constant, continuous foreign rate of interest i .
Commodity	Discrete flow	$\tilde{S}_T - (Se^{rT} - FVI)$	For most physical commodities, storage costs (e.g., warehouse rent and insurance) are paid and convenience yield accrues. In this case, FVI is the future value of convenience yield less the future value of the discrete storage cost payments, that is, $FVI = FV(\text{Convenience yield}) - FV(\text{Storage costs})$
			While storage costs can be modeled as discrete flows, convenience yield (e.g., the lease rate on gold) may be modeled as a rate.

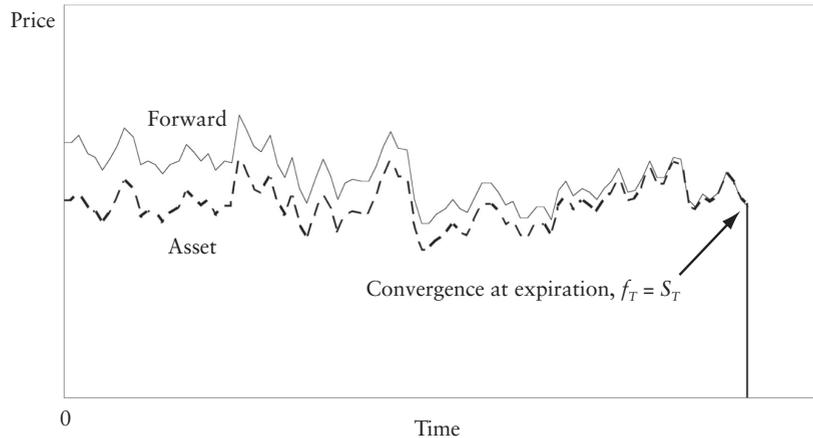


Figure 1 Price paths of forward contract and its underlying asset through time. Price convergence occurs at expiration.

decision about which strategy to take, the firm will compare the forward price with the future value had the euros been purchased today and carried until day T . If f exceeds $Se^{(r-i)T}$, the firm will buy the euros in the spot market and carry them. If f is less than $Se^{(r-i)T}$, the firm will buy the forward contract. Both alternatives provide the firm with EUR 1,000,000 in T days at a price locked in today. Since they are perfect substitutes, they must have the same price. The value of a forward in a constant continuous net cost of carry framework is

$$f = Se^{(r-i)T} \quad (2)$$

The relation (2) is sometimes called the *net cost of carry relation*. When the prices of the forward and the asset are such that (2) holds exactly, the forward market is said to be at full carry. Unless costless arbitrage is somehow impeded, we can be assured that the forward market will always be at full carry. Suppose, for an instant in time, $f > Se^{(r-i)T}$. Such a condition implies that there is a costless arbitrage opportunity. We should immediately sell the forward and buy the asset, financing the purchase of the asset with risk-free borrowing. Table 3 shows the outcome. With no investment today, we earn a certain outcome of $f - Se^{(r-i)T} > 0$ on day T . Naturally, the market cannot be in equilib-

rium. The costless arbitrage activity would continue until the selling pressure on the forward price and the buying pressure on the asset price makes the arbitrage profit equal to 0. Where no arbitrage opportunity exists, the cost of carry relation (2) holds.

The net cost of carry relation (2) is written in future value form, since both sides of the equation are values on day T , as shown in Table 3. The relation can also be expressed in present value form. Multiplying both sides of (2) by the discount factor e^{-rT} , we get

$$fe^{-rT} = Se^{-iT} \quad (3)$$

What (3) says is that the prepaid *forward contract*, fe^{-rT} , equals the initial cost of the asset position, Se^{-iT} .

Table 3 Costless Arbitrage Trades Where $f > Se^{(r-i)T}$

Trades	Initial Investment	Value on Day T
Buy e^{-iT} units of asset	$-Se^{-iT}$	\tilde{S}_T
Borrow (sell risk-free bonds)	Se^{-iT}	$-Se^{(r-i)T}$
Sell forward contract		$-(\tilde{S}_T - f)$
Net portfolio value	0	$f - Se^{(r-i)T}$

Discrete Flows

In the event that income or noninterest carry costs are more appropriately modeled as discrete cash flows, the net cost of carry relation is

$$f = Se^{rT} - FVI$$

where FVI is the future value of the promised income receipts. If the underlying asset is a physical asset, the future value of the income, FVI , may be negative as a result of storage cost payments. The relation can also be written in its present value form,

$$fe^{-rT} = S - PVI$$

where PVI is the present value of the promised income receipts, that is, $PVI = FVIe^{-rT}$. The prepaid forward price equals $S - PVI$, where the underlying asset distributes discrete known cash flows through time.

To illustrate, let's compute the value of a forward contract on a hypothetical dividend-paying stock, HAL Company. Specifically, we want to value a six-month forward contract on 3,000 shares of this company, assuming that the current share price is \$120 and that a \$3 cash dividend will be paid in two months and then again in five months. Assume the risk-free rate of interest is 5%. Since the cash dividend payments are discrete cash inflows, the cost of carry relation given by (1) is the most appropriate. The future value of the first dividend payment is $3e^{0.05(4/12)}$ given by (1) and the future value of the second dividend is $3e^{0.05(1/12)}$. The future value of all income received during the forward contract's life is therefore

$$FVI = 3e^{0.05(4/12)} + 3e^{0.05(1/12)} = 6.06$$

The value of the forward contract is therefore

$$f = 120e^{0.05(6/12)} - 6.06 = 116.97 \text{ per share}$$

or \$350,910 in total.

Hedging with Forwards

Before turning to futures contract valuation, it is worth considering the no-arbitrage portfolio in Table 3 more closely. It contains important

Table 4 Hedging a Stock Portfolio Using a Forward Contract

Trades	Initial Investment	Value on Day T
Own stock portfolio. Reinvest all dividend income into more shares of stocks.	$-S$	$\tilde{S}_T e^{iT}$
Sell e^{-iT} forward contract.	0	$-(\tilde{S}_T - f)e^{iT}$
Net portfolio value	0	fe^{iT}

intuition regarding hedging risk. Suppose that we hold a stock portfolio and fear that the market will decline over the next few months. To avoid the risk of a stock market decline, we can sell our stocks and buy risk-free bonds. Alternatively, we can sell a forward contract on our stock portfolio. These alternatives are perfect substitutes.

To see this, assume that our portfolio is sufficiently broad-based that it is reasonable to assume that the dividend yield is a constant continuous rate, i . If all dividend income is invested in more units of the stock portfolio, one unit in the stock portfolio today will grow to e^{iT} units on day T , as we discussed earlier and illustrated in Table 4. To hedge the price risk exposure of e^{iT} units of the stock portfolio on day T , we need to sell e^{iT} forward contracts today. The value of this forward position will be $-(\tilde{S}_T - f)e^{iT}$ on day T . Once the positions are netted, the terminal value of the portfolio is fe^{iT} . Note that the value is certain. The forward price, the dividend yield rate, and the hedge period horizon (i.e., the life of the forward contract) are all known on day 0. To see that the return on the hedged portfolio equals the risk-free return, substitute the net cost of carry relation, $f = Se^{(r-i)T}$, in the expression for the terminal value of the portfolio in Table 4. The net terminal value is $fe^{iT} = Se^{(r-i)T}e^{iT} = Se^{rT}$, exactly the amount we would have had if the stock portfolio had been liquidated and invested in risk-free bonds at the outset.

Table 5 Perfect Substitutes Implied by the Net Cost of Carry Relation

Position 1		Position 2
Buy asset/sell forward	=	Buy risk-free bonds (lend)
Buy risk-free bonds (lend)/buy forward	=	Buy asset
Buy asset/sell risk-free bonds (borrow)	=	Buy forward
Sell asset/buy forward	=	Sell risk-free bonds (borrow)
Sell risk-free bonds (borrow)/sell forward	=	Sell asset
Sell asset/buy risk-free bonds (lend)	=	Sell forward

Summary

A long forward position is a perfect substitute for buying the asset using risk-free borrowings. Consequently, the price of a forward equals the price of the asset plus net carry costs. But this is only one possible combination of positions in the asset, the forward, and risk-free bonds. Table 5 shows all possible pairings. Using the net cost of carry relation, we can demonstrate why Position 1 is a perfect substitute for Position 2 in all six rows of the table. A full understanding of each relation will prove invaluable in understanding valuation and risk management problems.

VALUING FUTURES

Futures contracts are like forward contracts, except that price movements are *marked-to-market* each day rather than waiting until contract expiration and having a single, once-and-for-all settlement. If the marking-to-market produces a gain during the futures contract's life, the gain can be reinvested in interest-bearing securities. Conversely, if the marking-to-market produces a loss, the loss must be covered with either existing interest-bearing assets or borrowing at the risk-free interest rate.

To distinguish between buying a forward and buying a futures, consider the futures position cash flows shown in Table 6. As we discussed earlier, a forward contract purchased today has a value $\tilde{S}_T - f$ on day T . In contrast, a futures contract is marked to market each day, and the daily gains/losses gather interest. If risk-free

rate of interest is 0%, the terminal value of the futures position (i.e., the sum of the mark-to-market gain/loss column) is the same as the terminal value of the forward position. If risk-free rate of interest is greater than 0%, however, the value of the futures position on day T may be greater or less than the terminal value of the forward position, depending on the path that futures prices follow over the life of the contract.

To illustrate, suppose that an investor needs £1,000,000 in three days and wants to lock in the price today. Suppose also that a three-day forward contract on British pounds is priced at \$1.60 per pound and that a British pound futures contract with three days remaining to expiration also has a price of \$1.60. Let's compare the terminal values of a long forward position with a long futures position at the end of three days assuming the domestic risk-free rate is 5%. Assume that the futures prices over

Table 6 Cash Flows of Long Futures Positions through Time

Day t	Futures Price	Mark-to-Market Gain/Loss on Day t	Value of Gain/Loss on Day T
0	F		
1	\tilde{F}_1	$\tilde{F}_1 - F$	$(\tilde{F}_1 - F)e^{r(T-1)}$
2	\tilde{F}_2	$\tilde{F}_2 - \tilde{F}_1$	$(\tilde{F}_2 - \tilde{F}_1)e^{r(T-2)}$
...			...
t	\tilde{F}_t	$\tilde{F}_t - \tilde{F}_{t-1}$	$(\tilde{F}_t - \tilde{F}_{t-1})e^{r(T-t)}$
...			...
$T-1$	\tilde{F}_{T-1}	$\tilde{F}_{T-1} - \tilde{F}_{T-2}$	$(\tilde{F}_{T-1} - \tilde{F}_{T-2})e^r$
T	\tilde{F}_T	$\tilde{F}_T - \tilde{F}_{T-1}$	$\tilde{F}_T - \tilde{F}_{T-1}$
Total		$\tilde{F}_T - \tilde{F}_1$	$\sum_{t=1}^T (\tilde{F}_t - \tilde{F}_{t-1})e^{r(T-t)}$

the next three days are \$1.71, \$1.67, and \$1.70, respectively.

The terminal value of a long forward position is simply the exchange rate on day 3, \$1.70, less the forward price, \$1.60, times one million, \$100,000, exactly equal to the sum of the mark-to-market gains/losses on the long futures position. The terminal value of the long futures position when the mark-to-market gains/losses are invested/financed at the risk-free rate of interest, however, is \$100,024.66, as is shown in the following table.

Day t	Futures Price	Mark-to-Market Gain/Loss on day t	Value of Gain/Loss on Day T
0	1.60		
1	1.71	110,000.00	110,030.14
2	1.67	-40,000.00	-40,005.48
3	1.70	30,000.00	30,000.00
Total		100,000.00	100,024.66

In general, the terminal value of a long forward and a long futures will be different. The reason that the terminal values are different is that the terminal value of the futures position depends on how the futures price evolves through time. Other futures price paths will produce different terminal values. If, for example, the futures price had been \$1.51 on day 1 rather than \$1.71, the terminal value of the futures position would have been \$99,997.26,

below (not above) the \$100,000 terminal value of the long forward.

Telescoping Futures Position

Interestingly, the fact that a long forward position does not have the same terminal value of a long futures position does not imply that the forward and futures prices are different. Indeed, as we will show shortly, they are equal. We can control the effect of the reinvestment of the mark-to-market proceeds by creating a “telescoping futures position.”

A telescoping futures position is created as follows. We begin, on day 0, with e^{-rT} futures contracts. Since we enter the position at the close of day 0, the marked-to-market gain for the day is 0. In preparation for day 1, we increase the size of the futures position by a factor e^r . At the end of day 1, the futures position is marked-to-market, generating proceeds of $e^{-r(T-1)}(\tilde{F}_1 - F)$. If this gain/loss is carried forward at the risk-free interest rate until day T , the terminal gain/loss will be $e^{-r(T-1)}(\tilde{F}_1 - F)e^{r(T-1)} = \tilde{F}_1 - F$, as shown in Table 7. On day 2, the position is again increased by a factor e^r and is marked-to-market at $e^{-r(T-2)}(F_2 - F_1)$. Carrying this amount forward to day T , we have $e^{-r(T-2)}(\tilde{F}_2 - \tilde{F}_1)e^{r(T-2)} = (\tilde{F}_2 - \tilde{F}_1)$, and so on. Because the number of futures is chosen to exactly offset the accumulated interest factor on the daily mark-to-market gain/loss, there will be exactly one futures contract on hand on day

Table 7 Cash Flows of Telescoping Futures Position Providing Same Terminal Value as Forward Position on Day T

Day t	Futures Prices	No. of Futures Contracts	Mark-to-Market Gain/Loss on Day t	Value of Gain/Loss on Day T
0	F	e^{-rT}		
1	\tilde{F}_1	$e^{-r(T-1)}$	$e^{-r(T-1)}(\tilde{F}_1 - F)$	$e^{-r(T-1)}(\tilde{F}_1 - F)e^{r(T-1)} = (\tilde{F}_1 - F)$
2	\tilde{F}_2	$e^{-r(T-2)}$	$e^{-r(T-2)}(\tilde{F}_2 - \tilde{F}_1)$	$\tilde{F}_2 - \tilde{F}_1$
...			...	
t	\tilde{F}_t	$e^{-r(T-t)}$	$e^{-r(T-t)}(\tilde{F}_t - \tilde{F}_{t-1})$	$\tilde{F}_t - \tilde{F}_{t-1}$
...			...	
$T-1$	\tilde{F}_{T-1}	e^{-r}	$e^{-r}(\tilde{F}_{T-1} - \tilde{F}_{T-2})$	$\tilde{F}_{T-1} - \tilde{F}_{T-2}$
T	\tilde{F}_T	1	$\tilde{F}_T - \tilde{F}_{T-1}$	$\tilde{F}_T - \tilde{F}_{T-1}$
Total				$\tilde{F}_T - F = \tilde{S}_T - F$

T , and the value of the futures position will be $S_T - F$. Assuming that the futures and forward contracts expire at the same time, the telescoping futures position will have exactly the same terminal value as the long forward position.

Using an illustration, let's compare terminal values of long forward and long telescoping futures positions. Suppose that an investor needs £1,000,000 in three days and wants to lock in the price today. Suppose also that a three-day forward contract on British pounds is priced at \$1.60 per pound and that a British pound futures contract with three days remaining to expiration also has a price of \$1.60. Assume that the domestic risk-free interest rate is 5% and that the futures prices over the next three days are \$1.71, \$1.67, and \$1.70, respectively.

As in the previous illustration, the terminal value of a long forward position is the exchange rate on day 3, \$1.70, less the forward price, \$1.60, times one million, or \$100,000. Because the initial futures position has less than 1 million units, the total of the mark-to-market gains/losses column is less than \$100,000. The terminal value of the telescoping futures position when the mark-to-market gains/losses are invested/financed at the risk-free rate of interest is exactly \$100,000, as is shown in the following table:

Day	Futures Price	Number of Units	Mark-to-Market Gain/Loss on day t	Value of Gain/Loss on Day T
0	1.60			
1	1.71	999,726.06	109,969.87	110,000.00
2	1.67	999,863.02	-39,994.52	-40,000.00
3	1.70	1,000,000.00	30,000.00	30,000.00
Total			99,975.35	100,000.00

The dynamic rebalancing of the futures position within the telescoping strategy ensures that the outcome is exactly the same as a long forward position.

Equivalence of Forward and Futures Prices

The fact that a long telescoping futures position has a terminal value of $\tilde{S}_T - F$ and that a long forward position has a terminal value of $S_T - F$ implies that the futures price and forward price must be equal to each other.¹ If they are not, a costless arbitrage profit would be possible by selling the forward and entering a long telescoping position in the futures (if $f > F$) or by buying the forward and entering a short telescoping position in the futures (if $f < F$). Given the equivalence of forward and futures prices, the valuation equations for a futures contract are the same as those of the forward, that is,

$$F = f = Se^{(r-i)T} \quad (4)$$

if all carry costs are constant continuous rates, and

$$F = f = Se^{rT} - FVI \quad (5)$$

if noninterest carry costs are discrete.

Let's illustrate how to short sell stocks synthetically using stock futures. Retail investors in the U.S. often find it costly to short sell shares of common stock. Consequently, stocks futures were recently launched. Assume that an investor wants to short sell a particular stock over the next T days. Its current share price is S , and a cash dividend of D has been declared and will be paid in t days. Let's demonstrate that selling a telescoping position in share futures is equivalent to short selling the stock.

First, the value in T days of a short position in the stock must be identified. Short selling a share of the stock generates proceeds of S . Assume that an investor can take the proceeds from the short sale and invest them at the risk-free rate of interest. In addition, the stock pays a cash dividend of D on day t . The investor is responsible for paying the cash dividend. On day T , the value of each security position

in the portfolio is as reported in the following table:

Trades	Initial Investment	Value on Day T
Short sell stock. Must pay cash dividends, if any.	S	$-\tilde{S}_T - De^{r(T-t)}$
Buy risk-free bonds	$-S$	Se^{rT}
Net portfolio value	0	$Se^{rT} - De^{r(T-t)} - \tilde{S}_T$

The net portfolio value on day T is $Se^{rT} - De^{r(T-t)} - \tilde{S}_T$.

From the discussion above, we know that selling a telescoping position in the share futures has a terminal value of $F - \tilde{S}_T$. But, from valuation equation (5), we know that, in the absence of costless arbitrage opportunities, $F = Se^{rT} - De^{r(T-t)}$. Substituting, we find that the value of the short futures position on day T is $Se^{rT} - De^{r(T-t)} - \tilde{S}_T$, an amount identical to that of the short stock position.

HEDGING WITH FUTURES

The telescoping futures position has implications in terms of hedging with futures contracts. For the hedge to be completely effective, the number of futures must equal the number of units of the underlying asset on day T . Under the continuous carry cost assumption, we know that one unit of the asset grows to e^{iT} units on day T . We also know that telescoping futures positions that starts with e^{-rT} futures contracts today has a single contract at time T . Consequently, to hedge the long asset position in Table 4, our futures hedge would start off with being short $e^{-(r-i)T}$ futures contract on day 0, and would scale up by a factor of e^r contracts per day over the life of the hedge. Assuming the futures expires on day T , the terminal value of the short telescoping position would be $-(\tilde{S}_T - F)e^{iT}$ and the net terminal value of the hedged portfolio would be Fe^{iT} . Substituting the

net cost of carry relation (4), the net terminal value of the hedged portfolio may be written Se^{rT} , which shows that hedging using a short telescoping futures position is equivalent to liquidating the asset position and buying risk-free bonds. The day-to-day increase in the size of the futures position by the interest factor e^r undoes the effects of interest on the daily marking to market of the futures gains/losses. In practice, this dynamic, day-to-day adjustment is called tailing the hedge.

SUMMARY

Futures contracts are like forward contracts except that price movements are marked to market daily. Because these daily gains/losses are allowed to accrue interest until the end of the contract's life, a long futures position will not in general have the same terminal value as a long forward position. The effects of the interest accrual on the mark-to-market gains/losses can be undone, however, using a telescoping futures position. Each day t , the number of futures is set equal to $e^{-r(T-t)}$ for each unit of the underlying asset at the end of the hedging interval. Set in this way, the terminal value of a long telescoping position in the futures equals the terminal value of a long forward. From a costless arbitrage perspective, therefore, the following are perfect substitutes:

Long telescoping futures position = Long forward position

Short telescoping futures position = Short forward position

The telescoping futures strategy also has implications for hedging. To undo the effects of interest on the daily marking to market of the futures gains/losses when the life of the futures matches the hedging horizon T , the size of a futures hedge starts at a level equal to the present value of the number of terminal units of that asset, that is, e^{-rT} for each unit of the asset and increases in size by a factor of e^r each day.

IMPLYING FORWARD NET CARRY RATES

Thus far, we have examined forward/futures contracts with a single maturity. A casual examination of the financial pages, however, shows multiple maturities for the same underlying asset. In these situations, we can use the net cost of carry relation (2) to deduce implied forward cost of carry rates.

VALUING SWAPS

A *swap* contract is an agreement to exchange a set of future cash flows. A plain-vanilla swap is usually regarded to be an exchange of a fixed payment for a floating payment, where the floating payment is tied to some reference rate, index level, or price. Like a forward contract, the underlying asset can be anything from a financial asset such as a stock or a bond to a physical asset such as crude oil or gold. Also, like a forward contract, a swap involves no up-front payment.

The key information needed to value a swap contract is the *forward curve* of the underlying asset and the zero-coupon yield curve for risk-free bonds. The forward curve refers to the relation between the price of a forward contract on the underlying asset and its time to expiration or settlement. Where the time to expiration is 0, the forward price equals the prevailing spot price. Figure 2 shows two possible forward curve relations. A *normal forward curve* is upward sloping, and an *inverted forward curve* is downward sloping. For financial assets, the slope will depend on the net difference between the risk-free rate and the income received on the underlying asset. Thus, a normal forward curve will arise in markets where the interest rate is greater than the income rate, and an inverted forward curve will arise in markets where the interest rate is less than the income rate. For physical assets or commodities, the nature of the forward curve depends also on

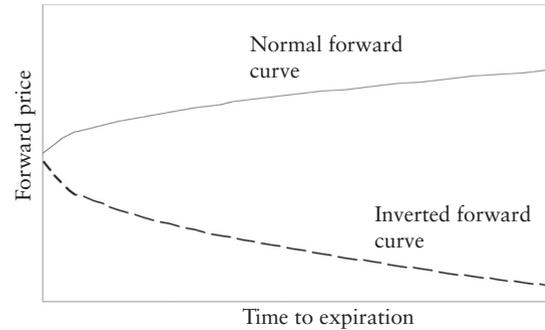


Figure 2 Forward curve: Relation between forward price and its time to expiration. Where time to expiration is 0, forward price equals spot price.

the cost of storage and convenience yield. The zero-coupon yield curve refers to the relation between interest rates and term to maturity.

In terms of swap valuation, the nature of the forward curve is irrelevant as long as the forward prices represent tradable prices. To see this, consider a jeweler (i.e., long hedger) who needs 1,000 troy ounces of gold each quarter over the next two years and wants to lock in his input cost today. One hedging alternative is to buy a strip of forward (or futures) contracts, one corresponding to each desired delivery date. The cost of the gold each quarter will be locked in; however, the cost of the gold will be different each quarter unless the forward curve is a horizontal line. The gold market, however, is typically in contango, so the cost, although certain, will escalate through time. A second alternative is to buy a swap contract that provides for the delivery of 1,000 ounces of gold each quarter, where there is single fixed price for all deliveries.² In the absence of costless arbitrage opportunities, it must be the case that the present value of the deliveries using the forward curve must be the same as the present value of the deliveries using the fixed price of the swap contract, that is,

$$\sum_{i=1}^n f_i e^{-r_i T_i} = \sum_{i=1}^n \bar{f} e^{-r_i T_i} \quad (6)$$

where n is the number of delivery dates, f_i is the price of a forward contract with time to expiration T_i , r_i is the risk-free rate of interest corresponding to time to expiration T_i ,³ and f is the fixed price in the swap agreement.⁴ In an instance where the right-hand side of (6) is greater (less) than the left-hand side, an arbitrageur would buy (sell) the swap and sell (buy) the strip of forward contracts, pocketing the difference. Because such free money opportunities do not exist, (6) must hold as an equality.

Equation (6) can be rearranged to isolate the fixed price of the swap agreement, that is,

$$\bar{f} = \frac{\sum_{i=1}^n f_i e^{-r_i T_i}}{\sum_{i=1}^n e^{-r_i T_i}} = \sum_{i=1}^n f_i \left(\frac{e^{-r_i T_i}}{\sum_{i=1}^n e^{-r_i T_i}} \right) \quad (7)$$

Expressed in this fashion, it becomes obvious that the fixed price of a swap is a weighted average of forward prices, one corresponding to each delivery date.⁵

KEY POINTS

- The net cost of carry is the cost of holding an asset over a period of time. One component of the cost of carry for all assets is the opportunity cost of funds. In order to buy the asset, an investor must pay for it.
- Beyond interest cost, however, carry costs may be positive or negative, depending upon the nature of the underlying asset. If the asset is a physical asset or commodity such as grain, the asset holder must pay storage costs such as warehouse rent and insurance. If the underlying asset is a financial asset or security such as a stock, a bond, or a currency, on the other hand, there are no storage costs. Instead, such assets produce a known income stream in the form of dividend payments or interest receipts, and this income can be used to subsidize the cost of borrowing.

- The interest cost is modeled as a constant continuous rate and the noninterest costs/benefits as either continuous rates or discrete cash flows, depending on the nature of the underlying asset.
- Given the assumption and definition of the cost of carry, pricing equations for forward and futures contracts can be developed. The price of a forward equals the price of a futures and both are equal to the asset price plus net carry costs. This is because if an investor needs an asset on hand at some future date at a price “locked-in” today, the investor can buy a forward contract, buy a futures, or buy the underlying asset and carry it.
- Perfect substitutes must have the same price.
- The relation between the forward curve and the fixed price of a swap is as follows. In the absence of costless arbitrage opportunities, the fixed price is a weighted average of the prices of the corresponding forward contracts, with the weights equal to the discount factor of each flow in relation to the sum of all discount factors.

NOTES

1. Cox, Ingersoll, and Ross (1981) use no-arbitrage arguments to demonstrate the equivalence of forward and futures prices when future interest rates are known. They go on to show, however, that if interest rates are uncertain, the futures price will be greater than or less than the forward price, depending upon whether the correlation between futures price changes and interest rate changes is negative or positive. See also Jarrow and Oldfield (1981).
2. As a practical matter, many swap agreements are cash-settled, so, instead of paying the fixed price per ounce and receiving 1,000 ounces in gold, we will receive in cash 1,000 times the difference between the prevailing (random) spot price of gold each

quarter and the fixed price. If the spot price is greater than the fixed price, we receive a cash payment from our counter-party, and vice versa.

3. Note that we are allowing for the fact that the risk-free rate may be term-specific.
4. The delivery quantity is irrelevant since it is the same on both sides of the equation. That is, equation (6) assumes that one unit is delivered on each delivery date.
5. For illustrations of how to compute the fixed rate of a swap based on the forward curve and the unwind price of swap based on

forward curve, see Chapter 4 in Whaley (2006).

REFERENCES

- Cox, J. C., Ingersoll, J. E., and Ross, S. (1981). The relation between forward and futures prices. *Journal of Financial Economics* 9: 321–346.
- Jarrow, R. A., and Oldfield, G. S. (1981). Forward contracts and futures contracts. *Journal of Financial Economics* 9: 373–382.
- Whaley, R. E. (2006). *Derivatives: Markets, Valuation, and Risk Management*. Hoboken, NJ: John Wiley & Sons.

No-Arbitrage Price Relations for Options

ROBERT E. WHALEY, PhD

Valere Blair Potter Professor of Management and Co-Director of the Financial Markets Research Center, Owen Graduate School of Management, Vanderbilt University

Abstract: For derivative instruments, in the absence of costless arbitrage price relations can be developed. In the case of options (calls and puts), there are three types of price relations that can be obtained. The first is the lower bound on the option's price. The second, and perhaps most important, no-arbitrage price relation is the one between the price of a put and the price of a call. This relation is called the put-call parity relation and arises from simultaneous trades in the call, the put, and the asset. The third price relation is the intermarket relation, which is the link between the prices of asset options and the prices of futures options. The price relations exist for European-style and American-style options and under both the continuous rate and discrete flow net cost of carry assumptions. Price relations are important for risk management strategies using options. Option pricing models go beyond these price relations to provide a fair value for an option.

The purpose of this entry is to develop *no-arbitrage price relations* for option contracts assuming that two perfect substitutes have the same price. In the absence of *costless arbitrage* opportunities, options have three types of no-arbitrage price relations—lower bounds, *put-call parity* relations, and intermarket relations. Each type of relation is developed in turn, for both European- and American-style options¹ and under both the continuous rate and discrete flow *net cost of carry* assumptions. Before deriving the no-arbitrage price relations for options, however, we focus on clearly distinguishing between the characteristics of option and forward contracts.

OPTIONS AND FORWARDS

Options differ from forwards in two key respects. First, the net cost of carry of a *forward* contract is zero since it involves no investment outlay. An option, on the other hand, involves investment. An option buyer pays the option premium for the right to buy or sell the underlying asset, and, like the buyer of any other asset, faces carry costs. For an option, however, the only carry cost is interest. Holding an option neither produces income like a dividend-paying stock nor requires storage costs like a commodity (i.e., a physical asset).

The effects of carry costs on the terminal profit functions of forward and option contracts are

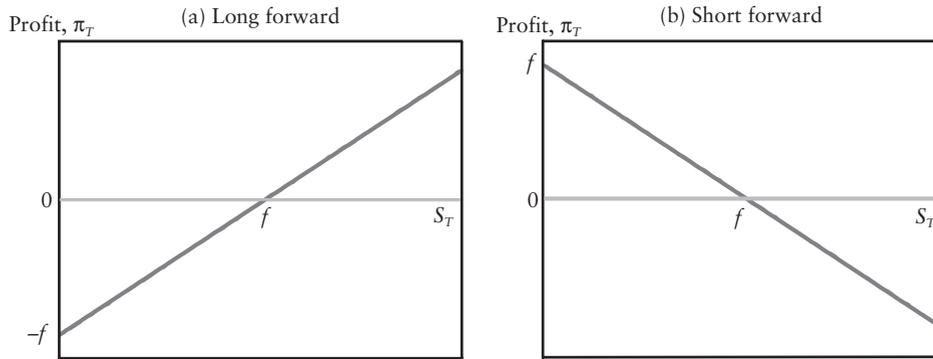


Figure 1 Terminal Profit of Long and Short Forward Positions

shown in Figures 1 through 3. The profit from a long forward position at expiration is

$$\pi_{\text{long forward}, T} = S_T - f \quad (1)$$

where S_T denotes the future price of the asset and f denotes the forward price.

On the other hand, the profit from a long call position is

$$\pi_{\text{long call}, T} = \begin{cases} S_T - X - ce^{rT}, & \text{if } S_T \geq X \\ -ce^{rT}, & \text{if } S_T < X \end{cases} \quad (2)$$

and from a long put position is

$$\pi_{\text{long put}, T} = \begin{cases} -pe^{rT}, & \text{if } S_T \geq X \\ X - S_T - pe^{rT}, & \text{if } S_T < X \end{cases} \quad (3)$$

where c and p are the prices of a European-style call and put, respectively; X is the exercise

price or strike price of the option. The opportunity cost of funds (i.e., the risk-free rate of interest) is denoted by r . Note that the profit functions for the long call and the long put (2) and (3) reflect the fact that the initial option premiums, c and p , are carried forward until the option's expiration at the risk-free interest rate. We have lost the opportunity cost of the funds we tied up in buying the option. Conversely, short call and short put positions (i.e., $\pi_{\text{short call}, T} = -\pi_{\text{long call}, T}$ and $\pi_{\text{short put}, T} = -\pi_{\text{long put}, T}$) reflect the fact that the option seller receives the premium payment and invests the cash at the risk-free interest rate. The profit function of a long forward position (1) has no interest component since the forward price is a promised payment on day T rather than a cash outlay today.

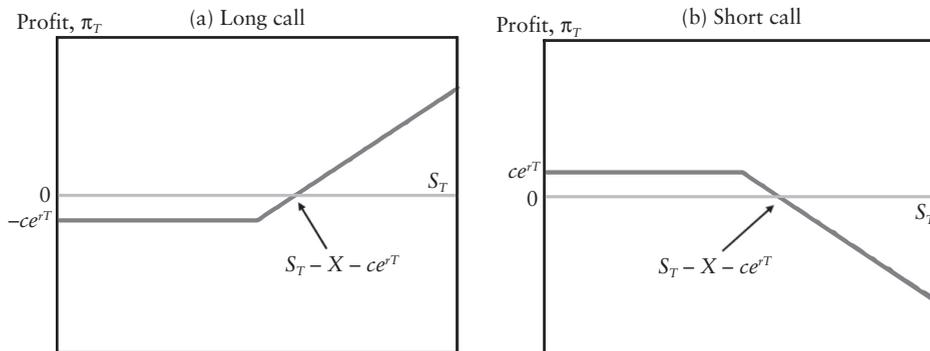


Figure 2 Terminal Profit of Long and Short Call Positions

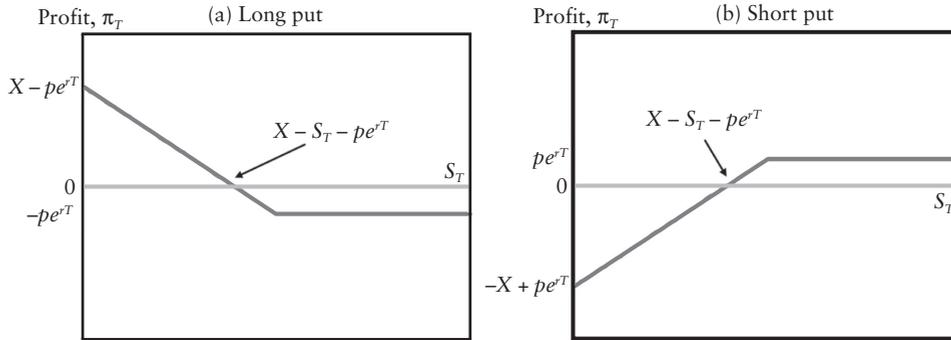


Figure 3 Terminal Profit of Long and Short Put Positions

The second key difference between forwards and options is that the buyer of a forward is obliged to buy the underlying asset at expiration, independent of whether or not the terminal asset price is greater than or less than the initial forward price. The buyer of an option, on the other hand, is not obliged to buy or sell the underlying asset, but will do so only when it is profitable. The profit function for the long call position (2), for example, shows that the option is exercised only when $S_T \geq X$. If $S_T < X$, the call option buyer chooses not to exercise, forfeiting only his original investment plus carry costs, ce^{rT} . The limited liability feature of the long call and long put positions are illustrated in Figures 2a and 3a, respectively. In the interest of completeness, the short positions in the respective instruments are illustrated in Figures 1b through 3b.

The profit functions of the *call* and the *put* show a certain complementarity to the profit function of a forward. Suppose we buy a call and sell a put at the same exercise price. The profit function for the overall position is

$$\begin{aligned} \pi_{c,T} - \pi_{p,T} &= \begin{cases} S_T - X - ce^{rT} + pe^{rT} & \text{if } S_T \geq X \\ S_T - X - ce^{rT} - pe^{rT} & \text{if } S_T < X \end{cases} \\ &= S_T - X - ce^{rT} - pe^{rT} \end{aligned}$$

Now, suppose that we chose the exercise price of the options such that $X = f - ce^{rT} + pe^{rT}$. The profit functions of the option portfolio and the long forward position will be exactly the same. If we buy the option portfolio and sell the

forward contract, the terminal value of the overall position must be 0. In the absence of costless arbitrage opportunities, the current value of the position must also be equal to 0, and, therefore, the call and put prices must be equal. Buying the call and selling the put (with the exercise price defined as above) is a perfect substitute for buying a forward. Viewed in this way, we can construct virtually any derivatives contract from any of the following pairs of basic instruments: (1) a forward and a call, (2) a forward and a put, and (3) a call and a put.

CONTINUOUS RATES

The *net cost of carry* refers to the difference between the costs and the benefits of holding an asset. One carry cost common to all assets is the opportunity cost of funds. We assume this cost is incurred at the risk-free rate of interest. Beyond interest cost, however, carry costs vary depending upon the nature of the asset. For a physical asset or commodity, we incur storage costs (e.g., rent and insurance). At the same time, certain benefits may accrue. By storing wheat we may avoid some costs of possible running out of our regular inventory before two months are up and having to pay extra for emergency deliveries. This is called *convenience yield*. Thus, the net cost of carry for a commodity equals interest cost plus storage costs less convenience yield.

Table 1 Arbitrage Portfolio Trades Supporting Lower Price Bound of European-Style Call Option Where the Underlying Asset Has a Continuous Net Carry Rate, $c \geq Se^{-iT} - Xe^{-rT}$

Trades	Initial Investment	Value on Day T	
		$S_T < X$	$S_T \geq X$
Sell asset	Se^{-iT}	$-\tilde{S}_T$	$-\tilde{S}_T$
Buy call option	$-c$	0	$-\tilde{S}_T - X$
Buy risk-free bonds	$-Xe^{-rT}$	X	X
Net portfolio value	$Se^{-iT} - Xe^{-rT} - c$	$X - \tilde{S}_T$	0

For a financial asset or security such as a stock or a bond, the carry costs/benefits are different. While borrowing costs remain, securities do not require storage costs and do not have convenience yields. What they do have, however, is income (yield) that accrues in the form of quarterly cash dividends or semiannual coupon payments. Thus, the net cost of carry for a security is equal to the cost of funds reduced by income. Carry costs and benefits are modeled either as continuous rates or as discrete flows. Some costs/benefits such as the cost of funds (i.e., the risk-free interest rate) are best modeled as continuous rates.

Under the continuous rate assumption, both interest cost and noninterest costs/benefits are modeled as continuous rates. Under the discrete flow assumption, interest cost is modeled as a continuous rate but noninterest costs/benefits are modeled as discrete cash flows. This section relies on the continuous rate assumption. The interest carry cost rate is represented by the notation r , and the noninterest carry benefit/cost rate is i . If the asset holder receives income from holding the asset such as the dividend yield on a stock portfolio or interest on a foreign currency investment, the income rate is positive (i.e., $i > 0$). If the asset holder pays costs in addition to interest in order to hold the asset (e.g., storage costs of holding a physical commodity), the income rate is negative (i.e., $i < 0$). Where $i = 0$, the only cost of carry is interest. As noted earlier in this section, the net cost of carry of an option is simply the interest rate.

Lower Price Bound of European-Style Call

Under the continuous rate assumption, the lower price bound of a European-style call option is

$$c \geq \max(0, Se^{-iT} - Xe^{-rT}) \quad (4)$$

The reason that the call price must be greater or equal to 0 is obvious—we do not have to be paid to take on a privilege. The reason the call price must exceed $Se^{-iT} - Xe^{-rT}$ is less obvious and is derived by means of an arbitrage portfolio. Suppose we form a portfolio by selling e^{-iT} units of the underlying asset² and buying a European-style call. In addition, to make sure that we have enough cash on hand to exercise the call at expiration, we buy Xe^{-rT} in risk-free bonds. The initial investment and terminal values of these positions are shown in Table 1. On day T , the net terminal value of the portfolio depends on whether the asset price is above or below the exercise price. If the asset price is less than the exercise price (i.e., $S_T < X$), we let the call expire worthless. We then use the risk-free bonds to buy one unit of the asset to cover the short sale obligation. What remains is $X - \tilde{S}_T$, which we know is greater than 0. If the asset price is greater than or equal to the exercise price (i.e., $S_T \geq X$), we exercise the call. This requires a cash payment of X . Fortunately we have exactly that amount on hand in the form of risk-free bonds. The unit of the asset that we receive upon exercising the call is used to retire

the short sale obligation. In this case, the net terminal value is certain to be 0.

What are the implications of this strategy? Well, we have formed a portfolio that is certain to have a terminal value of at least 0. In the absence of costless arbitrage opportunities, this implies that the greatest initial value is 0. More simply, we cannot reasonably expect to collect money at the outset without risk of loss. In the absence of costless arbitrage opportunities, $Se^{-iT} - Xe^{-rT} - c \leq 0$. Hence, a lower price bound for the European-style call is $c \geq Se^{-iT} - Xe^{-rT}$.³

In general, the lower price bound of an option is called its intrinsic value, and the difference between the option's market value (price)⁴ and its intrinsic value is called its time value. Thus a European-style call has an intrinsic value of $\max(0, Se^{-iT} - Xe^{-rT})$ and a time value of $c - \max(0, Se^{-iT} - Xe^{-rT})$. This entry deals with identifying intrinsic values by virtue of no-arbitrage arguments. Option pricing models uncover the determinants of time value.

To illustrate, suppose a three-month European-style call option written on a stock index portfolio has an exercise price of 70 and a market price of 4.25. Suppose also the current index level is 75, the portfolio's dividend yield rate is 4%, and the risk-free rate of interest is 5%. Is a costless arbitrage profit possible?

To test for the possibility of a costless arbitrage profit, substitute the problem parameters into the lower price bound (4), that is,

$$4.25 < \max[0, 75e^{-0.04(3/12)} - 70e^{-0.05(3/12)}] = 5.12$$

Since the lower bound relation is violated, a costless arbitrage profit of at least $5.12 - 4.25 = 0.87$ is possible. Since the violation may result from either the call being underpriced or the asset being overpriced, the arbitrage requires buying the call and selling the asset.⁵ The appropriate arbitrage trades are provided in Table 1. Substituting the prices and rates,

Trades	Initial Investment	Value at Time T	
		$ST < 70$	$ST \geq 70$
Sell index portfolio	74.25	$-\tilde{S}_T$	$-\tilde{S}_T$
Buy call option	-4.25	0	$\tilde{S}_T - 70$
Buy risk-free bonds	-69.13	70	70
Net portfolio value	0.87	$70 - \tilde{S}_T$	0

In examining the net portfolio value, note that you (a) earn an immediate profit of 0.87, and (b) have the potential of earning even more if the index level is below 70 at the option's expiration. If prices in the market were actually configured at such levels, you should expect that buying pressure on the call and selling pressure on the index portfolio would very quickly return the market to equilibrium. In the absence of costless arbitrage opportunities, $c \geq Se^{-iT} - Xe^{-rT}$.

Lower Price Bound of American-Style Call

American-style options are like European-style options except that they can be exercised at any time up to and including the expiration day. Since this additional right cannot have a negative value, the relation between the prices of American-style and European-style call options is

$$C \geq c \tag{5}$$

where the uppercase C represents the price of an American-style call option with the same exercise price and time to expiration and on the same underlying asset as the European-style call. The lower price bound of an American-style call option is

$$C \geq \max(0, Se^{-it} - Xe^{-rt}, S - X) \tag{6}$$

This is the same as the lower price bound of the European-style call (4), except that the term $S - X$ is added within the maximum value operator on the right-hand side. The reason is, of course, that the American-style call cannot sell for less than its immediate early exercise proceeds, $S - X$. If $C < S - X$, a costless arbitrage profit

of $S - X - C$ can be earned by simultaneously buying the call (and exercising it) and selling the asset.

As an illustration, suppose a three-month American-style call option written on a stock index portfolio has an exercise price of 70 and a market price of 4.25. Suppose also the current index level is 75, the portfolio's dividend yield rate is 4%, and the risk-free rate of interest is 5%. Is a costless arbitrage profit possible?

To test for the possibility of a costless arbitrage profit, substitute the problem information into (6), that is,

$$\begin{aligned} 4.25 &< \max[0, 75e^{-0.04(3/12)} \\ &\quad - 70e^{-0.05(3/12)}, 75 - 70] \\ &= \max(0, 5.12, 5) = 5.12 \end{aligned}$$

At the current call price of 4.25, two types of arbitrage are possible. A costless arbitrage profit of $5.00 - 4.25 = 0.75$ is possible simply by buying the call, exercising it, and selling the asset. The amount of this arbitrage profit, however, is less than the arbitrage profit of at least $5.12 - 4.25 = 0.87$ that can be earned by buying the call, selling the asset, buying risk-free bonds, and holding the portfolio until the call's expiration, as was shown in the previous arbitrage table. Under this second alternative, you earn an immediate profit of 0.87, and have the potential of earning even more if the asset price is below 70 at the option's expiration.

Early Exercise of American-Style Call Options

The structure of the lower price bound of the American-style call (6) can be used to provide important insight regarding the possibility of *early exercise*. The second term in the squared brackets, $Se^{-iT} - Xe^{-rT}$, is the minimum price at which the call can be sold in the marketplace.⁶ The third term is the value of the American-style if it is exercised immediately. If the value

of the second term is greater than the third term (for a certain set of call options), the call's price in the marketplace will be always exceed its exercise proceeds so it will never be optimal to exercise the call early.

To identify this set of calls, we must examine the conditions under which the relation

$$Se^{-iT} - Xe^{-rT} > S - X$$

holds. The job is easier if we rearrange the relation to read

$$S(e^{-iT} - 1) > -X(1 - e^{-rT}) \quad (7)$$

Since the risk-free interest rate is positive, the expression of the right-hand side is negative. If the left-hand side is positive or zero, the call option holder can always get more by selling his option in the marketplace than by exercising it; so early exercise will never be optimal and the value of the American-style call is equal to the value of the European-style call, $C = c$. This condition is met for calls whose underlying asset has a negative or zero noninterest carry rate, $i \leq 0$.

The intuition for this result can be broken down into two components—interest cost, r , and noninterest benefit (i.e., $i > 0$) or cost (i.e., $i < 0$). With respect to interest cost, recognize that exercising the call today requires that we pay X today. If we defer exercise until the call's expiration, on the other hand, we have the opportunity to earn interest (i.e., our liability is only the present value of the exercise cost, Xe^{-rT}). So, holding other factors constant, we always have an incentive to defer exercise.⁷ With respect to the noninterest costs, recall that assets with $i < 0$ are typically physical assets that require storage. If we exercise a call written on such an asset, we must take delivery, whereupon we immediately begin to incur storage costs. If we defer exercise, on the other hand, and continue to hold the claim on the asset rather than the asset itself, we avoid paying storage costs. Thus, where $i < 0$, there are two reasons not to exercise early. But even if storage

costs are zero (i.e., with $i = 0$), condition (7) holds since the interest cost incentive remains.

For American-style call options on assets with $i > 0$ (e.g., stock index portfolio with a nonzero dividend yield and foreign currencies with a nonzero foreign interest rate), on the other hand, early exercise may be optimal. The intuition is that, while there remains the incentive to defer exercise and earn interest on the exercise price, deferring exercise means forfeiting the income on the underlying asset (e.g., the dividend yield on a stock index portfolio). The only way to capture this income is by exercising the call and taking delivery of the asset. For American-style call options on assets with $i > 0$, early exercise may be optimal and, therefore, $C > c$.

Lower Price Bound of European-Style Put

The lower price bound of a European-style put option is

$$p \geq \max(0, Xe^{-rT} - Se^{-iT}) \tag{8}$$

Again, the reason that the option price must be greater or equal to 0 is obvious—we do not have to be paid to take on a privilege. The reason the put price must exceed the bound, $Xe^{-rT} - Se^{-iT}$, is given by the arbitrage trade portfolio in Table 2. If we buy e^{-iT} units of the asset and a put, and sell Xe^{-rT} risk-free bonds, the net terminal value of the portfolio is certain to be greater than or equal to 0. If the asset price is less than or equal to the exercise price at the

option’s expiration (i.e., $S_T \leq X$), we will exercise the put, delivering the asset and receiving X in cash. We will then use the exercise proceeds X to cover our risk-free borrowing obligation. In the event the asset price is greater than the exercise price (i.e., $S_T > X$), we will consider the put expire worthless. We still need to cover our risk-free borrowing, which we do by selling the asset. After repaying our debt, we have $\tilde{S}_T - X$ remaining.

For example, a three-month European-style put option written on a stock index portfolio has an exercise price of 70 and a market price of 8.80. Suppose also the current index level is 61, the portfolio’s dividend yield rate is 4%, and the risk-free rate of interest is 5%. Is a costless arbitrage profit possible?

To test for the possibility of a costless arbitrage profit, substitute the problem parameters into the lower price bound (8),

$$8.80 > \max[0, 70e^{-0.05(3/12)} - 61e^{-0.04(3/12)}] = 8.74$$

At the current price of 8.80, the no-arbitrage condition (8) holds, so no costless arbitrage opportunity exists.

Lower Price Bound for American-Style Put

An American-style put has an early exercise privilege, which means that the relation between the prices of American-style and European-style put options is

$$P \geq p \tag{9}$$

Table 2 Arbitrage Portfolio Trades Supporting Lower Price Bound of European-Style Put Option Where the Underlying Asset Has a Continuous Net Carry Rate, $p \geq Xe^{-rT} - Se^{-iT}$

Trades	Initial Investment	Value on Day T	
		$S_T < X$	$S_T \geq X$
Sell asset	$-Se^{-iT}$	\tilde{S}_T	\tilde{S}_T
Buy call option	$-p$	$X - \tilde{S}_T$	0
Buy risk-free bonds	Xe^{-rT}	$-X$	$-X$
Net portfolio value	$Xe^{-rT} - Se^{-iT} - p$	0	$\tilde{S}_T - X$

where uppercase P represents the price of an American-style put option with the same exercise price, time to expiration, and underlying asset as the European-style put. The lower price bound of an American-style put option is

$$p \geq \max(0, Xe^{-rT} - Se^{-iT}, X - S) \quad (10)$$

This is the same as the lower price bound of the European-style put (8), except that, because the American-style put may be exercised at any time including now, the exercise proceeds, $X - S$, is added within the maximum value operator on the right-hand side. If $P < X - S$, a costless arbitrage profit of $X - S - P$ can be earned by simultaneously buying the put (and exercising it) and buying the asset.

To illustrate, assume that a three-month American-style put option written on a stock index portfolio has an exercise price of 70 and a market price of 8.80. Suppose also the current index level is 61, the portfolio's dividend yield rate is 4%, and the risk-free rate of interest is 5%. Is a costless arbitrage profit possible?

To test for the possibility of a costless arbitrage profit, substitute the problem information into (10), that is,

$$\begin{aligned} 8.80 &< \max[0, 70e^{-0.05(3/12)} \\ &\quad - 61e^{-0.04(3/12)}, 70 - 61] \\ &= \max(0, 8.74, 9.00) = 9.00 \end{aligned}$$

At the current price of 8.80, the no-arbitrage relation (10) is violated, indicating the presence of a costless arbitrage opportunity. Since it is the early exercise condition (third term) on the right-hand side that is violated, you should buy the put (and exercise it) and buy the index portfolio. You would pay 8.80 for the put and 61 for the index portfolio, and receive 70 when you deliver the index portfolio upon exercising the put. The amount of the arbitrage profit is 0.20 and is earned immediately.

Early Exercise of American-Style Put Options

In the case of an American-style call, we found that if the underlying asset had carry costs and above interest (e.g., storage), the call option holder would never (rationally) exercise early. In the case of an American-style put, no comparable condition exists.⁸ There is always some prospect of early exercise, so the American-style put is always worth more than the European-style put, that is, $P > p$. The intuition is straightforward. Suppose, for whatever reason, the asset price falls to 0. The put option holder should exercise immediately. There is no chance that the asset price will fall further, so delaying exercise means forfeiting the interest income that can be earned on the exercise proceeds of the put, X . The interest-induced, early-exercise incentive works in exactly the opposite way for the put than it did for the call. For the put, we want to exercise early to get the cash and let it begin to earn interest. For the call, we want to defer exercise and let the cash continue to earn interest.

Put-Call Parity for European-Style Options

Perhaps the most important no-arbitrage price relation for options is put-call parity.⁹ The put-call parity price relation arises from the simultaneous trades in the call, the put, and the asset. Put-call parity for European-style options is given by

$$c - p = Se^{-iT} - Xe^{-rT} \quad (11)$$

The composition of the put-call parity arbitrage portfolio is given in Table 3. A portfolio that consists of a long position of e^{-iT} units of the asset, a long put, a short call, and a short position of Xe^{-rT} in risk-free bonds is certain to have a net terminal value of 0. If the terminal asset price is less than or equal to the exercise price of the options (i.e., $S_T \leq X$), we exercise the put and deliver the asset. The cash proceeds

Table 3 Arbitrage Portfolio Trades for European-Style Put-Call Parity Where the Underlying Asset Has a Continuous Net Carry Rate, $c - p = Se^{-iT} - Xe^{-rT}$

Trades	Initial Investment	Value at Time T	
		$S_T < X$	$S_T \geq X$
Buy asset	$-Se^{-iT}$	\tilde{S}_T	\tilde{S}_T
Buy put option	$-p$	$X - \tilde{S}_T$	0
Sell call option	c	0	$-(\tilde{S}_T - X)$
Sell risk-free bonds	Xe^{-rT}	$-X$	$-X$
Net portfolio value	$Xe^{-rT} - Se^{-iT} - p + c$	0	0

from exercise are used to repay our debt. The call option is out-of-the-money, so the call option holder will let it expire worthless. On the other hand, if the terminal asset price exceeds the exercise price (i.e., $S_T > X$), we will let our put expire worthless. The call option holder will exercise, requiring that we deliver a unit of the asset, which we just happen to have.¹⁰ The call option holder pays us X , which we use to retire our risk-free borrowings. Since the net terminal portfolio value is zero, the cost of entering into such a portfolio today must also be 0, otherwise costless arbitrage would be possible. If the initial investment is 0, the put-call parity relation (11) holds.

The set of arbitrage trades spelled out in Table 3 (i.e., buy the asset, buy the put, sell the call, and sell risk-free bonds) is called a conversion. If all of the trades are reversed (i.e., sell the asset, sell the put, buy the call, and buy risk-free bonds), it is called a reverse conversion. These names arise from the fact that we can create any position in the asset, options, or risk-free bonds by trading (or converting) the remaining secu-

rities, in the same manner we used a call and a put to create a forward contract at the beginning of the entry. Table 4 provides a complete list of the conversions that are possible using the put-call parity relation for European-style options. The first row says that buying the asset, buying a put, and selling a call is equivalent to buying risk-free bonds. We can check this by creating an arbitrage trade table, or by simply working through it mentally. If the asset price is less than the exercise price at expiration, we will exercise our put and sell the asset. If the asset price is greater than the exercise price, the call option holder will exercise, requiring that we deliver the asset. In both cases, we are certain to have X in cash when all is said and done. This is the same as the amount we would have had if we bought risk-free bonds.

Let's see how put-call parity is applied for European-style options. Suppose that a three-month call and put with an exercise price of 70 have prices of 5.00 and 4.50, respectively. Suppose also that the current level of the index

Table 4 Perfect Substitutes Implied by European-Style Put-Call Parity

Position 1		Position 2
Buy asset/buy put/sell call	=	Buy risk-free bonds (lend)
Buy asset/buy put/sell risk-free bonds	=	Buy call
Sell asset/buy call/buy risk-free bonds	=	Buy put
Sell put/buy call/buy risk-free bonds	=	Buy asset
Sell asset/sell put/buy call	=	Sell risk-free bonds (borrow)
Sell asset/sell put/buy risk-free bonds	=	Sell call
Buy asset/sell call/sell risk-free bonds	=	Sell put
Buy put/sell call/sell risk-free bonds	=	Sell asset

portfolio underlying the options is 70, the index portfolio has a dividend yield rate of 3%, and the risk-free rate of interest is 5%. Is a costless arbitrage profit possible?

To test for the possibility of a costless arbitrage profit, substitute the problem parameters into the put-call parity relation (11),

$$5.00 - 4.50 = 0.50 > 70e^{-0.03(3/12)} - 70e^{-0.05(3/12)} = 0.34$$

Since the equation does not hold, a costless arbitrage profit is possible. Since the violation may result from either the call being overpriced, the put being underpriced, or the asset being underpriced, the arbitrage will require all three trades: selling the call, buying the put, and buying the asset. Using the trades as set out in Table 3, we get:

Trades	Initial Investment	Value at Time T	
		$S_T < 70$	$S_T \geq 70$
Buy asset	-69.48	\tilde{S}_T	\tilde{S}_T
Buy put option	-4.50	$\tilde{S}_T - 70$	0
Sell call option	5.00	0	$-(\tilde{S}_T - 70)$
Sell risk-free bonds	69.13	-70	-70
Net portfolio value	0.16	0	0

By forming this portfolio, we generate a costless arbitrage profit of 0.16. The buying pressure on the index portfolio and the put will cause their prices to rise, and the selling pressure on the call will cause its price to fall. The arbitrage trading will stop when the initial value investment column sums to zero (i.e., the costless ar-

bitrage opportunity ceases to exist), or where $c - p = Se^{-iT} - Xe^{-rT}$.

Put-Call Parity for American-Style Options

The early exercise feature of American-style options complicates the put-call parity relation. The nature of the relation depends on the level of noninterest costs/benefits, i . Specifically, the put-call parity relations are

$$S - X \leq C - P \leq Se^{-iT} - Xe^{-rT} \quad \text{if } i = 0 \quad (12a)$$

and

$$Se^{-iT} - X \leq C - P \leq S - Xe^{-rT} \quad \text{if } i > 0 \quad (12b)$$

Each inequality in (12a) and in (12b) has a separate set of arbitrage trades. To illustrate, consider (12b), the case in which the asset pays some form of income, say, a stock index portfolio with a constant dividend yield rate, or a foreign currency with a constant foreign risk-free rate of interest. To establish the left-hand side inequality of (12b), consider the arbitrage portfolio trades in Table 5. To generate the table entries, assume the left-hand side inequality of (12b) is reversed. This means the asset price is overpriced, the put is overpriced, and/or the call is underpriced. Thus, the arbitrage portfolio must account for all three possibilities. We should sell the asset, sell the put, buy the call, and buy some risk-free bonds. At the options' expiration, the portfolio is certain to have

Table 5 Arbitrage Portfolio Trades Supporting American-Style Put-Call Parity Where the Underlying Asset Has a Continuous Net Carry Rate, $Se^{-iT} - X < C - P$

Trades	Initial Investment	Early Exercise at t	Value on Day T	
			$S_T < X$	$S_T \geq X$
Sell asset	$-Se^{-iT}$	$-\tilde{S}_t e^{-i(T-t)}$	$-\tilde{S}_T$	$-\tilde{S}_T$
Sell put option	P	$-(X - \tilde{S}_t)$	$-(X - \tilde{S}_T)$	0
Buy call option	$-C$	$-\tilde{C}_t$	0	$\tilde{S}_T - X$
Buy risk-free bonds	$-X$	Xe^{rt}	Xe^{rT}	Xe^{rT}
Net portfolio value	$Se^{-iT} + P - C - X$	$\tilde{S}_t[1 - e^{-i(T-t)}] + \tilde{C}_t + X(e^{rT} - 1)$	$X(e^{rT} - 1)$	$X(e^{rT} - 1)$

positive value $X(e^{rT} - 1)$. If $S_T < X$, the put option holder exercises, requiring that we pay X in return for a unit of the underlying asset. We pay the exercise price using a portion of our risk-free bonds, and use the delivered asset to cover our short position. On the other hand, if $S_T \geq X$, we exercise the call and receive the asset. The asset delivered on the call is used to cover the short position. We use some of the risk-free bonds to pay for the exercise price of the call.

The early exercise feature of the American-style options requires that we consider one other contingency within the arbitrage table, that is, what happens if the put option holder decides to exercise early at some arbitrary time t between now and expiration. Looking at Table 5, we see that our obligation should the put be exercised early is $-(X - \tilde{S}_t)$. But since we have Xe^{rt} in risk-free bonds, we have more than enough to cover the payment of X to the put option holder. In return, we receive \tilde{S}_t , which is more than enough to cover our short asset position in the asset that has value $-\tilde{S}_t e^{-i(T-t)}$. In addition, we have a long position in the call with value \tilde{C}_t . Because the net portfolio value is positive at expiration and also in the event the put is exercised early, the initial investment must be negative (since if it were zero or positive, there would be a certain arbitrage). And, if $Se^{-iT} - X - C + P < 0$, then $Se^{-iT} - X < C + P$.

To establish the right-hand side inequality of (12b), consider the arbitrage portfolio trades in Table 6. To generate the table entries, again assume the right-hand side inequality of (9b)

is reversed. This means the asset price is underpriced, the put is underpriced, and/or the call is overpriced. The arbitrage portfolio trades must account for all possibilities. We should buy the asset, buy the put, sell the call, and sell some risk-free bonds. At the options' expiration, the portfolio is certain to have positive value $\tilde{S}_T(e^{iT} - 1)$. If $S_T < X$, we exercise the put and sell the asset. The long asset position has a value $\tilde{S}_T e^{iT}$, which is more than enough to pay for the unit of the asset owed on the put. The cash received from exercising the put is used to cover our risk-free bond obligation. On the other hand, if $S_T \geq X$, the call option holder exercises, implying that we receive X in return for delivering one unit of the asset. We use the call received from the call option holder to retire the risk-free bond position. The value of our asset position, $\tilde{S}_T e^{iT}$, is more than we need to deliver on the put.

The early exercise feature of the American-style call must also be considered, that is, what happens if the call option holder decides to exercise early on day t ? Looking at Table 6, we see that the call exercise obligation is $-(\tilde{S}_t - X)$. But, if we receive X , that is more than enough to cover the balance of $-Xe^{-r(T-t)}$ in risk-free bonds. We must pay \tilde{S}_t , but we have more than one unit of the asset, that is, $\tilde{S}_t e^{i(T-t)}$. In addition, we have a long position in the put with value \tilde{P}_t . Since the net portfolio value is positive at expiration and in the event the call is exercised early, the initial investment must be negative. And, if $-S + Xe^{-rT} + C - P < 0$, $C - P < S - Xe^{-rT}$.

Table 6 Arbitrage Portfolio Trades Supporting American-Style Put-Call Parity Where the Underlying Asset Has a Continuous Net Carry Rate, $C - P < S - Xe^{-rT}$

Trades	Initial Investment	Early Exercise at t	Value on Day T	
			$S_T < X$	$S_T \geq X$
Buy asset	$-S$	$\tilde{S}_t e^{iT}$	$\tilde{S}_T e^{iT}$	$\tilde{S}_T e^{iT}$
Buy put option	$-P$	\tilde{P}_t	$X - \tilde{S}_T$	0
Sell call option	C	$-(\tilde{S}_t - X)$	0	$-(\tilde{S}_T - X)$
Sell risk-free bonds	Xe^{rt}	$-Xe^{-r(T-t)}$	$-X$	$-X$
Net portfolio value	$-S - P + Xe^{rT} + C$	$\tilde{S}_t(e^{-it} - 1) + \tilde{P}_t + X[1 - e^{-r(T-t)}]$	$\tilde{S}_T(e^{iT} - 1)$	$\tilde{S}_T(e^{iT} - 1)$

Table 7 No-Arbitrage Price Relations For European- and American-Style Options Where the Underlying Asset Has a Continuous Net Carry Rate

Description	European-Style Options	American-Style Options
Lower price bound for call	$c \geq \max(0, Se^{-iT} - Xe^{-rT})$	$C \geq \max(0, Se^{-iT} - Xe^{-rT}, S - X)$
Lower price bound for put	$p \geq \max(0, Xe^{-rT} - Se^{-iT})$	$P \geq \max[0, Xe^{-rT} - Se^{-iT}, X - S]$
Put-call parity relation	$c - p = Se^{-iT} - Xe^{-rT}$	$S - X < C - P < Se^{-iT} - Xe^{-rT}, \text{ if } i \leq 0$ $Se^{-iT} - X < C - P < S - Xe^{-rT}, \text{ if } i > 0$

Summary

This completes the derivations of no-arbitrage price relations for European-style and American-style options on assets with a continuous net carry rate. For convenience, a summary of the no-arbitrage relations is provided in Table 7.

DISCRETE FLOWS

With the no-arbitrage price relations for an underlying asset with a continuous carry cost rate in hand, the focus now turns to developing the same set of relations for an asset that has interest cost modeled as a continuous rate but noninterest costs/benefits modeled as a discrete flow. If the noninterest flow is income such as in the case of a cash dividend payment on a share of stock or a coupon payment on a bond, the income is represented as a positive value, that is, $I_t > 0$. If the flow is a cost such as, say, warehouse rent from storing an inventory of wheat, the income is represented as a negative value, that is, $I_t < 0$. Again, since this book deals primarily with financial assets, most of the illustrations will have I_t discussed as being a positive value. Although I_t represents a cash payment on any

type of asset, we will call I_t a dividend payment throughout this section for expositional convenience.

Lower Price Bound of European-Style Call

The lower price bound of a European-style call option on an asset that makes a single, discrete cash dividend payment during the option's life is

$$c \geq \max(0, S - I_t e^{-rt} - Xe^{-rT}) \quad (13)$$

In this relation, $I_t e^{-rt}$ is the present value of the promised dividend to be received at time t , where $t < T$. The arbitrage trading strategy that supports (13) is: sell the asset, buy a call, and buy risk-free bonds. The initial investment and terminal values are shown in Table 8. The first row in the table represents the short asset position. Today, we collect S , and, at the option's expiration, the short position must be covered at a cost of \tilde{S}_T . Shorting an asset, however, requires that we pay any dividends on the underlying asset. If we are short a stock and the stock pays a dividend, for example, we are obliged to pay the dividend out of our own pocket. Since the dividend is made during the option's life

Table 8 Arbitrage Portfolio Trades Supporting Lower Price Bound of European-Style Call Option Where the Underlying Asset Pays a Discrete Cash Dividend, $C - P < S - Xe^{-rT}$

Trades	Initial Investment	Cash Flow at t	Value on Day T	
			$S_T < X$	$S_T \geq X$
Buy asset	S	$-I_t$	$-\tilde{S}_T$	$-\tilde{S}_T$
Buy call option	$-c$		0	$\tilde{S}_T - X$
Buy risk-free bonds	$-Xe^{-rT} - I_t e^{-rt}$	I_t	X	X
Net portfolio value	$S - I_t e^{-rt} - Xe^{-rT} - c$	0	$X - \tilde{S}_T$	0

(i.e., $t < T$), the first row has a cash outflow of $-I_t$ paid on day t . The second row shows the long call position. On day t , the call is worth nothing if $S_t < X$ and $S_t - X$ if $S_t \geq X$. Finally, we buy some risk-free bonds. The amount necessary must be sufficient to cover the payment of the exercise price, X , on day T and the payment of the cash dividend, I_t , on day t , that is, $-Xe^{-rT} - I_t e^{-rt}$. Since the portfolio is certain to have a nonnegative net value on day t , the net portfolio value today must be less than or equal to 0, which implies $c \geq S - I_t e^{-rt} - Xe^{-rT}$.

Lower Price Bound of American-Style Call

A discrete cash dividend payment on the underlying asset affects the early exercise behavior of American-style call options differently than in the continuous carry rate case. In the case of an American-style call written on a stock, it may be optimal to exercise either just prior to the ex-dividend date (when the stock price falls by I_t) or at expiration. Early exercise between today and the ex-dividend instant and between the ex-dividend instant and expiration are not optimal because the call is worth more alive than dead.¹¹ The lower price bound of an American-style call is therefore the lower bound of a call expiring at the ex-dividend instant, $\max(0, S - Xe^{-rt})$, and the lower bound of the call expiring at expiration, $\max(0, S - I_t e^{-rt} - Xe^{-rT})$. Combining these two results,

$$c \geq \max(0, S - Xe^{-rt}, S - I_t e^{-rt} - Xe^{-rT}) \quad (14)$$

Early Exercise of American-Style Call Options

The last two terms on the right-hand side of (14) provide important guidance in deciding whether to exercise the American call option early, just prior to the ex-date. The second term in the parentheses is the present value of the early proceeds of the call. If this amount is less than the lower price bound of the call that ex-

pires normally, that is, if

$$S - Xe^{-rt} < S - I_t e^{-rt} - Xe^{-rT}$$

an American-style call will not be exercised early. To understand why, rewrite the expression as

$$I_t < X[1 - e^{-r(T-t)}] \quad (15)$$

The American-style call will not be exercised early if the cash flow (e.g., dividend or coupon payment) captured by exercising prior to the ex-date is less than the interest implicitly earned by deferring exercise from the ex-date until expiration.

The logic underlying the relation (15) also applies to the case where there are multiple known dividends paid during the call option's life. Take a stock option, for example. If the i th dividend is less than the present value of the interest income that can be implicitly earned as a result of deferring the payment of the exercise price until the next dividend payment, that is, if

$$I_i < X[1 - e^{-r(t_{i+1}-t_i)}] \quad (16)$$

exercising just prior to the i th dividend payment will not be optimal. This relation proves useful for simplifying the valuation of long-term stock options. The following example shows that dividend-induced early exercise on a long-term American-style call is most likely to occur just prior to the last dividend payment during the option's life.

Let's identify whether an American-style call option with an exercise price of 50 and one year remaining to expiration may be exercised early just prior to any of the dividend payments. Assume that the stock pays a quarterly dividend of 0.50 in 70 days, 161 days, 252 days, and 343 days. Assume the risk-free rate of interest is 5%.

Whether or not the call may be exercised early depends on the amount of the dividend payment in relation to the present value of the

interest income implicitly received by deferring the payment of the exercise price. For the first dividend, compute the values in expression (16) and find

$$0.50 < 50[1 - e^{-0.05(161/365-70/365)}] = 0.6194$$

Hence, the call will not optimally be exercised just prior to the first dividend payment. The same is true for the second and third dividend payments, as shown in the table below.

Quarter	Cash Dividend	Days to Dividend Payment	Years to Dividend Payment	PV of Interest Income
1	0.50	70	0.1918	0.6194
2	0.50	161	0.4411	0.6194
3	0.50	252	0.6904	0.6194
4	0.50	343	0.9397	0.1505

For the last dividend payment in 353 days, condition (13) is violated, that is,

$$0.50 > 50[1 - e^{-0.05(365-343)/365}] = 0.1505$$

This implies that exercise just prior to the last dividend payment during this option's life may be optimal.

Lower Price Bound of European-Style Put

The lower price bound for the European-style put option is

$$p \geq \max(0, Xe^{-rT} - S + I_t e^{-rt}) \quad (17)$$

Again, the asset price is reduced by the present value of the promised cash dividend on the asset. Unlike the call, however, the dividend payment increases the lower price bound of the European-style put. Because the put option is the right to sell the underlying asset at a fixed price, a discrete drop in the asset price such as one induced by the payment of a dividend on a stock serves to increase the value of the option. The arbitrage trades driving this relation are buy a put, buy a share of stock, and sell $I_t e^{-rt} + Xe^{-rT}$ risk-free bonds.

Lower Price Bound of American-Style Put

The lower price bound of the American-style put is

$$P \geq \max(0, Xe^{-rt} - S + I_t e^{-rt}, X - S) \quad (18)$$

The second term on the right-hand side is the present value of the exercise proceeds if the put is exercised just after the dividend payment. This lower price bound is supported by the arbitrage trades listed above for the European-style put. The third term on the right is the exercise proceeds if the put is exercised immediately. If $P < X - S$, a costless arbitrage profit can be earned by buying the put and the asset, and then exercising the put. The arbitrage profit is $X - S - P > 0$.

Early Exercise of American-Style Put Options

The early exercise behavior induced by the discrete cash dividend on the asset is different for the American-style put than it was for the call. If the third term exceeds the second in (18), the put will not be exercised early prior to the payment date. In that period the interest earned on the exercise proceeds of the option is less than the drop in the stock price from the payment of the dividend. For the third term to be larger than the second, that is,

$$Xe^{-rt} - S + I_t e^{-rt} > X - S$$

it must be the case that

$$I_t > X(e^{rt} - 1) \quad (19)$$

In other words, if the amount of the dividend amount exceeds the interest income that will accrue on the cash received if the put is exercised immediately, the put will not optimally be exercised early.

As in the case of the call, this argument can be generalized to handle the multiple dividends during the life of an American-style put. Again, consider a stock option. If the i th dividend is

greater than the interest that will accrue over the period,

$$I_t > X[e^{r(t-t_{i-1})} - 1] \tag{20}$$

the put will not be exercised before the dividend payment, as the illustration below shows.

We'll use an example to identify whether an American-style put option with an exercise price of 50 and one year remaining to expiration may be exercised early just after any of the dividend payments. Assume that the stock pays a quarterly dividend of 0.50 in 70 days, 161 days, 252 days, and 343 days. Assume the risk-free rate of interest is 5%.

Whether or not the put may be exercised early depends on the amount of the dividend payment in relation to the interest income that could be earned if the put were exercised immediately. For the first dividend, compute the values in expression (20), that is,

$$0.50 > 50[e^{0.05(70/365)} - 1] = 0.4818$$

This implies that the put will not be exercised before the first dividend payment in 70 days.

The computation for the second dividend is

$$0.50 > 50[e^{0.05(161/365 - 70/365)} - 1] = 0.6272$$

This implies that the put may be exercised in the period between the first and second dividends. The same is true between the second and third dividends, and the third and fourth dividends, as indicated below. Early exercise after

the fourth dividend is paid may also be optimal since no more dividends are paid during the option's life.

Quarter	Cash Dividend	Days to Dividend Payment	Years to Dividend Payment	Accrued Interest
1	0.50	70	0.1918	0.4818
2	0.50	161	0.4411	0.6272
3	0.50	252	0.6904	0.6272
4	0.50	343	0.9397	0.6272

Put-Call Parity for European-Style Options

Put-call parity for European-style options on assets with discrete noninterest cash flows is

$$c - p = S - I_t e^{-rt} - X e^{-rT} \tag{21}$$

To see this, assume the left-hand side of (21) is less than the right-hand side. If such is the case, an arbitrage profit can be made by selling the asset, selling the put, buying the call, and buying some risk-free bonds. The arbitrage is shown in Table 9. On day t , the net portfolio value is certain to be 0. The same is true on day T , when the cash dividend is made. Thus the value at time 0, $S - I_t e^{-rt} - X e^{-rT} + p - c$, represents the arbitrage profit and is positive if the left-hand side of (21) is less than the right-hand side. Since the market cannot be in equilibrium, arbitrage will continue until the net portfolio value goes to 0. When it does, the market is in equilibrium and (21) holds.

Table 9 Arbitrage Portfolio Trades Supporting European-Style Put-Call Parity Where the Underlying Asset Pays a Discrete Cash Dividend, $c - p = S - I_t e^{-rt} - X e^{-rT}$

Trades	Initial Investment	Cash Flow at t	Value on Day T	
			$S_T < X$	$S_T \geq X$
Sell asset	S	$-I_t$	$-\tilde{S}_T$	$-\tilde{S}_T$
Sell put option	p		$-(X - \tilde{S}_T)$	0
Buy call option	$-c$		0	$\tilde{S}_T - X$
Buy risk-free bonds	$-X e^{-rT} - I_t e^{-rt}$	I_t	X	X
Net portfolio value	$S - I_t e^{-rt} - X e^{-rT} + p - c$	0	0	0

Table 10 Arbitrage Trades Supporting American-Style Put-Call Parity Where the Underlying Asset Pays a Discrete Cash Dividend, $S - I_t e^{-rt} - X < C - P$

Trades	Initial Value	Ex-Day Value (t)	Put Exercised Early, Intermediate Value (τ)	Put Exercised Normally, Terminal Value (T)	
				$\tilde{S}_T \leq X$	$\tilde{S}_T < X$
Buy call	$-C$		\tilde{C}_τ	0	$\tilde{S}_T - X$
Sell put	P		$-(X - \tilde{S}_\tau)$	$-(X - \tilde{S}_T)$	0
Sell asset	S	$-I_t$	$-\tilde{S}_\tau$	$-\tilde{S}_T$	$-\tilde{S}_T$
Buy risk-free bonds	$-I_t e^{-rt} - X$	I_t	$Xe^{r\tau}$	Xe^{rT}	Xe^{rT}
Net portfolio value	$-C + P + S - I_t e^{-rt} - X$	0	$\tilde{C}_\tau + X(e^{r\tau} - 1)$	$X(e^{rT} - 1)$	$X(e^{rT} - 1)$

Put-Call Parity for American-Style Options

The put-call parity for American-style options on assets with discrete cash dividends is

$$S - I_t e^{-rt} - X \leq C - P \leq S - I_t e^{-rt} - X e^{-rT} \tag{22}$$

To understand why, we consider each inequality in (22) in turn. The inequality on the left can be derived by considering the values of a portfolio that consists of buying a call, selling a put, selling the stock, and buying $X + I_t e^{-rt}$ in risk-free bonds. Table 10 contains these trades as well as the net portfolio value.

In Table 10, we see that, if all positions stay open until expiration, the net portfolio value is positive independent of whether the terminal asset price is above or below the exercise price of the options. If the terminal asset price is above the exercise price, the call option is exercised, and the asset acquired at exercise price

X is used to deliver, in part, against the short asset position. If the terminal asset price is below the exercise price, the put is exercised. The asset received in the exercise of the put is used to cover the short stock position. In the event the put is exercised early at time τ , the investment in the risk-free bonds is more than sufficient to cover the payment of the exercise price, and the asset received upon delivery can be used to cover the short asset position. In addition, the call position remains open and has a nonnegative value. In other words, the combination of securities described in Table 10 will never have a negative future value. And, if the future value is certain to be nonnegative, the sum of the initial investment column must be nonpositive. In the absence of costless arbitrage opportunities, the left-hand inequality of (22) must hold.

The right inequality of (19) may be derived using the same portfolio used to prove European-style put-call parity. Table 11 contains the

Table 11 Arbitrage Trades Supporting American-Style Put-Call Parity Where the Underlying Asset Pays a Discrete Cash Dividend, $C - P < S - I_t e^{-rt} - X e^{-rT}$

Trades	Initial Value	Ex-Day Value (t)	Call Exercised Early, Intermediate Value (τ)	Call Exercised Normally, Terminal Value (T)	
				$\tilde{S}_T \leq X$	$\tilde{S}_T < X$
Sell call	C		$-(\tilde{S}_\tau - X)$	0	$-(\tilde{S}_T - X)$
Buy put	$-P$		\tilde{P}_τ	$X - \tilde{S}_\tau$	0
Buy stock	$-S$	I_t	\tilde{S}_τ	$-\tilde{S}_T$	\tilde{S}_T
Sell risk-free bonds	$-I_t e^{-rt} + X e^{-rT}$	$-I_t$	$-X e^{-r(T-\tau)}$	$-X$	$-X$
Net portfolio value	$C - P - S + I_t e^{-rt} + X$	0	$\tilde{P}_\tau + X(1 - e^{r(T-\tau)})$	0	0

Table 12 No-Arbitrage Price Relations For European- and American-Style Options on Assets Where the Underlying Asset Pays a Discrete Cash Dividend

Description	European-Style Options	American-Style Options
Lower price bound for call	$c \geq \max(0, S - I_t e^{-rt} - X e^{-rT})$	$c \geq \max[0, S - X e^{-rt}, S - I_t e^{-rt} - X]$
Lower price bound for put	$p \geq \max(0, X e^{-rT} - S + I_t e^{-rt})$	$P \geq \max(0, X - S, X e^{-rT} - S + I_t e^{-rt})$
Put-call parity relation	$c - p = S - I_t e^{-rt} - X e^{-rT}$	$S - I_t e^{-rt} - X \leq C - P$ $\leq S - I_t e^{-rt} - X e^{-rT}$

arbitrage portfolio trades. In this case, the net portfolio value at expiration is certain to be 0 should the option positions stay open until that time. In the event the American call option holder decides to exercise early, the portfolio holder delivers his share of stock, receives cash in the amount of the exercise price, and then uses the cash to retire his outstanding debt. After these actions are taken, the portfolio holder still has an open long put position and cash in the amount of $X[1 - e^{-r(T-\tau)}]$. Since the portfolio is certain to have nonnegative outcomes, the initial value must be negative or the right-hand inequality of (22) must hold.

Summary

This completes our derivations of arbitrage relations for European-style and American-style options on assets with discrete cash dividends. Options on dividend-paying stocks and on coupon-bearing bonds fall into this category. Before proceeding with a discussion of arbitrage relations for futures options, we summarize our results in Table 12.

NO-ARBITRAGE FUTURES OPTIONS RELATIONS

A *futures option* is like an asset option, except that if the option is exercised, a futures po-

sition is entered. Exercising a call option on a futures contract, for example, means that the holder will receive a long position in the futures at a price equal to the exercise price of the call.

Developing the lower bounds and put-call parity for European- and American-style futures options follows directly from the previous discussions. All we need to do is recall the prepaid version of the net cost of carry relations for futures: $Fe^{-rT} = Se^{-iT}$ where noninterest costs are modeled as a continuous rate, and $Fe^{-rT} = S - Ie^{-rt}$ where noninterest costs are modeled as a discrete flow. Substituting $Fe^{-rT} = Se^{-iT}$ into the no-arbitrage price relations summarized in Table 7 or $Fe^{-rT} = S - Ie^{-it}$ in the relations summarized in Table 12 produces the no-arbitrage price relations for futures options summarized in Table 13. The arbitrage portfolios supporting each of these relations are the same as those used to derive the relations for the asset throughout the entry.

NO-ARBITRAGE INTERMARKET RELATIONS

In many cases, both asset options and futures options trade concurrently. The Chicago Board Options Exchange, for example, lists options on the S&P 500 index, while the Chicago Mercantile Exchanges lists options on the S&P 500

Table 13 No-Arbitrage Price Relations For European- and American-Style Options on Futures Contracts

Description	European-Style Options	American-Style Options
Lower price bound for call	$c \geq \max[0, e^{-rT}(F - X)]$	$C \geq \max(0, F - X)$
Lower price bound for put	$p \geq \max[0, e^{-rT}(X - F)]$	$P \geq \max(0, X - F)$
Put-call parity relation	$c - p = e^{-rT}(F - X)$	$F e^{-rT} - X < C - P < F - X e^{-rT}$

futures (which, in turn, is written on the S&P 500 index). The prices of asset options are inextricably linked to the prices of futures options. Under the assumption that the futures and options expire simultaneously and that the exercise prices of the asset and futures options are the same, a number of no-arbitrage price relations may be derived. Next we present such relations for European-style and American-style options.

European-Style Options

The price of a European-style asset option is equal to the price of the corresponding futures option, that is,

$$c(S) = c(F) \quad (23a)$$

and

$$p(S) = p(F) \quad (23b)$$

The reason is that at expiration the payoffs of the asset option and the futures option are identical. Suppose, for the sake of illustration, that the price of a call on a futures exceeds the price of a call on an asset. In such a situation, costless arbitrage profits may be earned by buying the asset call and selling the futures call, as is shown in Table 14. The long asset option position pays nothing at expiration if the terminal asset price is less than the exercise price and pays $\tilde{S}_T - X$ if the terminal asset price exceeds the exercise price. At the same time, the short futures option position expires worthless at expiration if the terminal futures (asset) price is less than the exercise price and costs $-(\tilde{F}_T - X)$ if the terminal futures (asset) price exceeds the exercise price. But, since $\tilde{F}_T = \tilde{S}_T$, the net port-

folio value is certain to be zero. A portfolio that is certain to pay nothing on day T must cost nothing. Hence, in the absence of costless arbitrage opportunities, European-style asset options and European-style futures options have the same price.

American-Style Options

The relation between the price of an American-style asset option and the price of the corresponding futures option depends on whether the futures price is greater than the asset price or not. If $F > S$,

$$C(S) < C(F) \quad (24a)$$

and

$$P(S) > P(F) \quad (24b)$$

To see this, consider the American-style call options. Since both the call on the futures and the call on the asset may be exercised early, we can compare the early exercise proceeds to establish which has greater value. The call on the asset has immediate early exercise proceeds of $S - X$ and the call on the futures has early exercise proceeds of $F - X > S - X$. Thus as long as there is some chance of early exercise, the call on the futures is worth more than the call on the asset and the put on the asset is worth more than the put on the futures.

For cases where futures price is less than the asset price, the opposite results hold, that is,

$$C(S) > C(F) \quad (25a)$$

and

$$P(S) < P(F) \quad (25b)$$

Table 14 Arbitrage Portfolio Trades Demonstrating the Equivalence of Prices of European-Style Call Options on an Asset and a Futures, $c(F) = c(S)$

Trades	Initial Investment	Value on Day T	
		$S_T < X$	$S_T \geq X$
Buy call option on asset	$-c(S)$	0	$\tilde{S}_T - X$
Sell call option on futures	$c(F)$	0	$-(\tilde{F}_T - X) = -(\tilde{S}_T - X)$
Net portfolio value	$c(F) - c(S)$	0	0

Table 15 No-Arbitrage Relations Between the Prices of Asset Options and Futures Options

Description	European-Style Options	American-Style Options
Call	$c(S) = c(F)$	$C(S) < C(F)$ if $F > S$ $C(S) > C(F)$ if $F < S$
Put	$p(S) = p(F)$	$P(S) > P(F)$ if $F > S$ $P(S) < P(F)$ if $F < S$

The previous arbitrage argument is merely reversed. Table 15 summarizes the results.

KEY POINTS

- Under the assumption that no costless arbitrage (i.e., free money) opportunities are available in the marketplace, no-arbitrage price relations for European- and American-style options can be developed.
- The net cost of carry of the underlying asset plays an important role. Consequently, it is necessary to model interest cost as a constant continuous rate and the noninterest cost as a continuous rate or as a discrete flow, depending on the nature of the underlying asset.

For options on stock indexes, currencies, and some commodities, the continuous rate assumption is most appropriate. For options on stocks, bonds, and other commodities, the discrete flow assumption is preferred.

- With the assumptions regarding net cost of carry, lower price bounds, put-call parity price relations, and intermarket price relations can be derived for both European-style and American-style options on an asset and on a forward/futures.
- For American-style options, there is always the prospect of early exercise. Under certain circumstances regarding the cost of carry, the holder of an American-style call option would never (rationally) exercise early. In the case of an American-style put, there is always some prospect of early exercise, so the American-style put is always worth more than the European-style put.

- Perhaps most important is the no-arbitrage price relation between the price of a put and the price of a call. This relation, called the put-call parity relation, arises from simultaneous trades in the call, the put, and the asset.
- With respect to intermarket price relations, the prices of asset options are inextricably linked to the prices of futures options. Under the assumption that the futures and options expire simultaneously and that the exercise prices of the asset and futures options are the same, a number of no-arbitrage price relations may be derived.

NOTES

1. *European-style* options can be exercised only on expiration day, while *American-style* options can be exercised at any time up to and including the expiration day. Both types of options are traded on exchanges and in OTC markets.
2. Under the continuous cost of carry rate assumption, the continuously paid income received from holding the asset is immediately reinvested in more units of the asset, so that e^{-iT} units on day 0 grows to one unit on day T . For a short asset position, the reverse applies in the sense that our liability (in terms of number of units owed) grows at rate i .
3. It is also worthwhile to note that the lower price bound of the call can be re-expressed relative to the forward/futures prices. The net cost of carry relation for forwards/futures prices is $fe^{-rT} = Se^{-iT}$. Substituting the cost of carry relation into (4), $c \geq \max(0, fe^{-rt} - Xe^{-rT})$.
4. The distinction between value and price is subtle, but important. A price is what we observe for the security in the marketplace; a value is what we believe a security is worth. If the value exceeds the price, the security is underpriced, and, if the value is less than the price, the security is overpriced.
5. Note that we are not making any judgment on whether the call price is too high or too

low *per se*. We are saying only that the call is incorrectly priced (in this case it is priced too low) *relative* to the price of the underlying asset. To execute the arbitrage, we must trade both the call and the underlying asset, so that we make money when their prices come back into line relative to each other. In this example, the prices come back into line with each other for certain at the option's expiration.

6. To exit a long position in an American-style call option, we have three alternatives. First, we can hold it to expiration, at which time we will (a) let it expire worthless if it is out of the money or (b) exercise it if it is in the money. Second, we can exercise it immediately, receiving the difference between the current asset price and the exercise price. Third, we can sell it in the marketplace. There is, after all, an active secondary market for standard calls and puts.
7. This point was first demonstrated by Merton (1973) for call options on nondividend-paying stocks. He refers to such options as being worth more "alive" than "dead."
8. In the expression on the right-hand side of (10), the third term is greater than the second term over some range for S , independent of the level of i .
9. The term, "put-call parity," was first coined by Stoll (1969) in the first academic study to develop and test the relation.
10. If we buy a put option, we pay the premium today for the right to sell the underlying asset at the exercise price. If we sell the put, we collect the premium today but have the obligation to deliver the asset and receive the exercise price if the put option buyer chooses to exercise.
11. By not exercising in the period prior to ex-dividend, the call option holder enjoys the benefits of implicitly earning interest on the dividend and the exercise price of the call. By not exercising after the ex-dividend date but before expiration, the call option holder enjoys the benefit of implicitly earning interest on the exercise price of the call.

REFERENCES

- Merton, Robert C. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4: 141–183.
- Smith Jr., Clifford W. (1976). Option pricing: A review. *Journal of Financial Economics* 3 (January/March): 3–51.
- Stoll, Hans R. (1969). The relationship between put and call prices. *Journal of Finance* 24: 802–824.
- Stoll, Hans R., and Robert E. Whaley. (1986). New option instruments: Arbitrageable linkages and valuation. *Advances in Futures and Options Research* 1(A): 25–62.

Introduction to Contingent Claims Analysis

EDWIN H. NEAVE, PhD

Professor Emeritus, School of Business, Queen's University, Kingston, Ontario

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Contingent claims are a tool for valuing securities and for analyzing the effects of risky financial decisions. Contingent claims analysis can be used to value any kind of financial instrument, including such apparently exotic instruments as put and call options and convertible securities. Contingent claim analysis defines risky outcomes relative to states of the world, and uses claims to represent and value state outcomes. Thus given a definition of risky states, all financial instruments and arrangements can be represented as combinations of contingent claims on those states. Theoretically complete markets assume claims can be traded on every state of the world, but in practice markets are not likely to be complete at any point in time. Since in practice market incompleteness will inhibit certain risk management strategies, in so doing it also provides incentives to create new instruments that can be used to manage and to value claims on additional states of the world.

Contingent claims analysis is used in financial modeling to value any financial instrument, including such apparently exotic instruments as put and call options and convertible securities. In this entry, we discuss this important tool. We begin by explaining the notion of states of the world, a way of classifying risky outcomes whose value can then be represented using contingent claims. After providing examples of valuation using contingent claims, we introduce the concept of incomplete markets and consider its importance for modeling real-world financial arrangements. We then examine some financial instruments and arrangements that can be used to trade or to manage risks.

STATES OF THE WORLD

The idea of *states of the world* is useful for thinking about convenient ways to model risky payoffs. In a two-time-point model, states of the world are defined as those future events that matter to the decision problem being considered. These states of the world are defined by the decision maker to be mutually exclusive and collectively exhaustive. Using an example given by Savage (1951), if one is about to break a ninth egg into a bowl already containing eight other eggs, the relevant states of the world could be whether the ninth egg is rotten and would hence spoil the others. (Here we presume the

rotteness of an egg is not discernible until the egg has been broken and fallen into the bowl.)

In a second example more closely related to finance, an investor might be concerned with the future price of a share of stock, and this price might in turn depend on economic conditions. Suppose the investor defines (1) “states” to represent economic conditions, and (2) “future prices” to be the following list of possible share prices that may obtain at the time a given state is actually realized:

State	Future Price
1	\$10
2	\$8
3	\$6

For example, state 1 might mean that the industry in which the firm operates faces buoyant market conditions; state 2, conditions that are neither good nor bad; and state 3, conditions that are depressed. In each state, the effect is registered on the stock price.

We shall usually associate probabilities with the states; for example, p_i might represent the probability that state i will actually occur; that is, $i = 1, 2, 3$. Because the states are mutually exclusive, only one can actually occur; because they are collectively exhaustive, one of the three must occur. Hence $\sum_i p_i = 1$.

Note that although in this chapter we make less use of multiperiod models using contingent claims, we can also define states at different points in time, for example, the states of the world at different times.

CONTINGENT CLAIMS AND THEIR VALUE

A *unit contingent claim* is a security that will pay an amount of \$1 if a certain state of the world is actually realized, but nothing otherwise. A claim that pays \$1 if state i is realized is frequently called a *unit claim* on state i . A unit contingent claim is also referred to as a *primary security* or *Arrow-Debreu security* (so named after

the economists who introduced them—Arrow [1964] and Debreu [1959]).

Accordingly, the future stock price described earlier may be regarded as equivalent to a package containing all of the following:

Ten unit claims on state 1
Eight unit claims on state 2
Six unit claims on state 3

The idea of a contingent claim is thus useful for expressing, in terms of fundamental units, exactly what a given security’s payoff may be in different possible states of the world.

It may take a little imagination to come up with real-world examples of claims, and those real-world examples are not numerous. (A ticket to win on a horse race is an example of a claim; a fire insurance policy is another. One example of a unit claim is an option that pays off \$1 if the value of some underlying asset exceeds a fixed dollar value.) But packages of unit claims represent perfect substitutes for the more ordinary types of securities such as stocks or bonds, and we shall frequently find it useful to employ claims to help understand price relations between securities. For example, if we assume a perfectly competitive financial market along with a description of future events in terms of states of the world, certain price relationships between securities and contingent claims must obtain. This means in turn that certain predictable relationships between securities prices must also obtain.

To see these relationships, suppose that we can describe the world using two states and that two stocks are available, stock A and stock B. We assume the stocks’ future prices have the following distributions:

	Stock A	Stock B
1	\$10	\$7
2	\$8	\$9

Let $A(0)$ = denote the time 0 price of stock A and $B(0)$ = the time 0 price of stock B, and suppose these prices admit no arbitrage opportunities. Now if we let C_1 and C_2 represent the

time 0 prices of unit claims on states 1 and 2, we can use the foregoing information about stock prices and payoffs to find the time 0 prices C_1 and C_2 . Purchasing stock A for \$6 is equivalent to buying a package of 10 unit claims on state 1 and 8 unit claims on state 2, while buying stock B for \$5 is equivalent to buying a package of 7 unit claims on state 1 and 9 unit claims on state 2. Since the unit claims comprising the two stocks are perfect substitutes, they must sell for the same prices in a perfect market. Hence we can write

$$10C_1 + 8C_2 = \$6$$

$$7C_1 + 9C_2 = \$5$$

which can be solved to obtain

$$C_1 = \$\frac{7}{17}, \quad C_2 = \$\frac{4}{17}$$

We can use the same reasoning to find the risk-free rate of return that must obtain in this market. Since a risk-free instrument is one that offers the same payoff irrespective of which state of the world obtains, we wish to find a combination of the two stocks that gives the same time 1 payoff, here denoted k , in either state of the world. That is, the following equation must be solved for α :

$$\alpha \begin{pmatrix} 10 \\ 8 \end{pmatrix} + (1 - \alpha) \begin{pmatrix} 7 \\ 9 \end{pmatrix} = \begin{pmatrix} k \\ k \end{pmatrix}$$

We can write the payoff k as equal to either of the following payoffs:

$$10\alpha + 7(1 - \alpha) = 8\alpha + 9(1 - \alpha)$$

which implies that

$$2\alpha = 2(1 - \alpha)$$

so that $\alpha = \frac{1}{2}$. The riskless payoff is then $\frac{1}{2}(10) + \frac{1}{2}(7) = \8.50 , and this can be obtained for a price equal to $\frac{1}{2}(6) + \frac{1}{2}(5) = \5.50 , since a portfolio composed of equal proportions of the two stocks creates the riskless investment. Accordingly, the risk-free rate of return is

$$\frac{\$8.50 - \$5.50}{\$5.50} = \frac{6}{11} = 54.55\%$$

Of course, this is not necessarily a realistic number for a risk-free rate of interest. (Whether it is realistic or not depends on the length of the time period under consideration, a matter we have left unspecified.) However, our purpose here is to develop illustrative calculations to display relations between contingent claims, and for this purpose particular sizes of numbers are not really important.

Another way of making a riskless investment is to buy one of each available unit claim, that is, one claim on state 1 and one claim on state 2. Such a portfolio gives a certain payoff of \$1 for an investment cost of

$$\$ \frac{4}{17} + \$ \frac{7}{17} = \$ \frac{11}{17}$$

The rate of return on this investment is then

$$\frac{\$1 - \$ \frac{11}{17}}{\$ \frac{11}{17}} = \frac{17 - 11}{11} = \frac{6}{11} = 54.55\%$$

just as before.

INVESTOR'S UTILITY MAXIMIZATION IN CONTINGENT CLAIMS MARKETS

In this section, we describe how an investor may solve the utility maximization problem when facing risk in a market for contingent claims. For our illustration, we shall continue with stocks A and B from the previous section. Further, we shall assume the investor's initial wealth to be \$600. This scenario is summarized in Table 1. We let w_1 represent wealth if state 1 occurs and correspondingly for w_2 , and we may plot these data in (w_1, w_2) space, as shown in Figure 1. Note that the previously determined riskless position of dividing the purchases to obtain an equal number of each security (54.5 of each) is also shown and generates a riskless terminal wealth position of $w_1 = w_2 = \$926.50$.

Table 1 Summary of Terminal Wealth in Two States

	No. of Shares Purchased	Terminal Wealth	
		State 1	State 2
Purchases A only	100	\$1,000	\$ 800
Purchases B only	120	\$ 840	\$1,080

We can also use another way to calculate the value of the claims' combinations at time 1. We can write the equation of the straight line in Figure 1 as

$$w_2 = a - bw_1$$

so that for the time 1 price of stock A we have

$$\$800 = a - \$1,000b$$

while for the time 1 price of stock B we have

$$\$1,080 = a = \$840b$$

Solving these two simultaneous equations, we find $b = 0.175$ and $a = \$2,550$. Thus, when $w_1 = 0$, $w_2 = \$2,550$, while when $w_2 = 0$, $w_1 = \$1,457$, which are the two intercepts of the line on their respective axes in Figure 1.

Now if $w_2 = 0$, we have the case of a claim (primary security) on state 1. (The security pays \$1,457 in state 1 and nothing otherwise.) The price of this claim can be calculated by dividing initial wealth by the maximum wealth obtained if state 1 occurs, or $\$600/\$1,457 = 0.41$ ($= \frac{7}{17}$). Similarly, the price of primary security 2 is $\$600/\$2,550 = 0.24$ ($= \frac{4}{17}$), and our earlier results are confirmed.

Note that in Figure 1 the investor's time 1 position is some point on the line from A to B. How could the investor obtain a terminal wealth position lying beyond these points? The investor could engage in short sales, that is, selling shares not currently owned, for delivery when the unknown future state of the world is revealed. In this transaction the investor obtains cash from the time 0 sale of one security and uses it to buy the other. In so doing, the investor promises later to buy the security sold short at whatever price will be prevailing and deliver it. Note that there is a potential for large gains or losses in such transactions. Here the initial wealth will be used as a constraint and we shall require that at worst the investor will

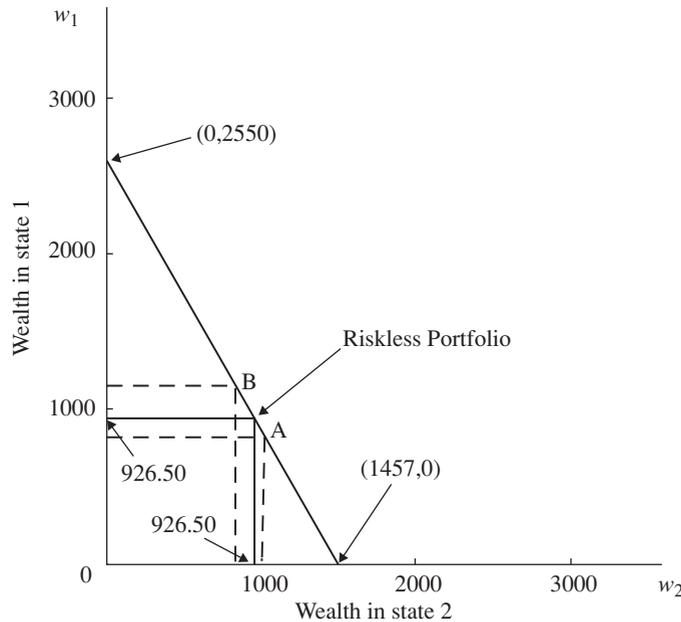


Figure 1 Market Opportunity Line, Showing Implied Prices of Unit Claims. Note: w_1 represents wealth if state 1 occurs and correspondingly for w_2 .

have zero terminal wealth if he or she guesses incorrectly. That is, no net borrowing is permitted at the end of the period so that the investor cannot go beyond the intercepts in Figure 1.

To illustrate, consider point $w_1 = \$1,457$, $w_2 = 0$. Let n_A be the number of shares of stock A and n_B the number of shares of stock B purchased. If state 1 occurs, the terminal wealth will be

$$10n_A + 7n_B = \$1,457$$

while if state 2 occurs, we must have

$$8n_A + 9n_B = 0$$

Solving these equations simultaneously, we find $n_B = 343$. If the investor sells short 343 shares of stock B at the current price of \$5, he or she will receive \$1,715. Combining this with the initial wealth of \$600 gives \$2,315, so this investor may buy $\$2,315/\$6 = 386 n_A$ at \$6 per share. If state 1 eventuates, the investor will receive \$3,860 (\$10 for the 386 shares) but now must pay \$2,401 (\$7 for 343 shares) for stock B shares to cover the short position. The net terminal wealth is $\$3,860 - \$2,401 = \$1,459$ (difference due to rounding), as required. In state 2, the terminal wealth will be equal to \$3,088 (386 shares times \$8 per share) reduced by the cost to repurchase stock B to cover the short position, 343 shares at \$9 per share or \$3,087. Therefore, the net terminal wealth is equal to zero (the calculations show it is \$1 but that is due to rounding).

Note that none of the points we have considered will necessarily be a utility-maximizing point. To determine this point, it is necessary to know the investor's utility function in (w_1, w_2) space. The optimal portfolio for the investor satisfies the tangency condition that the slope of the wealth constraint (the ratio of the prices of the unit claims) equals the slope of the indifference curve (marginal rate of substitution of state 1 consumption for state 2 consumption).

The point of the foregoing demonstration is to show first that every security can be viewed as a bundle of unit claims and thus represents a combination of positions regarding future

states of the world. Moreover, in these circumstances an investor can attain any point along the market opportunity line. If, on the other hand, there are fewer securities than the number of distinct states, the individual's optimal consumption choices may not be attainable. The significance of this will be explored in the next section.

Although we do not discuss it here, the real power of the contingent claim analysis is in providing the basis for valuing complex financial instruments and financial arrangements.

INCOMPLETE MARKETS FOR CONTINGENT CLAIMS

A market is said to be a *complete market* when economic agents can structure any set of future state payoffs by investing in a portfolio of unit contingent claims (i.e., primary securities). A financial market is said to be *incomplete* if the number of (linearly) independent securities traded in it is smaller than the number of distinct states of the world. Clearly, market incompleteness depends on how states of the world are defined. However, since the number of states of the world needed to describe a typical financial market is likely to be large, the possibility that real-world financial markets will be incomplete is a very real one.

The importance of market incompleteness is best introduced by an example. Let us consider an economy with three possible states of the world and suppose only two securities (taking the form of unit claims for ease of exposition) are traded in it. We describe the securities in terms of their time 1 market value, for each state of the world, as in Table 2. It is apparent from the table that weighted averages of the two unit claims can be used to create packages with time 1 distributions of values ranging between zero and unity, the actual outcome depending on whether state 1 or state 2 obtains. However, an investor cannot create an income claim of other than zero in state 3 by using just the existing two unit claims. Moreover, no

Table 2 Market Values of Two Securities at Time 1

Security	States of the World		
	1	2	3
1	1	0	0
2	0	1	0

investor can arrange a risk-free investment in this example, because it is not possible to guarantee the same return in every state of the world by using just the available securities.

The situation is quite different if a third unit claim worth \$1 in state 3 and zero in the other states becomes available. Now the number of claims equals the number of distinct states, and a risk-free investment can now be arranged.

We are now ready to discuss some practical implications of market incompleteness. It is obvious from the foregoing example that investor choice is restricted in incomplete markets. Moreover if investor choices are restricted, the investors will never be better off, and are likely to be worse off, than would be the case if markets were complete (i.e., if the restrictions were removed). In such situations, it is to be expected that if ways of completing the market can be found, those possibilities are likely to be utilized. That is, in the context of incomplete financial markets the appearance of new instruments might be regarded as attempts to provide investors with financial opportunities not otherwise available. The appearance of *derivatives* (options, futures, and swaps) might be examples of such attempts. Mossin (1977) argues that the preference existing firms show for organizing new activities as separate corporations may be another indication of attempts to deal with market incompleteness.

FINANCIAL INSTRUMENTS AS CONTINGENT CLAIMS

Most financial instruments can be bought or sold, but not all of them are actively traded in

financial markets. For example, a common form of contingent claim (and one that is close in concept to a unit claim) is a lottery ticket. In its simplest form this claim results in its holder winning either a positive prize or zero. Accordingly, this lottery ticket represents a claim that can be valued using two states of the world. Obviously, if a lottery has several different prizes, several states of the world may need to be defined in order to describe it completely. But lottery tickets, once issued, are rarely traded again. The same is true of such other contingent claims as the tickets obtained when betting on horse races or similar contests.

An insurance policy is a contingent claim that comes closer to our usual notions of a financial instrument, but again it is rarely traded in the financial markets. On the other hand, put or call options, representing contingent claims for selling or buying securities or financial indexes at prespecified prices, trade actively on such organized exchanges. Rights and warrants are other examples of contingent claims in that they permit, but do not require, the holder to buy securities on prespecified terms.

There are also securities that have *embedded derivatives* in them, derivatives that are not traded separately from the instrument itself. For example, a *callable bond* is a bond that grants the issuer the right to redeem the bond at some time in the future and at a specified price. That is, a callable bond can be viewed as a straight bond with an embedded call option granted to the issuer. A *puttable bond* is a bond that grants the investor the right to sell (i.e., put) the bond to the issuer in the future at a specified price. Hence, the bond structure can be viewed as a straight bond with an embedded put option. *Convertible securities*, which include convertible bonds or convertible preferred stocks, represent contingent claims in that they typically allow the owner to exchange the original issue for other securities, usually common stock, and they are callable. Some convertible securities even include an embedded put option.

KEY POINTS

- Contingent claims analysis and contingent strategies are tools for dealing with risk in financial decision making.
- Contingent claims analysis uses the notion of states of the world in assessing future risky payoffs.
- A unit contingent claim (also known as a primary security or Arrow-Debreu security) is a security that has a payoff of \$1 if a certain state of the world is actually realized, but nothing in all other states.
- A contingent claim that pays off \$1 if state i is realized is also referred to as a unit claim on state i .
- Although few unit contingent claims exist in reality, claims represent a useful tool to employ in valuing securities and in understanding relations among them.
- An investor may solve the utility maximization problem when facing risk in a market for contingent claims.
- Using contingent claims analysis, an investor can obtain a terminal wealth position beyond what can be obtained by simply buying securities with initial wealth by engaging in short sales (i.e., selling shares not currently owned, for delivery when the unknown future state of the world is revealed). The outcomes in this case are more risky than they would be in the absence of short selling.
- Every security can be viewed as representing a bundle of unit claims and thereby further represents a combination of positions (long and short) regarding future states of the world.
- If the number of (linearly) independent securities traded is smaller than the number of distinct states of the world, the financial market is said to be incomplete.
- Because the number of states of the world necessary to describe a well-functioning financial market is likely to be large, the possibility that real-world financial markets will be incomplete is a very real one.
- Although most financial instruments representing contingent claims can be bought or sold, there are financial instruments or financial arrangements that are not actively traded in financial markets.

REFERENCES

- Arrow, K. J. (1964). The role of securities in the optimal allocation of risk-bearing. *Review of Economic Studies* 31: 91–96.
- Debreu, G. (1959). *The Theory of Value*. New York: Wiley.
- Mossin, J. (1977). *The Economic Efficiency of Financial Markets*. Lexington, MA: Heath.
- Savage, L. J. (1951). *The Foundations of Statistics*. New York: Wiley.

Black-Scholes Option Pricing Model

SVETLOZAR T. RACHEV, PhD, DrSci

Frey Family Foundation Chair Professor, Department of Applied Mathematics and Statistics, Stony Brook University; and Chief Scientist, FinAnalytica

CHRISTIAN MENN, Dr Rer Pol

Managing Partner, RIVACON

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: The most popular continuous-time model for option valuation is based on the Black-Scholes theory. Although certain drawbacks and pitfalls of the Black-Scholes option pricing model have been understood shortly after its publication in the early 1970s, it is still by far the most popular model for option valuation. The Black-Scholes model is based on the assumption that the underlying follows a stochastic process called geometric Brownian motion. Besides pricing, every option pricing model can be used to calculate sensitivity measures describing the influence of a change in the underlying risk factors on the option price. These risk measures are called the “Greeks” and their use will be explained and described.

In this entry, we look at the most popular model for pricing options, the Black-Scholes model, and look at the assumptions and their importance. We also explain the various Greeks that provide information about the sensitivity of the option price to changes in the factors that the model tells us affects the value of an option.

MOTIVATION

Let us assume that the price of a certain stock in June of Year 0 ($t = 0$) is given to be $S_0 = \$100$. We want to value an option with strike price

$X = \$110$ maturing in June of Year 1 ($t = T$). As additional information we are given the continuously compounded one-year risk-free interest rate $r = 5\%$. Figure 1 visualizes potential paths of the stock between $t = 0$ and $t = T$. How can we define a reasonable model for the stock price evolution?

It is clear that the daily changes or the change between the current and the next quotes cannot be predicted and can consequently be seen as realizations of random variables. However, we know that if we are investing in stocks then we can expect a rate of return in the long run which is higher than the risk-free rate. Let us denote

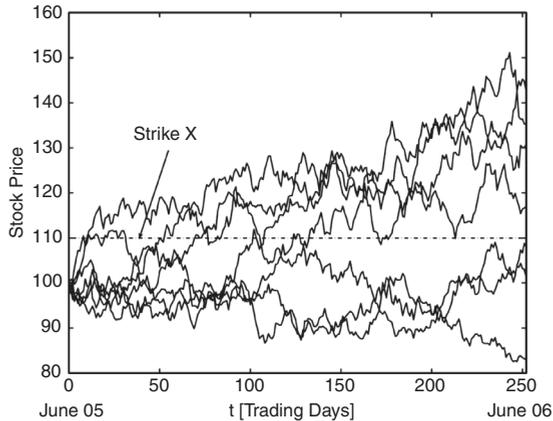


Figure 1 Possible Paths of the Stock Price Evolution over One Year with $S_0 = \$100$ and $X = \$110$

that unknown expected rate of return as μ . Here and in the rest of this entry, we assume that the stock pays no dividend.

Furthermore, we know that stock returns exhibit random fluctuations called volatility. Let σ denote the unknown yearly rate of volatility. Here and below we have implicitly assumed that the expected return and the volatility of the stock are time independent. This assumption might be violated in practice. Formalizing our ideas about the stock price we come up with the following equation for the return of the stock in a small time interval of length Δt :

$$\underbrace{\frac{S_{t+\Delta t} - S_t}{S_t}}_{\text{Return in period } [t, t+\Delta t]} = \mu \cdot \Delta t + \underbrace{\sigma \cdot \varepsilon_t^{\Delta t}}_{\text{“Stochastic noise”}}$$

The stochastic noise $\sigma \cdot \varepsilon_t^{\Delta t}$ should have the following properties:

- No systematic influence: $E(\varepsilon_t^{\Delta t}) = 0$.
- No dependence between the noise of different dates: The random variables ε_t and ε_s are independent for $s \neq t$.
- The variance of the noise is proportional to the length of the time interval Δt .

One possible model for the noise process is provided by a stochastic process called Brownian motion. A detailed discussion of Brownian motion is beyond the scope of this entry, but

we provide a brief discussion of its defining properties.

Brownian motion is a stochastic process $(W_t)_{t \geq 0}$ in continuous time that has the following four properties:

1. $W_0 = 0$, that is, Brownian motion starts at zero.
2. $(W_t)_{t \geq 0}$ is a process with homogeneous and independent increments.
3. Any increment $W_{t+h} - W_t$ is normally distributed with mean zero and variance h .
4. The paths of $(W_t)_{t \geq 0}$ are continuous.

The increments of Brownian motion are an appropriate candidate for the stochastic noise in our stock price model and we define:

$$\varepsilon_t^{\Delta t} = W_{t+\Delta t} - W_t$$

From its defining properties, we know that the increments of the Brownian motion are independent and that the variance of the increments is proportional to the length of the considered time interval. Additionally, the expectation of the increments is zero.

With this definition, it is possible to write the equation for the return process in the following form:

$$\frac{S_{t+\Delta t} - S_t}{S_t} = \mu \Delta t + \sigma (W_{t+\Delta t} - W_t)$$

If we decrease the length Δt of the time interval over which the increment is considered constant, then we can switch to a “differential type” notation:

$$\frac{dS_t}{S_t} = \mu \cdot dt + \sigma \cdot dW_t, \quad t \geq 0$$

The process defined in the above equation is called *geometric Brownian motion*. In explicit notation the geometric Brownian motion possesses the following form

$$S_t = S_0 e^{\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma W_t}$$

and S_t is lognormally distributed. This process is used in the Black-Scholes model to describe the stock price dynamic. Additionally,

the model assumes the existence of a risk-free asset—called money market account or bond—with the following dynamic:

$$\frac{dB_t}{B_t} = r \cdot dt, \quad t \geq 0 \Leftrightarrow B_t = B_0 e^{rt} \cdot t \geq 0 \quad (1)$$

BLACK-SCHOLES FORMULA

Black and Scholes (1973) have shown that it is possible—under some assumptions discussed in this section—to duplicate the payoff of a European call option with a continuously rebalanced portfolio consisting of the two assets S and B . This means that the price of the call option equals the initial costs for starting the hedging strategy.

The *Black-Scholes option pricing model* computes the fair (or theoretical) price of a European call option on a nondividend-paying stock with the following formula:

$$C = S\Phi(d_1) - Xe^{-rT}\Phi(d_2) \quad (2)$$

where

$$d_1 = \frac{\ln(S/X) + (r + 0.5\sigma^2T)}{\sigma\sqrt{T}} \quad (3)$$

$$d_2 = d_1 - \sigma\sqrt{T} \quad (4)$$

where

- $\ln(\cdot)$ = natural logarithm
- C = call option price
- S = current stock price
- X = strike price
- r = short-term risk-free interest rate in percent per annum
- $e = 2.718$ (natural antilog of 1)
- T = time remaining to the expiration date (measured as a fraction of a year)
- σ = expected return volatility for the stock (standard deviation of the stock's return in percent per annum)
- $\Phi(\cdot)$ = the cumulative distribution function of a standard normal distribution

The option price derived from the Black-Scholes option pricing model is “fair” in the sense that if any other price existed in a market

where all the assumptions of the Black-Scholes model are fulfilled, it would be possible to earn riskless arbitrage profits by taking an offsetting position in the underlying stock. That is, if the price of the call option in the market is higher than that derived from the Black-Scholes option pricing model, an investor could sell the call option and buy a certain number of shares in the underlying stock. If the reverse is true, that is, the market price of the call option is less than the “fair” price derived from the model, the investor could buy the call option and sell short a certain number of shares in the underlying stock. This process of hedging by taking a position in the underlying stock allows the investor to lock in the riskless arbitrage profit. The number of shares necessary to hedge the position changes as the factors that affect the option price change, so the hedged position must be changed constantly.

COMPUTING A CALL OPTION PRICE

To illustrate the Black-Scholes option pricing formula, assume the following values:

- Strike price = \$45
- Time remaining to expiration = 183 days
- Current stock price = \$47
- Expected return volatility = Standard deviation = 25% per annum
- Risk-free rate = 10% per annum

In terms of the values in the formula:

- $S = 47$
- $X = 45$
- $T = 0.5$ (183 days/365, rounded)
- $\sigma = 0.25$
- $r = 0.10$

Substituting these values into equations (3) and (4):

$$d_1 = \frac{\ln(47/45) + (0.10 + 0.5[0.25]^2)0.5}{0.25\sqrt{0.5}} = 0.6172$$

$$d_2 = 0.6172 - 0.25\sqrt{0.5} = 0.440443$$

From a normal distribution table,

$$\Phi(0.6172) = 0.7315 \quad \text{and} \quad \Phi(0.4404) = 0.6702$$

Then

$$C = \$47(0.7315) - \$45(e^{-(0.10)(0.5)})(0.6702) = \$5.69$$

Table 1 shows the option value as calculated from the Black-Scholes option pricing model for different assumptions concerning (1) the standard deviation for the stock's return (that is, expected return volatility); (2) the risk-free rate; and (3) the time remaining to expiration. Notice that the option price varies directly with three

Table 1 Comparison of Black-Scholes Call Option Price Varying One Factor at a Time

Base Case	
Call option:	
Strike price = \$45	
Time remaining to expiration = 183 days	
Current stock price = \$47	
Expected return volatility = Standard deviation of a stock's return = 25%	
Risk-free rate = 10%	
Holding All Factors Constant except Expected Return Volatility	
Expected Price Volatility	Call Option Price [\$]
15% per annum	4.69
20	5.17
25 (base case)	5.59
30	6.26
35	6.84
40	7.42
Holding All Factors Constant Except the Risk-Free Rate	
Risk-Free Interest Rate, % per annum	Call Option Price [\$]
7%	5.27
8	5.41
9	5.50
10 (base case)	5.69
11	5.84
12	5.99
13	6.13
Holding All Factors Constant except Time Remaining to Expiration	
Time Remaining to Expiration	Call Option Price [\$]
30 days	2.85
60	3.52
91	4.15
183 (base case)	5.69
273	6.99

variables: expected return volatility, the risk-free rate, and time remaining to expiration. That is: (1) the lower (higher) the expected volatility, the lower (higher) the option price; (2) the lower (higher) the risk-free rate, the lower (higher) the option price; and (3) the shorter (longer) the time remaining to expiration, the lower (higher) the option price.

SENSITIVITY OF OPTION PRICE TO A CHANGE IN FACTORS: THE GREEKS

In employing options in investment strategies, an asset manager or trader would like to know how sensitive the price of an option is to a change in any one of the factors that affect its price. Sensitivity measures for assessing the impact of a change in factors on the price of an option are referred to as the *Greeks*. In this section, we will explain these measures for the factors in the Black-Scholes model. Specifically, we discuss measures of the sensitivity of a call option's price to changes in the price of the underlying stock, the time to expiration, expected volatility, and interest rate. These factors can be divided into "market factors" and "model factors." The underlying price and the time to expiration are market factors, whereas the volatility and the interest rate are model factors. The special aspect about the latter variables is that they are assumed to be constant within the model. By admitting that these parameters can change as well, we are admitting that the model is inconsistent with reality. Table 2 gives an overview and lists the sensitivities of the option price with respect to all parameters of the Black-Scholes model.

Price of a Call Option Price and Price of the Underlying: Delta and Gamma

In developing an option-pricing model, we have seen the importance of understanding the relationship between the option price and the

Table 2 Sensitivities of the Option Price with Respect to Each Parameter of the Black-Scholes Model

Parameter	Corresponding Greek	Analytic Expression
Stock price S	Delta	$\Delta = \frac{\partial C}{\partial S} = \Phi(d_1)$
Stock price S (convexity adjustment)	Gamma	$\Gamma = \frac{\partial^2 C}{\partial S^2} = \frac{\varphi(d_1)}{S\sigma\sqrt{T}}$
Volatility σ	Vega	$v = \frac{\partial C}{\partial \sigma} = S \cdot \varphi(d_1) \cdot T$
Time	Theta	$\Theta = -\frac{\partial C}{\partial T} = -\frac{S\varphi(d_1)\sigma}{2\sqrt{T}} - rXe^{-rT}\Phi(d_2)$
Interest rate r	Rho	$\rho = \frac{\partial C}{\partial r} = X \cdot T \cdot e^{-rT} \cdot \Phi(d_2)$

price of the underlying stock. Moreover, an asset manager employing options for risk management wants to know how the value of an option position will change as the price of the underlying changes.

One way to do so is to determine the derivative of the call option price with respect to the spot price of the underlying stock:

$$\Delta = \frac{\partial C}{\partial S} = \Phi(d_1) \tag{5}$$

This quantity is called the “delta” of the option, and can be used in the following way to determine the expected price change in the option if the stock increases by about \$1:

$$\Delta C = C(S + \$x) - C(S) \approx \frac{\partial C}{\partial S} \Delta S = \$x\Phi(d_1) \tag{6}$$

The relation given by (6) holds true for small changes in the price of the underlying. For large changes the assumed linear relationship between call and option price is not valid and we must apply a so-called convexity adjustment:

$$\Delta C = C(S + \$x) - C(S) \approx \frac{\partial C}{\partial S} \$x + \frac{1}{2} \cdot \underbrace{\frac{\partial^2 C}{\partial S^2}}_{=\Gamma} (\$x)^2$$

Here, Γ denotes the “options gamma,” which measures the curvature of the option price as a function of the price of the underlying stock.

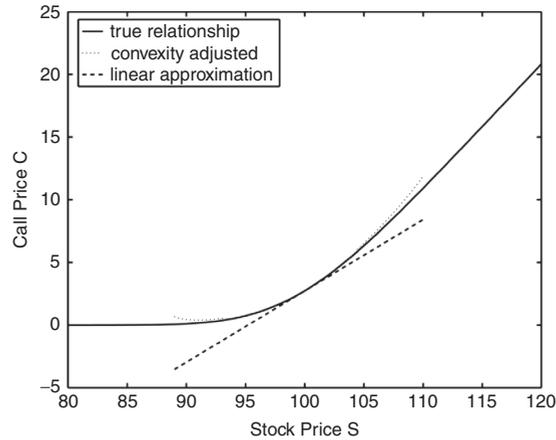


Figure 2 Accuracy of Simple Delta Approximation and Convexity-Adjusted Approximation
Note: The example is calculated for a one-month option with strike $X = \$100$ and current stock price $S = \$100$ with an interest of 10% per annum and volatility of 20% per annum.

Figure 2 visualizes this effect. We see that for small variations in the stock price the “true price” and both approximations nearly coincide. But for medium-sized variations, only the convexity-adjusted approximation is still accurate. For large variations in the underlying stock price both approximations fail.

The impact of the parameters stock price, interest rate, time to maturity, and volatility on the option’s delta is visualized in Figure 3. We can recognize that the influence of a change in the underlying on the option value measured by the option’s delta increases with increasing stock price. Intuitively, this is clear as for large values of the underlying stock the option behaves like the stock itself, whereas for values of the underlying stock near zero, the option is virtually worthless. Also, we can see that if the option is at the money, the impact of a change in the value of the underlying stock increases with increasing time to maturity and with increasing interest rate, which is not as obvious. The delta of the option that is at the money decreases with increasing volatility. The reason is as follows. Imagine that you possess an option on an underlying which is virtually nonrandom.

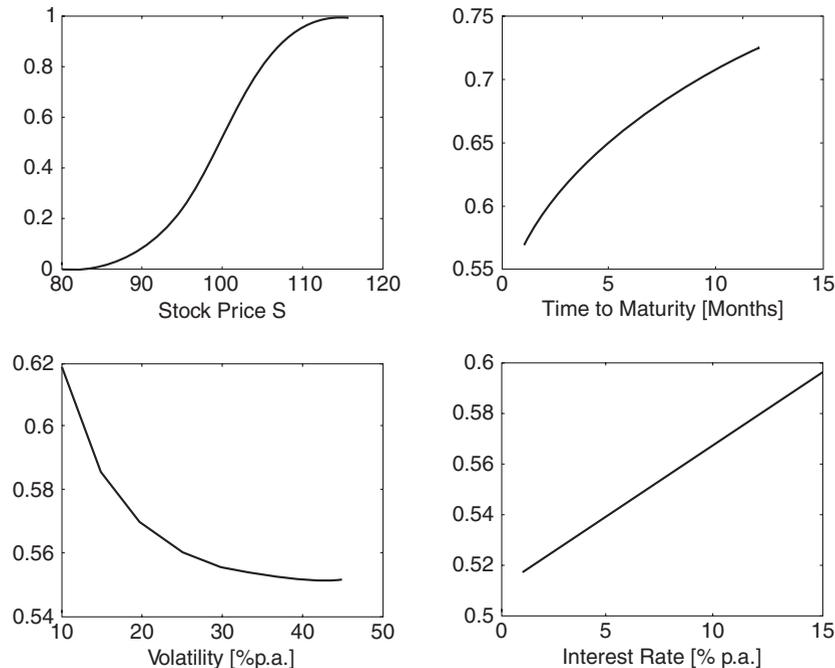


Figure 3 Delta as a Function of the Parameters

Note: The example is calculated for a one-month option with strike $X = \$100$ and current stock price $S = \$100$ with an interest of 10% per annum and a volatility of 20% per annum.

In this case, the value of the option equals its intrinsic value and therefore a change in the underlying stock price has a large impact on the value of the option provided that the current stock price is above the strike. In a stochastic environment (that is, nonzero volatility), every movement of the stock can be immediately followed by a movement in the opposite direction. This is why the option price is not as sensitive to stock price movements when volatility is high (that is, delta decreases with increasing volatility).

For gamma, it is clear that the impact of a change in the price of the underlying is the highest if the option is at the money. If the option is far out or far in the money, we have $C \approx 0$ or $C \approx S$ and, therefore, the second derivative with respect to S will vanish.

Below we will give a brief overview of the remaining sensitivity measures called theta, vega, and rho. Figure 4 visualizes the effect of the cur-

rent stock price on the Greeks gamma, theta, rho, and vega.

The Call Option Price and Time to Expiration: Theta

All other factors constant, the longer the time to expiration, the greater the option price. Since each day the option moves closer to the expiration date, the time to expiration decreases. The theta of an option measures the change in the option price as the time to expiration decreases, or equivalently, it is a measure of time decay.

Assuming that the price of the underlying stock does not change (which means that the intrinsic value of the option does not change), theta measures how quickly the time value of the option changes as the option moves toward expiration.

Buyers of options prefer a low theta so that the option price does not decline quickly as it

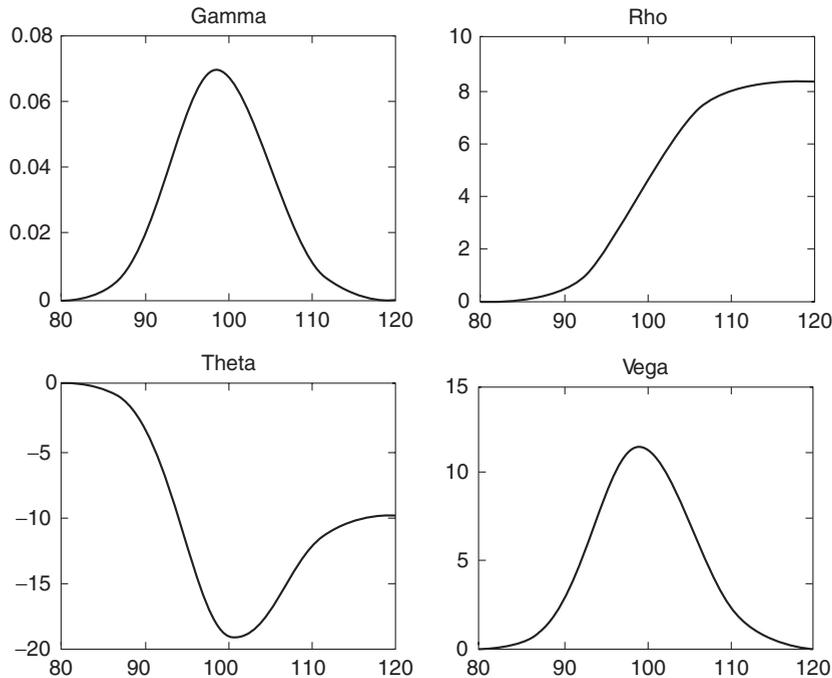


Figure 4 Variation of the Greeks with Respect to the Current Price of the Underlying Stock
Note: The example is calculated for a one-month option with strike $X = \$100$ and spot price $S = \$100$ with an interest of 10% per annum and a volatility of 20% per annum.

moves toward the expiration date. An option writer benefits from an option that has a high theta.

Option Price and Expected Volatility: Vega

All other factors constant, a change in the expected volatility will change the option price. The vega (also called “kappa”) of an option measures the dollar price change in the price of the option for a 1% change in the expected price volatility. (Vega is not a Greek letter. Vega is used to denote *volatility*, just as theta is used for *time* and rho is used for *interest rate*.) The option price is most sensitive with respect to a change in volatility when the option is at or near the money. This can be easily understood as follows. Imagine the option is very deep out of the money (that is, the option is virtually worthless). In this case, any small change in the volatility of the underlying will have no impact on the option price. It will still be nearly zero.

The same holds true if the option is far in the money (that is, it is nearly sure that the option will end in the money and the price of the option equals nearly the price of the stock). In this case, the impact of a small change in the volatility of the stock is negligible as well and, therefore, vega will be small. The situation is different if the option ranges near at the money. In this case, the option is very sensitive to volatility changes as they change the probability of ending in or out of the money dramatically. That is why we have a high vega for an option near the money.

Call Option Price and Interest Rate: Rho

The sensitivity of the option price to a change in the interest rate is called “rho.” The option’s rho is the least popular among the Greeks. Nevertheless, it is of practical value as it can be used to immunize a trader’s position against interest rate risk. An equivalent concept which might be familiar to some readers is the duration of

a bond. For our purposes, rho plays a minor role, and we have introduced it for the sake of completeness.

The Greeks and Portfolio Applications

In practical applications, the Greeks are used to hedge portfolios with respect to certain risk exposures. Because a portfolio is a linear combination of assets and as the derivative of a linear combination of functions equals the linear combination of the derivatives, we can simply calculate the Greek of a portfolio of options or other assets as the linear combination of the individual Greeks. When we seek to build a portfolio in a way that one or several of the Greeks equal zero, then the portfolio is said to be hedged with respect to the respective risk factor. A zero-delta portfolio, for example, is insensitive with respect to small changes in the value of S , and similarly for the other factors.

COMPUTING A PUT OPTION PRICE

We have focused our attention on call options. How do we value put options? This is done by using the following *put-call parity* relationship, which gives the relationship among the price of the common stock, the call option price, and the put option price. By simple no-arbitrage considerations, it can be shown that the following price identity must hold true for a European call and put option with the same strike and maturity:

$$\begin{aligned} \text{Call price} - \text{Put price} &= \text{Stock price} \\ &\quad - \text{Present value of dividends} \\ &\quad - \text{Present value of the strike} \end{aligned}$$

If we can calculate the fair value of a call option, the fair value of a put with the same strike price and expiration on the same stock can be calculated from the put-call parity relationship.

ASSUMPTIONS UNDERLYING THE BLACK-SCHOLES MODEL AND BASIC EXTENSIONS

The Black-Scholes model is based on several restrictive assumptions. These assumptions were necessary to develop the hedge to realize riskless arbitrage profits if the market price of the call option deviates from the price obtained from the model. Here, we will look at these assumptions and mention some basic extensions of the model that make pricing more realistic.

Taxes and Transactions Costs

The Black-Scholes model ignores taxes and transactions costs. The model can be modified to account for taxes, but the problem is that there is not one unique tax rate. Transactions costs include both commissions and the bid-ask spreads for the stock and the option, as well as other costs associated with trading options. This assumption, together with the next two, is the most important for the validity of the Black-Scholes model. The derivation of the price depends mainly on the existence of a replicating portfolio. When transaction costs exist, even if they are negligibly small, then the hedge portfolio can no longer be built and the argument leading to the uniqueness of the price fails.

Trading in Continuous Time, Short Selling, and Trading Arbitrary Fractions of Assets

One crucial assumption underlying the Black-Scholes model is the opportunity to (1) perform trades in continuous time; (2) buy a negative number of all traded assets (short selling); and (3) buy and sell arbitrary fractions of all traded assets. Only these more or less unrealistic assumptions together with the previously discussed absence of transaction costs and taxes allow the derivation of the unique call option

price by the hedging argument. The portfolio, consisting of certain fractions of the bond and the underlying stock, needs an ongoing rebalancing that is only possible in a market that allows continuous-time trading. Additionally, the number of stocks and bonds needed in the portfolio to replicate the option can be an arbitrary real number, possibly negative.

Variance of the Stock's Return

The Black-Scholes model assumes that the variance of the stock's return is (1) constant over the life of the option and (2) known with certainty. If (1) does not hold, an option pricing model can be developed that allows the variance to change. The violation of (2), however, is more serious. As the Black-Scholes model depends on the riskless hedge argument and, in turn, the variance must be known to construct the proper hedge, if the variance is not known, the hedge will not be riskless.

Stochastic Process Generating Stock Prices

To derive an option pricing model, an assumption is needed about the way stock prices move. The Black-Scholes model is based on the assumption that stock prices are generated by a geometric Brownian motion. Geometric Brownian motion is a stochastic process with continuous paths. In reality, one can sometimes observe that the market exhibits large fluctuations that cannot be explained by a continuous-time process with constant volatility as the Brownian motion. In theory, there are two possibilities to overcome this problem. Either one introduces the previously mentioned stochastic volatility or one allows for jumps in the stock price.

Risk-Free Interest Rate

In deriving the Black-Scholes model, two assumptions were made about the risk-free inter-

est rate. First, it was assumed that the interest rates for borrowing and lending were the same. Second, it was assumed that the interest rate was constant and known over the life of the option. The first assumption is unlikely to hold because borrowing rates are higher than lending rates. The effect on the Black-Scholes model is that the option price will be bound between the call price derived from the model using the two interest rates. The model can handle the second assumption by replacing the risk-free rate over the life of the option by the geometric average of the period returns expected over the life of the option. Returns on short-term Treasury bills cannot be known with certainty over the long term. Only the expected return is known, and there is a variance around it. The effects of variable interest rates are considered in Merton (1973).

BLACK-SCHOLES MODEL APPLIED TO THE PRICING OF OPTIONS ON BONDS: IMPORTANCE OF ASSUMPTIONS

While the Black-Scholes option pricing model was developed for nondividend paying stocks, it has been applied to options on bonds. We conclude this entry by demonstrating the limitations of applying the model to valuing options on bonds. This allows us to appreciate the importance of the assumptions on option pricing. To do so, let us look at the values that would be derived in a couple of examples.

We know that there are coupon-paying bonds and zero-coupon bonds. In our illustration we will use a zero-coupon bond. The reason is that the original Black-Scholes model was for common stock that did not pay a dividend and so a zero-coupon bond would be the equivalent type of instrument. Specifically, we look at how the Black-Scholes option pricing model would

value a zero-coupon bond with three years to maturity assuming the following:

Strike price = \$88.00

Time remaining to expiration = 2 years

Current bond price = \$83.96

Expected return volatility = Standard deviation
= 10% per annum

Risk-free rate = 6% per annum

The Black-Scholes model would give an option value of \$8.116. There is no reason to suspect that this value generated by the model is incorrect. However, let us change the problem slightly. Instead of a strike price of \$88, let us make the strike price \$100.25. The Black-Scholes option pricing model would give a fair value of \$2.79. Is there any reason to believe this is incorrect? Well, consider that this is a call option on a zero-coupon bond that will never have a value greater than its maturity value of \$100. Consequently, a call option with a strike price of \$100.25 must have a value of zero. Yet, the Black-Scholes option pricing model tells us that the value is \$2.79! In fact, if we assume a higher expected volatility, the Black-Scholes model would give an even greater value for the call option.

Why is the Black-Scholes model off by so much in our illustration? The answer is that there are three assumptions underlying the Black-Scholes model that limit its use in pricing options on fixed income instruments.

The first assumption is that the probability distribution for the underlying asset's prices assumed by the Black-Scholes model permits some probability—no matter how small—that the price can take on any positive value. But in the case of a zero-coupon bond, the price cannot take on a value above \$100. In the case of a coupon bond, we know that the price cannot exceed the sum of the coupon payments plus the maturity value. For example, for a five-year 10% coupon bond with a maturity value of \$100, the price cannot be greater than \$150 (five coupon payments of \$10 plus the maturity value of \$100). Thus, unlike stock prices, bond

prices have a maximum value. The only way that a bond's price can exceed the maximum value is if negative interest rates are permitted. While there have been instances where negative interest rates have occurred outside the United States, users of option pricing models assume that this outcome cannot occur. Consequently, any probability distribution for prices assumed by an option pricing model that permits bond prices to be higher than the maximum bond value could generate nonsensical option prices. The Black-Scholes model does allow bond prices to exceed the maximum bond value (or, equivalently, assumes that interest rates can be negative).

The second assumption of the Black-Scholes model is that the short-term interest rate is constant over the life of the option. Yet the price of an interest rate option will change as interest rates change. A change in the short-term interest rate changes the rates along the yield curve. Therefore, for interest rate options it is clearly inappropriate to assume that the short-term rate will be constant.

The third assumption is that the variance of returns is constant over the life of the option. As a bond moves closer to maturity, its price volatility declines and therefore its return volatility declines. Therefore, the assumption that variance of returns is constant over the life of the option is inappropriate.

KEY POINTS

- The most popular option pricing model is the Black-Scholes model.
- The factors that affect the value of an option include the current price of the asset, the strike price, the short-term risk-free interest rate, the time remaining to the expiration date of the option, and the expected return volatility.
- Option pricing models depend on the assumption regarding the distribution of returns.

- The option price derived from the Black-Scholes option pricing model is “fair” in the sense that if any other price existed in a market where all the assumptions of the Black-Scholes model are satisfied, riskless arbitrage profits can be realized by taking an offsetting position in the underlying asset.
- The sensitivity of the price of an option to a change in the value of a factor that affects the option’s price can be computed for any option pricing model. These sensitivity measures are referred to as the Greeks (delta, gamma, vega, theta, and rho).
- As with any economic model, there are assumptions that are made. When these assumptions are violated, the model value can depart significantly from the true value of the option.
- Using the Black-Scholes option pricing model to value an option on a bond is a good example where the model assumptions are not consistent with the realities of the bond market.

REFERENCES

- Black, F., and Scholes, M. (1973). The pricing of corporate liabilities. *Journal of Political Economy* 81: 637–659.
- Merton, R. (1973). The theory of rational option pricing. *Bell Journal of Economics and Management Science* 4: 141–183.

Pricing of Futures/Forwards and Options

FRANK J. FABOZZI, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

Abstract: There are various models that have been proposed to value financial assets in the cash market. Models for valuing derivatives such as futures, forwards, options, swaps, caps, and floors are valued using arbitrage principles. Basically, the price of a derivative is one that does not allow market participants to generate riskless profits without committing any funds. In developing a pricing model for derivatives, the model builder begins with a strategy (or trade) to exploit the difference between the cash price of the underlying asset for a derivative. The market price for the derivative is the cost of the package to replicate the payoff of the derivative.

Derivative instruments play an important role in financial markets as well as commodity markets by allowing market participants to control their exposure to different types of risk. When using derivatives, a market participant should understand the basic principles of how they are valued. While there are many models that have been proposed for valuing financial instruments that trade in the cash (spot) market, the valuation of all derivative models is based on arbitrage arguments. Basically, this involves developing a strategy or a trade wherein a package consisting of a position in the underlying (that is, the underlying asset or instrument for the derivative contract) and borrowing or lending so as to generate the same cash flow profile as the derivative. The value of the package is then equal to the theoretical price of the deriva-

tive. If the market price of the derivative deviates from the theoretical price, then the actions of arbitrageurs will drive the market price of the derivative toward its theoretical price until the arbitrage opportunity is eliminated.

In developing a strategy to capture any mispricing, certain assumptions are made. When these assumptions are not satisfied in the real world, the theoretical price can only be approximated. Moreover, a close examination of the underlying assumptions necessary to derive the theoretical price indicates how a pricing formula must be modified to value specific contracts.

In this entry, how futures, forwards, and options are valued is explained. The valuation of other derivatives such as swaps, caps, and floors is described in other entries.

PRICING OF FUTURES/FORWARD CONTRACTS

The pricing of futures and forward contracts is similar. If the underlying asset for both contracts is the same, the difference in pricing is due to differences in features of the contract that must be dealt with by the pricing model. To understand the differences, we begin with a definition of the two contracts.

A *futures contract* and a *forward contract* are agreements between a buyer and a seller, in which the buyer agrees to take delivery of the underlying at a specified price at some future date and the seller agrees to make delivery of the underlying at the specified price at the same future date. The buyer and the seller of the contract refers to the obligation that the party has in the future since neither party is obligated to transact in the underlying at the time of the trade. The futures price in the case of a futures contract or forward price in the case of a forward contract is the price at which the parties have agreed to transact in the future. The *settlement date* or *delivery date* is the future date when the two parties have agreed to transact (that is, buy or sell the underlying).

Differences between Futures and Forward Contracts

Futures contracts are standardized agreements as to the delivery date (or month) and quality of the deliverable, and are traded on organized exchanges. Associated with every futures exchange is a *clearinghouse*. The clearinghouse plays an important function: It guarantees that both parties to the trade will perform in the future. In the absence of a clearinghouse, the risk that the two parties face is that in the future when both parties are obligated to perform one of the parties will default. This risk faced in any derivative contract is referred to as counterparty risk. The clearinghouse allows the two parties to enter into a trade without

worrying about counterparty risk with respect to the counterparty to the trade. The reason is that after the trade is executed by the parties, the relationship between the two parties is terminated. The clearinghouse interposes itself as the buyer for every sale and the seller for every purchase. Consequently, the two parties to the trade are free to liquidate their positions without involving the original counterparty.

To protect itself against the counterparty risk of both the buyer and seller to the trade, the exchange where the contract is traded requires that when a position is first taken in a futures contract, both parties must deposit a minimum dollar amount per contract. This amount is specified by the exchange and referred to as *initial margin*. The parties have a choice of providing the initial margin in the form of cash or an interest-bearing security such as a Treasury bill. As the price of the futures contract fluctuates each trading day, the value of the equity of each party in the position changes. The equity in a futures margin account is measured by the sum of all margins posted and all daily gains less all daily losses to the account. To further protect itself against counterparty risk, the exchange specifies that the parties satisfy minimum equity positions. *Maintenance margin* is the minimum level that the exchange specifies that a party's equity position may fall as a result of an unfavorable price movement before a party is required to deposit additional margin. *Variation margin* is the additional margin that a party is required to provide in order to bring the equity in the margin account back to its initial margin level. If a party fails to furnish variation margin within 24 hours, the exchange closes the futures position out. Unlike initial margin, variation margin must be in cash rather than an interest-bearing security. Any excess margin in a party's margin account may be withdrawn.

In pricing futures contracts, the potential interim cash flows of futures contracts that are due to variation margin, in the case of adverse price movements, or withdrawal of cash for a party that experiences a favorable price

movement that results in the margin account's having excess margin must be taken into account.

We'll now compare these characteristics of a futures contract to a forward contract. A forward contract is an over-the-counter (OTC) instrument. That is, it is not an exchange-traded product. A forward contract is usually nonstandardized because the terms of each contract are negotiated individually between the parties to the trade. Also, there is no clearinghouse for trading forward contracts, and secondary markets are often nonexistent or extremely thin.

As just explained, futures contracts are marked to market at the end of each trading day. A forward contract may or may not be marked to market, depending on the wishes of the two parties. For example, both parties to a forward contract may be high-credit-quality entities. The parties may feel comfortable with the counterparty risk up to some specified amount and not require margin. Or one party may be satisfied with the high quality of the counterparty but the other party may not. In such cases, the forward contract may call for the marking to market of the position of only one of the counterparties. For a forward contract that is not marked to market, there are no interim cash-flow effects because no additional margin is required.

Other than these differences, which reflect the institutional arrangements in the two markets, most of what we say about the pricing of futures contracts applies equally to the pricing of forward contracts.

Basic Futures Pricing Model

We will illustrate the basic model for pricing futures contracts here. By "basic" we mean that we are extrapolating from the nuances of the underlying for a specific contract. The issues associated with applying the basic pricing model to some of the more popular futures contracts are described in other entries. Moreover, while the model described here is said to be a model

for pricing futures, technically, it is a model for pricing forward contracts with no mark-to-market requirements.

Rather than deriving the formula algebraically, the basic pricing model will be demonstrated using an illustration. We make the following six assumptions for a futures contract that has no initial and variation margin and which the underlying is asset U:

1. The price of asset U in the cash market is \$100.
2. There is a known cash flow for asset U over the life of the futures contract.
3. The cash flow for asset U is \$8 per year paid quarterly (\$2 per quarter).
4. The next quarterly payment is exactly three months from now.
5. The futures contract requires delivery three months from now.
6. The current three-month interest rate at which funds can be lent or borrowed is 4% per year.

The objective is to determine what the futures price of this contract should be. To do so, suppose that the futures price in the market is \$105. Let's see if that is the correct price. We can check this by implementing the following simple strategy:

- Sell the futures contract at \$105.
- Purchase asset U in the cash market for \$100.
- Borrow \$100 for three months at 4% per year (\$1 per quarter).

The purchase of asset U is accomplished with the borrowed funds. Hence, this strategy does not involve any initial cash outlay. At the end of three months, the following occurs

- \$2 is received from holding asset U.
- Asset U is delivered to settle the futures contract.
- The loan is repaid.

This strategy results in the following outcome:

From settlement of the futures contract:

Proceeds from sale of asset U to settle the futures contract	= \$105
Payment received from investing in asset U for three months	<u>= 2</u>
Total proceeds	= \$107

From the loan:

Repayment of principal of loan	= \$100
Interest on loan (1% for three months)	<u>= 1</u>
Total outlay	= \$101
<i>Profit from the strategy</i>	= \$6

The profit of \$6 from this strategy is guaranteed regardless of what the cash price of asset U is three months from now. This is because in the preceding analysis of the outcome of the strategy, the cash price of asset U three months from now never enters the analysis. Moreover, this profit is generated with no investment outlay; the funds needed to acquire asset U are borrowed when the strategy is executed. In financial terms, the profit in the strategy we have just illustrated arises from a riskless arbitrage between the price of asset U in the cash market and the price of asset U in the futures market.

In a well-functioning market, arbitrageurs who could realize this riskless profit for a zero investment would implement the strategy described above. By selling the futures and buying asset U in order to implement the strategy, this would force the futures price down so that at some price for the futures contract, the arbitrage profit is eliminated.

This strategy that resulted in the capturing of the arbitrage profit is referred to as a *cash-and-carry trade*. The reason for this name is that implementation of the strategy involves borrowing cash to purchase the underlying and “carrying” that underlying to the settlement date of the futures contract.

From the cash-and-carry trade we see that the futures price cannot be \$105. Suppose instead that the futures price is \$95 rather than \$105.

Let’s try the following strategy to see if that price can be sustained in the market:

- Buy the futures contract at \$95.
- Sell (short) asset U for \$100.
- Invest (lend) \$100 for three months at 1% per year.

We assume once again that in this strategy that there is no initial margin and variation margin for the futures contract. In addition, we assume that there is no cost to selling the asset short and lending the money. Given these assumptions, there is no initial cash outlay for the strategy just as with the cash-and-carry trade. Three months from now,

- Asset U is purchased to settle the long position in the futures contract.
- Asset U is accepted for delivery.
- Asset U is used to cover the short position in the cash market.
- Payment is made of \$2 to the lender of asset U as compensation for the quarterly payment.
- Payment is received from the borrower of the loan of \$101 for principal and interest.

More specifically, the strategy produces the following at the end of three months:

From settlement of the futures contract:

Price paid for purchase of asset U to settle futures contract	= \$95
Proceeds to lender of asset U to borrow the asset	<u>= 2</u>
Total outlay	= \$97

From the loan:

Principal from loan	= \$100
Interest earned on loan (\$1 for three months)	<u>= 1</u>
Total proceeds	= \$101
<i>Profit from the strategy</i>	= \$4

As with the cash and trade, the \$4 profit from this strategy is a riskless arbitrage profit. This strategy requires no initial cash outlay, but will generate a profit whatever the price of asset U is in the cash market at the settlement date. In real-world markets, this opportunity would lead arbitrageurs to buy the futures contract and short

asset U. The implementation of this strategy would be to raise the futures price until the arbitrage profit disappeared.

This strategy that is implemented to capture the arbitrage profit is known as a *reverse cash-and-carry trade*. That is, with this strategy, the underlying is sold short and the proceeds received from the short sale are invested.

We can see that the futures price cannot be \$95 or \$105. What is the theoretical futures price given the assumptions in our illustration? It can be shown that if the futures price is \$99 there is no opportunity for an arbitrage profit. That is, neither the cash-and-carry trade nor the reverse cash-and-carry trade will generate an arbitrage profit.

In general, the formula for determining the theoretical futures price given the assumptions of the model is:

$$\begin{aligned} &\text{Theoretical futures price} \\ &= \text{Cash market price} + (\text{Cash market price}) \\ &\quad \times (\text{Financing cost} - \text{Cash yield}) \quad (1) \end{aligned}$$

In the formula given by (1), “Financing cost” is the interest rate to borrow funds and “Cash yield” is the payment received from investing in the asset as a percentage of the cash price. In our illustration, the financing cost is 1% and the cash yield is 2%.

In our illustration, since the cash price of asset U is \$100, the theoretical futures price is:

$$\$100 + \$100 \times (1\% - 2\%) = \$99$$

The future price can be above or below the cash price depending on the difference between the financing cost and cash yield. The difference between these rates is called the *net financing cost*. A more commonly used term for the net financing cost is the *cost of carry*, or simply, *carry*. *Positive carry* means that the cash yield exceeds the financing cost. (Note that while the difference between the financing cost and the cash yield is a negative value, carry is said to be positive.) *Negative carry* means that the financing cost exceeds the cash yield. Below is a summary

of the effect of carry on the difference between the futures price and the cash market price:

Positive carry	Futures price will sell at a discount to cash price.
Negative carry	Futures price will sell at a premium to cash price.
Zero	Futures price will be equal to the cash price.

Note that at the settlement date of the futures contract, the futures price must equal the cash market price. The reason is that a futures contract with no time left until delivery is equivalent to a cash market transaction. Thus, as the delivery date approaches, the futures price will converge to the cash market price. This fact is evident from the formula for the theoretical futures price given by (1). The financing cost approaches zero as the delivery date approaches. Similarly, the yield that can be earned by holding the underlying approaches zero. Hence, the cost of carry approaches zero, and the futures price approaches the cash market price.

A Closer Look at the Theoretical Futures Price

In deriving theoretical futures price using the arbitrage argument, several assumptions had to be made. These assumptions as well as the differences in contract specifications will result in the futures price in the market deviating from the theoretical futures price as given by (1). It may be possible to incorporate these institutional and contract specification differences into the formula for the theoretical futures price. In general, however, because it is oftentimes too difficult to allow for these differences in building a model for the theoretical futures price, the end result is that one can develop bands or boundaries for the theoretical futures price. So long as the futures price in the market remains within the band, no arbitrage opportunity is possible.

Next, we will look at some of the institutional and contract specification differences that cause

prices to deviate from the theoretical futures price as given by the basic pricing model.

Interim Cash Flows

In the derivation of a basic pricing model, it is assumed that no interim cash flows arise because of changes in futures prices (that is, there is no variation margin). As noted earlier, in the absence of initial and variation margins, the theoretical price for the contract is technically the theoretical price for a forward contract that is not marked to market rather than a futures contract.

In addition, the model assumes implicitly that any dividends or coupon interest payments are paid at the settlement date of the futures contract rather than at any time between initiation of the cash position and settlement of the futures contract. However, we know that interim cash flows for the underlying for financial futures contracts do have interim cash flows. Consider stock index futures contracts and bond futures contracts.

For a stock index, there are interim cash flows. In fact, there are many cash flows that are dependent upon the dividend dates of the component companies. To correctly price a stock index future contract, it is necessary to incorporate the interim dividend payments. Yet, the dividend rate and the pattern of dividend payments are not known with certainty. Consequently, they must be projected from the historical dividend payments of the companies in the index. Once the dividend payments are projected, they can be incorporated into the pricing model. The only problem is that the value of the dividend payments at the settlement date will depend on the interest rate at which the dividend payments can be reinvested from the time they are projected to be received until the settlement date. The lower the dividend, and the closer the dividend payments to the settlement date of the futures contract, the less important the reinvestment income is in determining the futures price.

In the case of a Treasury futures contract, the underlying is a Treasury note or a Treasury bond. Unlike a stock index futures contract, the timing of the interest payments that will be made by the U.S. Department of the Treasury for a given issue that is acceptable as deliverable for a contract is known with certainty and can be incorporated into the pricing model. However, the reinvestment interest that can be earned from the payment dates to the settlement of the contract is unknown and depends on prevailing interest rates at each payment date.

Differences in Borrowing and Lending Rates

In the formula for the theoretical futures price, it is assumed in the cash-and-carry trade and the reverse cash-and-carry trade that the borrowing rate and lending rate are equal. Typically, however, the borrowing rate is higher than the lending rate. The impact of this inequality is important and easy to quantify.

In the cash-and-carry trade, the theoretical futures price as given by (1) becomes:

$$\begin{aligned} \text{Theoretical futures price based on borrowing rate} \\ = \text{Cash market price} + (\text{Cash market price}) \\ \times (\text{Borrowing rate} - \text{Cash yield}) \end{aligned} \quad (2)$$

For the reverse cash-and-carry trade, it becomes

$$\begin{aligned} \text{Theoretical futures price based on lending rate} \\ = \text{Cash market price} + (\text{Cash market price}) \\ \times (\text{Lending rate} - \text{Cash yield}) \end{aligned} \quad (3)$$

Formulas (2) and (3) together provide a band between which the actual futures price can exist without allowing for an arbitrage profit. Equation (2) establishes the upper value for the band while equation (3) provides the lower value for the band. For example, assume that the borrowing rate is 6% per year, or 1.5% for three months, while the lending rate is 4% per year, or 1% for three months. Using equation (2), the

upper value for the theoretical futures price is \$99.5 and using equation (3) the lower value for the theoretical futures price is \$99.

Transaction Costs

The two strategies to exploit any price discrepancies between the cash market and theoretical price for the futures contract will require the arbitrageur to incur transaction costs. In real-world financial markets, the costs of entering into and closing the cash position as well as round-trip transaction costs for the futures contract affect the futures price. As in the case of differential borrowing and lending rates, transaction costs widen the bands for the theoretical futures price.

Short Selling

The reverse cash-and-strategy trade requires the short selling of the underlying. It is assumed in this strategy that the proceeds from the short sale are received and reinvested. In practice, for individual investors, the proceeds are not received, and, in fact, the individual investor is required to deposit margin (securities margin and not futures margin) to short sell. For institutional investors, the underlying may be borrowed, but there is a cost to borrowing. This cost of borrowing can be incorporated into the model by reducing the cash yield on the underlying. For strategies applied to stock index futures, a short sale of the components stocks in the index means that all stocks in the index must be sold simultaneously. This may be difficult to do and therefore would widen the band for the theoretical future price.

Known Deliverable Asset and Settlement Date

In the two strategies to arbitrage discrepancies between the theoretical futures price and the cash market price, it is assumed that (1) only one asset is deliverable and (2) the settlement date occurs at a known, fixed point in the future. Neither assumption is consistent with the

delivery rules for some futures contracts. For U.S. Treasury note and bond futures contracts, for example, the contract specifies that any one of several Treasury issues that is acceptable for delivery can be delivered to satisfy the contract. Such issues are referred to as deliverable issues. The selection of which deliverable issue to deliver is an option granted to the party who is short the contract (that is, the seller). Hence, the party that is long the contract (that is, the buyer of the contract) does not know the specific Treasury issue that will be delivered. However, market participants can determine the cheapest-to-deliver issue from the issues that are acceptable for delivery. It is this issue that is used in obtaining the theoretical futures price. The net effect of the short's option to select the issue to deliver to satisfy the contract is that it reduces the theoretical future price by an amount equal to the value of the delivery option granted to the short.

Moreover, unlike other futures contracts, the Treasury bond and note contracts do not have a delivery date. Instead, there is a delivery month. The short has the right to select when in the delivery month to make delivery. The effect of this option granted to the short is once again to reduce the theoretical futures price below that given by equation (1). More specifically,

$$\begin{aligned} &\text{Theoretical futures price adjusted for delivery options} \\ &= \text{Cash market price} + (\text{Cash market price}) \\ &\quad \times (\text{Financing cost} - \text{Cash yield}) - \text{Value of the} \\ &\quad \text{delivery options granted to the short} \end{aligned} \quad (4)$$

Deliverable as a Basket of Securities

Some futures contracts have as the underlying a basket of assets or an index. Stock index futures are the most obvious example. At one time, municipal futures contracts were actively traded and the underlying was a basket of municipal securities. The problem in arbitraging futures contracts in which there is a basket of assets or an index is that it may be too expensive to buy or sell every asset included in the basket or index. Instead, a portfolio containing a smaller

number of assets may be constructed to track the basket or index (which means having price movements that are very similar to changes in the basket or index). Nonetheless, the two arbitrage strategies involve a tracking portfolio rather than a single asset for the underlying, and the strategies are no longer risk-free because of the risk that the tracking portfolio will not precisely replicate the performance of the basket or index. For this reason, the market price of futures contracts based on baskets of assets or an index is likely to diverge from the theoretical price and have wider bands.

Different Tax Treatment of Cash and Futures Transaction

Participants in the financial market cannot ignore the impact of taxes on a trade. The strategies that are implemented to exploit arbitrage opportunities between prices in the cash and futures markets and the resulting pricing model must recognize that there are differences in the tax treatment under the tax code for cash and futures transactions. The impact of taxes must be incorporated into the pricing model.

PRICING OF OPTIONS

Now we will look at the basic principles for valuing options. There are two parties to an option contract: the buyer and the writer or seller. The writer of the option grants the buyer of the option the right, but not the obligation, to either purchase from or sell to the writer something at a specified price within a specified period of time (or at a specified date). In exchange for the right that the writer grants the buyer, the buyer pays the writer a certain sum of money. This sum is called the option price or option premium. The price at which the underlying may be purchased or sold is called the *exercise price* or *strike price*. The option's *expiration date* (or maturity date) is the last date at which the option buyer can exercise the option. After the

expiration date, the contract is void and has no value.

There are two types of options: call options and put options. A *call option*, or simply call, is one in which the option writer grants the buyer the right to purchase the underlying. When the option writer grants the buyer the right to sell the underlying, the option is called a *put option*, or simply, a put.

The timing of the possible exercise of an option is an important characteristic of an option contract. An *American option* allows the option buyer to exercise the option at any time up to and including the expiration date. A *European option* allows the option buyer to exercise the option only on the expiration date.

As with futures and forward contracts, the theoretical price of an option is also derived based on arbitrage arguments. However, as will be explained, the pricing of options is not as simple as the pricing of futures and forward contracts.

Basic Components of the Option Price

The theoretical price of an option is made up of two components: the intrinsic value and a premium over intrinsic value.

Intrinsic Value

The *intrinsic value* is the option's economic value if it is exercised immediately. If no positive economic value would result from exercising immediately, the intrinsic value is zero. An option's intrinsic value is easy to compute given the price of the underlying and the strike price.

For a call option, the intrinsic value is the difference between the current market price of the underlying and the strike price. If that difference is positive, then the intrinsic value equals that difference; if the difference is zero or negative, then the intrinsic value is equal to zero. For example, if the strike price for a call option is \$100 and the current price of the underlying

is \$109, the intrinsic value is \$9. That is, an option buyer exercising the option and simultaneously selling the underlying would realize \$109 from the sale of the underlying, which would be covered by acquiring the underlying from the option writer for \$100, thereby netting a \$9 gain.

An option that has a positive intrinsic value is said to be *in-the-money*. When the strike price of a call option exceeds the underlying's market price, it has no intrinsic value and is said to be *out-of-the-money*. An option for which the strike price is equal to the underlying's market price is said to be *at-the-money*. Both at-the-money and out-of-the-money options have intrinsic values of zero because it is not profitable to exercise them. Our call option with a strike price of \$100 would be (1) in the money when the market price of the underlying is more than \$100; (2) out of the money when the market price of the underlying is less than \$100, and (3) at the money when the market price of the underlying is equal to \$100.

For a put option, the intrinsic value is equal to the amount by which the underlying's market price is below the strike price. For example, if the strike price of a put option is \$100 and the market price of the underlying is \$95, the intrinsic value is \$5. That is, the buyer of the put option who simultaneously buys the underlying and exercises the put option will net \$5 by exercising. The underlying will be sold to the writer for \$100 and purchased in the market for \$95. With a strike price of \$100, the put option would be (1) in the money when the underlying's market price is less than \$100, (2) out of the money when the underlying's market price exceeds \$100, and (3) at the money when the underlying's market price is equal to \$100.

Time Premium

The *time premium* of an option, also referred to as the time value of the option, is the amount by which the option's market price exceeds its intrinsic value. It is the expectation of the option buyer that at some time prior to the ex-

piration date changes in the market price of the underlying will increase the value of the rights conveyed by the option. Because of this expectation, the option buyer is willing to pay a premium above the intrinsic value. For example, if the price of a call option with a strike price of \$100 is \$12 when the underlying's market price is \$104, the time premium of this option is \$8 (\$12 minus its intrinsic value of \$4). Had the underlying's market price been \$95 instead of \$104, then the time premium of this option would be the entire \$12 because the option has no intrinsic value. All other things being equal, the time premium of an option will increase with the amount of time remaining to expiration.

An option buyer has two ways to realize the value of an option position. The first way is by exercising the option. The second way is to sell the option in the market. In the first example above, selling the call for \$12 is preferable to exercising, because the exercise will realize only \$4 (the intrinsic value), but the sale will realize \$12. As this example shows, exercise causes the immediate loss of any time premium. It is important to note that there are circumstances under which an option may be exercised prior to the expiration date. These circumstances depend on whether the total proceeds at the expiration date would be greater by holding the option or exercising and reinvesting any received cash proceeds until the expiration date.

Put-Call Parity Relationship

For a European put and a European call option with the same underlying, strike price, and expiration date, there is a relationship between the price of a call option, the price of a put option, the price of the underlying, and the strike price. This relationship is known as the *put-call parity relationship*. The relationship is:

$$\begin{aligned} \text{Put option price} - \text{Call option price} &= \text{Present value} \\ &\text{of strike price} + \text{Present value of cash distribution} \\ &- \text{Price of underlying} \end{aligned}$$

Factors That Influence the Option Price

The factors that affect the price of an option include the:

- Market price of the underlying.
- Strike price of the option.
- Time to expiration of the option.
- Expected volatility of the underlying over the life of the option.
- Short-term, risk-free interest rate over the life of the option.
- Anticipated cash payments on the underlying over the life of the option.

The impact of each of these factors may depend on whether (1) the option is a call or a put, and (2) the option is an American option or a European option. Table 1 summarizes how each of the six factors listed above affects the price of a put and call option. Here, we briefly explain why the factors have the particular effects.

Market Price of the Underlying Asset

The option price will change as the price of the underlying changes. For a call option, as the underlying's price increases (all other factors being constant), the option price increases. The opposite holds for a put option: As the price of the underlying increases, the price of a put option decreases.

Table 1 Summary of Factors That Affect the Price of an Option

Factor	Effect of an Increase of Factor On	
	Call Price	Put Price
Market price of underlying	Increase	Decrease
Strike price	Decrease	Increase
Time to expiration of option	Increase	Increase
Expected volatility	Increase	Increase
Short-term, risk-free interest rate	Increase	Decrease
Anticipated cash payments	Decrease	Increase

Strike Price

The strike price is fixed for the life of the option. All other factors being equal, the lower the strike price, the higher the price for a call option. For put options, the higher the strike price, the higher the option price.

Time to Expiration of the Option

After the expiration date, an option has no value. All other factors being equal, the longer the time to expiration of the option, the higher the option price. This is because, as the time to expiration decreases, less time remains for the underlying's price to rise (for a call buyer) or fall (for a put buyer), and therefore the probability of a favorable price movement decreases. Consequently, as the time remaining until expiration decreases, the option price approaches its intrinsic value.

Expected Volatility of the Underlying over the Life of the Option

All other factors being equal, the greater the expected volatility (as measured by the standard deviation or variance) of the underlying, the more the option buyer would be willing to pay for the option, and the more an option writer would demand for it. This occurs because the greater the expected volatility, the greater the probability that the movement of the underlying will change so as to benefit the option buyer at some time before expiration.

Short-Term, Risk-Free Interest Rate over the Life of the Option

Buying the underlying requires an investment of funds. Buying an option on the same quantity of the underlying makes the difference between the underlying's price and the option price available for investment at an interest rate at least as high as the risk-free rate. Consequently, all other factors being constant, the higher the short-term, risk-free interest rate, the

greater the cost of buying the underlying and carrying it to the expiration date of the call option. Hence, the higher the short-term, risk-free interest rate, the more attractive the call option will be relative to the direct purchase of the underlying. As a result, the higher the short-term, risk-free interest rate, the greater the price of a call option.

Anticipated Cash Payments on the Underlying over the Life of the Option

Cash payments on the underlying tend to decrease the price of a call option because the cash payments make it more attractive to hold the underlying than to hold the option. For put options, cash payments on the underlying tend to increase the price.

Option Pricing Models

Earlier in this entry, it was explained how the theoretical price of a futures contract and forward contract can be determined on the basis of arbitrage arguments. An option pricing model uses a set of assumptions and arbitrage arguments to derive a theoretical price for an option. Deriving a theoretical option price is much more complicated than deriving a theoretical futures or forward price because the option price depends on the expected volatility of the underlying over the life of the option.

Several models have been developed to determine the theoretical price of an option. The most popular one was developed by Fischer Black and Myron Scholes (1973) for valuing European call options on common stock. The Black-Scholes model requires as input the six factors discussed above that affect the value of an option. Several modifications to the Black-Scholes model followed. One such model is the lattice model suggested by Cox, Ross, and Rubinstein (1979), Rendleman and Bartter (1979), and Sharpe (1981).

Basically, the idea behind the arbitrage argument is that if the payoff from owning a call option can be replicated by (1) purchasing the underlying for the call option and (2) borrowing funds to purchase the underlying, then the cost of creating the replicating strategy (position) is the theoretical price of the option.

KEY POINTS

- For futures and forward contracts, the theoretical price can be derived using arbitrage arguments. Specifically, a cash-and-carry trade can be implemented to capture the arbitrage profit for an overpriced futures or forward contract while a reverse cash-and-carry trade can be implemented to capture the arbitrage profit for an underpriced futures or forward contract.
- The basic model states that the theoretical futures price is equal to the cash market price plus the net financing cost. The net financing cost, also called the cost of carry, is the difference between the financing cost and the cash yield on the underlying.
- Because of institutional and contract specification differences, the market price for the futures or forward contract can deviate from the theoretical price without any arbitrage opportunities being possible. Basically, a band can be established for the theoretical futures price and as long as the market price for the futures contract is not outside of the band, there is no arbitrage opportunity.
- The two components of the price of an option are the intrinsic value and the time premium. The former is the economic value of the option if it is exercised immediately, while the latter is the amount by which the option price exceeds the intrinsic value.
- The option price is affected by six factors: (1) the market price of the underlying; (2) the strike price of the option; (3) the time

remaining to the expiration of the option; (4) the expected volatility of the underlying as measured by the standard deviation; (5) the short-term, risk-free interest rate over the life of the option; and (6) the anticipated cash payments on the underlying.

- It is the uncertainty about the expected volatility of the underlying that makes valuing options more complicated than valuing futures and forward contracts.
- There are various models for determining the theoretical price of an option. These include the Black-Scholes model and the lattice model.

REFERENCES

- Black, F., and Scholes, M. (1973). Pricing of options and corporate liabilities. *Journal of Political Economy* 81, 3: 637–654.
- Collins, B., and Fabozzi, F. J. (1999). *Derivatives and Equity Portfolio Management*. New York: John Wiley & Sons.
- Cox, J. C., Ross, S. A., and Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of Financial Economics* 7: 229–263.
- Rendleman, R. J., and Bartter, B. J. (1979). Two-state option pricing. *Journal of Finance* 34, 5: 1093–1110.
- Sharpe, W. F. (1981). *Investments*. Englewood Cliffs, NJ: Prentice Hall.

Pricing Options on Interest Rate Instruments

RADU TUNARU, PhD

Professor of Quantitative Finance, Business School, University of Kent

BRIAN EALES

Academic Leader (Retired), London Metropolitan University

Abstract: Interest rate modeling has quickly become one of the main areas in financial markets. The models have grown in sophistication in response to development of new products and structures. Almost all pricing of securities and the risk management function, including marking-to-market, relies on interest rate modeling of some description. The information on interest rates, usually conveyed from the options markets, is important for other markets as well, such as the more established credit risk, commodities, equities, and the more recent ones such as inflation derivatives and insurance derivatives. Many models have been developed over the years, and their advantages and disadvantages should be appreciated and understood when they are applied.

Throughout the world, *interest rates* serve as instruments of control. When inflation rises to an undesirable or politically unacceptable level, the appropriate authorities raise interest rates to curb expenditure. In times when economic activity and corporate and consumer confidence is less buoyant, the policy is to lower rates. Interest rate derivatives were among the first contracts to be offered on derivative exchanges and have their origins in the period following the breakdown of the Bretton Woods Agreement. In today's sometimes volatile markets, they continue to be extremely useful tools for corporates, banks, and individuals from hedging, financial engineering, and speculative perspectives.

Two of the early prime movers in the interest rate derivatives market were the Chicago Board

of Trade (CBOT) and the Chicago Mercantile Exchange (CME). (In July 2007 the CBOT and CME merged to form the CME group.) Some of the contracts that were introduced in the 1970s are still popular today, as evidenced by the high volume they enjoy.

At the short end of the yield curve, the CME has the world's most actively traded exchange-based interest rate option contracts: Eurodollar options. Each Eurodollar option has as its underlying a Eurodollar time deposit futures contract with a principal value of \$1 million, which will be cash settled at maturity. Another high-volume contract available on the CME is the option on the one-month Eurodollar futures contract. This, too, is cash settled at maturity.

CME also, has option contracts on U.S. Treasury bonds and notes in its portfolio of interest rate derivative products. There are American-style options available on bonds with a maturity of at least 25-years, between 15 but less than 25 years, 10-year, 5-year, 3-year and 2-year notes, the most heavily traded of which is the 10-year Treasury note. The option contract has as its deliverable one U.S. Treasury 10-year note futures contract with a face value of \$100,000 at maturity. Whereas the CME futures contracts on short-term interest rates are cash settled, the bond/note futures require physical settlement. The corresponding option contracts have similar settlement requirements identified in their contract specifications. The CME Group lists seven international data vendors who provided quotes for call/put options across a range of strike prices and maturities.

Although option prices are easy to read and interpret from vendor screens, there is a mass of academic and practitioner research literature, which provides a platform from which *bond option prices* in general can be calculated with integrity. The literature on modeling interest rate derivatives in this arena is frequently divided into one- or two-factor (or multifactor) models.

- Calculating option prices in a one-factor model usually proposes that the process is driven by the short rate, often with a mean-reversion feature linked to the short rate. There are several popular models that fall into this category, for example, the *Vasicek model* and the Cox, Ingersoll, and Ross (CIR) model, both of which will be discussed in more detail later.
- Calculating option prices in a two-factor model involves both the short- and long-term rates linked by a mean-reversion process.

The problem with some of the preceding models is that they generate their own term structures, which, in the absence of adjustment, do not match the term structure observed in the market. A category of arbitrage-free models proposed by Ho and Lee (1986); Hull and White

(1990); and Black, Derman, and Toy (1990) seeks to eliminate this problem. For example, the Black, Derman, and Toy model enjoys a degree of popularity among market practitioners, since it takes account of and matches the term structure observed in the market, it eliminates the possibility of generating negative interest rates, and it models the observed interest rate volatility. These models together with other propositions will be discussed in more detail in this entry.

In order to examine some of the major developments in option/derivative pricing in the interest rate field, it is appropriate at this point to establish a working framework.

MODELING THE TERM STRUCTURE AND BOND PRICES

Let $(\Omega, \Sigma, \{F_t\}_{t \geq 0}, Q)$ be a filtered probability space modeling a financial market, where the filtration $F = \{F_t\}_{t \geq 0}$ describes the flux of information and the probability measure Q denotes the risk-neutral measure; the real-world or physical measure will be denoted by P . The starting point in modeling bond prices is the assumption that there is a bank account $B = \{B(t)\}_{t \geq 0}$ that is linked to the bank instantaneous interest rate (also called short rate, spot rate) process $r = \{r(t)\}_{t \geq 0}$ through

$$dB(t) = r(t)B(t)dt \quad \text{or} \\ B(t) = B(0) \exp \left[\int_0^t r(s)ds \right] \quad (1)$$

From a practical point of view, we can safely assume that the majority of stochastic processes representing prices of traded financial assets are adapted to the filtration F and that the short-rate process $r = \{r(t)\}_{t \geq 0}$ is a predictable process, meaning that $r(t)$ is F_{t-1} measurable. This implies that $B(t)$ is also F_{t-1} measurable and this condition is automatically satisfied for continuous or left-continuous processes.

In this entry we consider only default-free securities. We shall denote by $p(t, T)$ the price at time t of a pure discount bond with maturity T and obviously $p(t, t) = p(T, T) = 1$.

The following relationships are well known in the fixed-income area:

$$0 < p(t, T) \leq 1, \quad r(t) = \frac{\partial \ln p(t, T)}{\partial t} \Big|_{T=t}$$

$$= - \frac{\partial \ln p(t, T)}{\partial T} \Big|_{T=t}, \quad \text{for any } t < T \quad (2)$$

Let $f(t, s)$ be the *forward rate* at time $s > 0$ calculated at time $t < s$. The instantaneous forward rate at time t to borrow at time T can be calculated from the bond prices using

$$f(t, T) = - \frac{\partial \ln p(t, T)}{\partial T} \quad (3)$$

The reverse works as well; if forward rates are known, then bond prices can be calculated via $p(t, T) = e^{-\int_t^T f(t, s) ds}$. The short rate is intrinsically related to the forward rates because $r(t) \equiv f(t, t)$.

Short-Rate Models of Term Interest Rate Structure

Many models proposed for the short-rate process $r = \{r(t)\}_{t \geq 0}$ are particular cases of the general diffusion equation:

$$dr(t) = a(t, r(t))dt + b(t, r(t))dW(t) \quad (4)$$

where $W = \{W(t)\}_{t \geq 0}$ is a standard Wiener process defined on $(\Omega, \Sigma, \{F_t\}_{t \geq 0}, Q)$. The following list of models describes a chronological evolution without claiming that it is an exhaustive list:

The Merton model (Merton, 1973) is

$$dr(t) = \alpha dt + \sigma dW(t) \quad (5)$$

The Vasicek model (Vasicek, 1977) model is

$$dr(t) = (\alpha - \beta r(t))dt + \sigma dW(t) \quad (6)$$

One advantage of the Vasicek model is that the conditional distribution of r at any future time, given the current interest rates at time t , is

normally distributed. The main moments are

$$E_t(r(s)) = \frac{\alpha}{\beta} + \left(r(t) - \frac{\alpha}{\beta} \right) e^{-\beta(s-t)}, \quad t \leq s$$

$$\text{var}_t[r(s)] = \frac{\sigma^2}{2\beta} (1 - e^{-2\beta(s-t)}), \quad t \leq s \quad (7)$$

$$\text{cov}_t[r(u), r(s)] = \frac{\sigma^2}{2\beta} e^{-\beta(s+u-2t)} (e^{2\beta(u-t)} - 1),$$

$$t \leq u \leq s$$

Another advantage is that this model can be also derived within a general equilibrium framework as illustrated by Campbell (1986).

One disadvantage that is often discussed in the *interest rate modeling* literature is that there is a long-run possibility of negative interest rates. However, Rabinovitch (1989) proved that when the initial interest rate $r(0)$ is positive and the parameter estimates have reasonable values, the expected first-passage time of the process through the origin is longer than nine months. This result supports the use of the Vasicek model in practice since the majority of options traded on the organized exchanges expire in less than nine months.

The Dothan model (Dothan, 1978) is

$$dr(t) = \alpha r(t)dt + \sigma r(t)dW(t) \quad (8)$$

This is the same model as Rendleman and Bartter's model (Rendleman and Bartter, 1980). This model is the only lognormal single-factor model that leads to closed formulae for pure discount bonds. Nonetheless, there is no closed formula for a European option on a pure discount bond.

The *Cox-Ingersoll-Ross (CIR) models* (Cox, Ingersoll, and Ross, 1980, 1985) are

$$dr(t) = \beta(r(t))^{3/2}dW(t)$$

$$dr(t) = (\alpha - \beta r(t))dt + \sigma(r(t))^{1/2}dW(t) \quad (9)$$

CIR wrote arguably the first of several papers developing one-factor models of the term structure of interest rates. Around the same time models in the same spirit include the Vasicek, Dothan, Courtadon (1982), and Brennan and Schwartz (1979) models. The movements of

longer-maturity instruments are perfectly correlated with the instantaneous short-term rates.

The Ho-Lee model (Ho and Lee, 1986) is

$$dr(t) = \alpha(t)dt + \sigma dW(t) \quad (10)$$

This is the continuous version of the original model that was probably the first model designed to match exactly the observable term structure of interest rates.

The Black-Derman-Toy (BDT) model (Black, Derman, and Toy, 1990) is

$$dr(t) = \alpha(t)r(t)dt + \sigma(t)dW(t) \quad (11)$$

The Hull-White (HW) models (Hull and White, 1990, 1994, 1996) are

$$\begin{aligned} dr(t) &= [\alpha(t) - \beta(t)r(t)]dt + \sigma(t)dW(t) \\ dr(t) &= [\alpha(t) - \beta(t)r(t)]dt + \sigma(t)(r(t))^{1/2}dW(t) \end{aligned} \quad (12)$$

These models are two more general families of models incorporating the Vasicek model and CIR model, respectively. The first one is more often used and it can be calibrated to the observable term structure of interest rates and the volatility term structure of spot or forward rates. However, its implied volatility structures may be unrealistic. Hence, it may be wise to use a constant coefficient $\beta(t) = \beta$ and a constant volatility parameter $\sigma(t) = \sigma$ and then calibrate the model using only the term structure of market interest rates. It is still theoretically possible that the short rate r may go negative. The risk-neutral probability for the occurrence of such an event is

$$Q(r(t) < 0) = N\left(-\frac{\tilde{f}(0, t) + \frac{\sigma^2}{2\beta^2}(1 - e^{\beta t})^2}{\sqrt{\frac{\sigma^2}{2\beta^2}(1 - e^{2\beta t})}}\right) \quad (13)$$

where $\tilde{f}(0, t)$ is the market instantaneous forward rate. In practice, this probability seems to be rather small, as empirical evidence illustrated by Brigo and Mercurio (2007) shows. However, the probability is not zero, and this may bother some analysts.

An example will provide an idea of how a variation of one of the models proposed by Hull

and White described above by the first of (12) models can be used to price an option on a zero-coupon bond. If the assumptions are made that both β , the reversion rate, and σ , volatility, are constant, then the model can be restated as:

$$dr(t) = [\alpha(t) - \beta r(t)]dt + \sigma dW(t) \quad (14)$$

and the function $\alpha(t)$ can be calculated from a given term structure using:

$$\alpha(t) = f_T(0, t) + \beta f(0, t) + \frac{\sigma^2}{2\beta}(1 - e^{-2\beta t}) \quad (15)$$

The future market price of a zero-coupon bond in this framework can be found by defining the reversion rate, β , the volatility, and the time period involved.

$$p(T_0, T) = A(T_0, T)e^{-B(T_0, T)r(T_0)} \quad (16)$$

where T_0 represents the forward date at which the bond is to be priced, T represents the bond's maturity date, t is a time period index typically taken to be equal to zero (that is, representing the current point in time)

$$B(T_0, T) = \frac{1}{\beta}(1 - e^{-\beta(T-T_0)}) \quad (17)$$

$$\begin{aligned} \ln A(T_0, T) &= \ln\left(\frac{p(t, T)}{p(t, T_0)}\right) \\ &\quad - B(T_0, T) \frac{\partial \ln p(t, T_0)}{\partial T} \\ &\quad - \frac{1}{4\beta^3} \sigma^2 \left(e^{-\beta(T-t)} - e^{-\beta(T_0-t)}\right)^2 \\ &\quad \times \left(e^{2\beta(T_0-t)} - 1\right) \end{aligned} \quad (18)$$

and $r(T)$ is the prevailing short rate at the forward date.

To illustrate how this works consider the case where we wish to find the 1-year forward price of a bond with 4 years remaining to maturity. Assume that the yield curve offers 4.00% continuously compounded for all maturities, volatility is 2.00%, and the reversion rate is 0.1. In this example T is 4 and T_0 is 1. The price of the bond can be found using $p(1, 4) = A(1, 4)e^{-B(1, 4)(0.04)}$. Clearly, $A(1, 4)$ and $B(1, 4)$

must be evaluated. Starting with $B(1,4)$ we have $B(1, 4) = \frac{1}{0.1} (1 - e^{-0.1(4-1)}) = 2.5918 \dots$

The next step requires the evaluation of $A(1, 4)$ and the expression for $\ln A(1, 4)$ can be broken down into a series of relatively straightforward calculations:

$$\begin{aligned} \ln \left(\frac{p(t, T)}{p(t, T_0)} \right) &= \ln \left(\frac{p(0, 4)}{p(0, 1)} \right) = \ln \left(\frac{e^{-(4)(0.04)}}{e^{-(1)(0.04)}} \right) \\ &= \ln \left(\frac{0.8521}{0.9607} \right) = -0.12 \end{aligned}$$

$B(1, 4)$ has already been calculated and is equal to 2.5918. Moreover, $\frac{\partial \ln p(t, T_0)}{\partial T_0}$ can be approximated by $\left(\frac{\ln p(t, T_0 + \Delta t) - \ln p(t, T_0 - \Delta t)}{2\Delta t} \right)$ which, if a time interval, Δt , is assumed to be 0.1 years yields $\left(\frac{\ln p(0, 1 + 0.1) - \ln p(0, 1 - 0.1)}{2(0.1)} \right) = -0.04$. This leaves the expression:

$$\begin{aligned} &\frac{1}{4\beta^3} \sigma^2 \left(e^{-\beta(T-t)} - e^{-\beta(T_0-t)} \right)^2 \left(e^{2\beta(T_0-t)} - 1 \right) \\ &= \frac{1}{4(0.1)^3} 0.02^2 \left(e^{-(0.1)(4)} - e^{-(0.1)(1)} \right)^2 \\ &\quad \times \left(e^{2(0.1)(1)} - 1 \right) = 0.001217 \end{aligned}$$

Combining all the above calculations we find $\ln A(1, 4) = -0.01754$ and then the one-year forward bond price is $p(1, 4) = e^{-0.01754 \dots - 2.5918 \dots (0.04)} = 0.8858$.

The Black-Karasinski (BK) model (Black and Karasinski, 1991) is

$$dr(t) = r(t)[\alpha(t) - \beta(t) \ln r(t)]dt + \sigma(t)r(t)dW(t) \tag{19}$$

The BDT, HW, and BK models extended the Ho-Lee model to match a term structure volatility curve (e.g., the *cap* prices) in addition to the term structure. The BK model is a generalization of the BDT model, and it overcomes the problem of negative interest rates, assuming that the short rate r is the exponential of an OU process having time-dependent coefficients. It is popular with practitioners because it fits well the swaption volatility surface. Nevertheless, it does not have closed formulae for bonds or options on bonds.

The Sandmann-Sondermann model (Sandmann and Sondermann, 1993) is

$$\begin{aligned} r(t) &= \ln(1 + \eta(t)) \\ d\eta(t) &= \eta(t)(\alpha(t)dt + \sigma(t)dW(t)) \end{aligned} \tag{20}$$

The Dothan model, BKi model, and the exponential Vasicek model given below imply that r is lognormally distributed. While this finding may seem reasonable, it is the cause for the explosion of the bank account; that is, from a single unit of money, one may be able to make in an infinitesimal interval of time an infinite amount of money. The Sandmann-Sondermann model overcomes this problem by modeling the short rates as above.

The Chen model (Chen, 1995) is

$$\begin{aligned} dr(t) &= (\alpha(t) - r(t))dt + (\sigma(t)r(t))^{1/2} dW^1(t) \\ d\alpha(t) &= (\alpha - \alpha(t))dt + (\alpha(t))^{1/2} dW^2(t) \\ d\sigma(t) &= (\gamma - \sigma(t))dt + (\sigma(t))^{1/2} dW^3(t) \end{aligned} \tag{21}$$

where α, γ are constants and W^1, W^2 , and W^3 are independent Wiener processes. This is an example of a three-factor model.

The Schmidt model (Schmidt, 1997) is

$$r(t) = H[f(t) + g(t)W(T(t))] \tag{22}$$

where $T = T(t)$ and $H = H(x)$ are continuous nonnegative strictly increasing functions of $t \geq 0$ and real x , while $f = f(t)$ and $g = g(t) > 0$ are continuous functions.

The exponential Vasicek model is

$$dr(t) = r(t) [\eta - a \ln r(t)] dt + \sigma r(t) dW(t) \tag{23}$$

This model is similar to the Dothan model, being a lognormal short-rate model. This model does not lead to explicit formulae for pure discount bonds or for options contingent on them. In addition, this is an example of a nonaffine term-structure model.

The Mercurio-Moraleda model (Mercurio and Moraleda, 2000) is

$$\begin{aligned} dr(t) &= r(t) \left[\eta(t) - \left(\lambda - \frac{\gamma}{1 + \gamma t} \right) \ln r(t) \right] dt \\ &\quad + \sigma r(t) dW(t) \end{aligned} \tag{24}$$

The CIR++ model (Brigo and Mercurio, 2007) is

$$\begin{aligned} r(t) &= x(t) + \varphi(t) \\ dx(t) &= k[\theta - x(t)]dt + \sigma\sqrt{x(t)}dW(t) \end{aligned} \quad (25)$$

The extended exponential Vasicek model (Brigo and Mercurio, 2007) is

$$\begin{aligned} r(t) &= x(t) + \varphi(t) \\ dx(t) &= x(t)[\eta - \lambda \ln x(t)]dt + \sigma x(t)dW(t) \end{aligned} \quad (26)$$

Two-factor models were based on a second source of random shocks. Two-factor models were developed by Brennan and Schwartz (1982), Fong and Vasicek (1992), and Longstaff and Schwartz (1992a). However, Hogan (1993) proved that the solution to the Brennan and Schwartz model explodes, that is, reaches infinity in a finite amount of time with positive probability. This shows that adding more factors may cause unseen problems. More complex multifactor models are described by Rebonato (1998) and by Brigo and Mercurio (1997).

Therefore, the *short-rate models* lead to two main problems. Mean-reverting models such as Vasicek or Hull and White may produce negative interest rates. From a computational perspective, if the risk-neutral probability of producing such negative rates is negligible, then those scenarios can simply be ignored in a Monte Carlo setup. The so-called lognormal models ensure nonnegativity of interest rates but may become explosive due to the change of scale in the short-rate modeling. Multifactor short-rate models become rapidly computationally infeasible, and they may produce volatility surfaces that do not match those observed in the markets.

The problems signaled above for the short-rate models led to the development of new classes of models, more notably the LIBOR market models or BGM developed by Brace, Gatarek, and Musiela (1997); Jamshidian (1997); and Musiela and Rutkowski (1997). This model starts with a geometric Brownian motion for the forward LIBOR rate $L_i(t) := L(t; T_i, T_{i+1})$,

where $0 = T_0 < T_1 < \dots < T_n$ to acquire positivity of rates

$$dL_i(t) = \mu_i^Q(t)L_i(t)dt + \sigma_i(t)L_i(t)dW_i^Q(t) \quad (27)$$

where Q is the martingale measure corresponding to the numeraire $N(t) = p(t, T_n)$, also called the terminal measure because the numeraire is the price of the bond with the last tenor. Now $\prod_{k=i}^{n-1} (1 + (T_{k+1} - T_k)L_k(t)) = \frac{p(t, T_i)}{p(t, T_n)}$ is the numeraire rebased price of a traded asset, the zero-coupon bond with maturity T_i . Hence, it should be a martingale and its drift must be zero. Calculating the drift with Ito calculus for all consecutive indexes $i, i + 1$ allows the drift determination

$$\mu_i^Q(t) = \sum_{\substack{k \geq i+1 \\ k \leq n}} \frac{(T_{k+1} - T_k)L_k(t)}{[1 + (T_{k+1} - T_k)L_k(t)]} \sigma_i(t)\sigma_k(t)\rho_{i,k}(t) \quad (28)$$

for all $i \in \{0, \dots, n - 1\}$.

Other numeraires are also feasible but lead to a different style of *calibration*. The pricing of interest rate derivatives is realized with Monte Carlo simulation.

The quest for ensuring positiveness of the short rates motivated the development of a new class sometimes called Markov functional models. Important contributions in this area are Flesaker and Hughston (1996), Rogers (1997), and Rutkowski (1997), although some seminal ideas are also contained in Constantinides (1992). In a nutshell, given a strictly positive diffusion process $\{D(t)\}_{t \geq 0}$ adapted to the filtration of the probability space, the term-structure model described by $p(t, T) = \frac{E_t^P[D(T)]}{D(t)}$ is arbitrage free, and if the diffusion process is also a supermartingale, then the short-rate process $\{r(t)\}_{t \geq 0}$ is positive with probability P one.

MODELING IN PRACTICE

One popular way of turning theory into practice is to use a *tree* approach to modeling. The

Table 1 Market Spot and Forward Rates

Time (Months)	Implied Spot Zero Rates	Implied Forward Rates
6	5.0000%	5.0000%
12	5.1266%	5.2533%
18	5.2544%	5.5103%
24	5.3835%	5.7714%
30	5.5141%	6.0371%
36	5.6462%	6.3080%

tree can be either binomial or trinomial in its construction. To illustrate the idea, consider first the binomial approach. The tree could be set up to reflect observed or estimated market short rates, and the data provided in Table 1 will help to demonstrate this idea.

The process starts from the first six-month period where the rate is known to be 5.000%. At the end of the six-month period, the following six-month forward rates are treated as being the short rates and are split, allowing interest rates to rise with a probability of 0.5 or fall with a probability of 0.5, but also taking into account the short-rate volatility. For a description of how this is achieved, see Eales (2000). Figure 1 shows how the rates would appear in a binomial tree once the procedure has been performed.

When the rates have been established, they must then be calibrated. The calibration procedure is achieved using the observed market price of a bullet government bond and pricing the bond using the “tree” calculated rates to obtain the appropriate discount fac-

tors. Consider a three-year-to-maturity government bond trading at par and offering a coupon of 5.625% paid semiannually as an example. On maturity, the bond will be redeemed for 102.8125, which is made up of the bond’s face value, say 100, and one half of the annual coupon, 2.8125.

Figure 2 illustrates how, moving back through the tree, the discounting process of the terminal payment taken together with the discounted interim coupons generate a bond price of 100.013. Given that the observed bond price is 100, the rates in the tree will need to be adjusted to ensure that the backward calculated price agrees with the market price of the bond. In this example the adjustment factor is 0.6 basis points, and this will be added to every node in the tree with the exception of the starting value. The resulting rates will then be as displayed in Figure 3.

The calibrated tree can now be used to calculate corporate bond spreads as well as bond options. The outlined procedure is close to that advanced by Black, Derman, and Toy in that the process fits observed market rates and short-rate volatility. There is, however, a danger that interest rates could go negative in this procedure.

As an alternative to this binomial approach, Hull and White (1994) have suggested a two-stage methodology that uses a mean-reverting process with the short rate as the source of uncertainty and calculated in a trinomial tree framework. The first stage in the approach

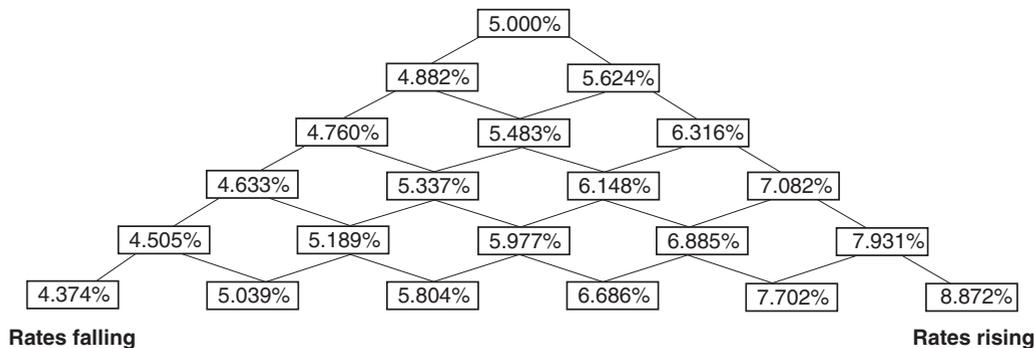


Figure 1 Term Structure Evolution: Binomial Tree

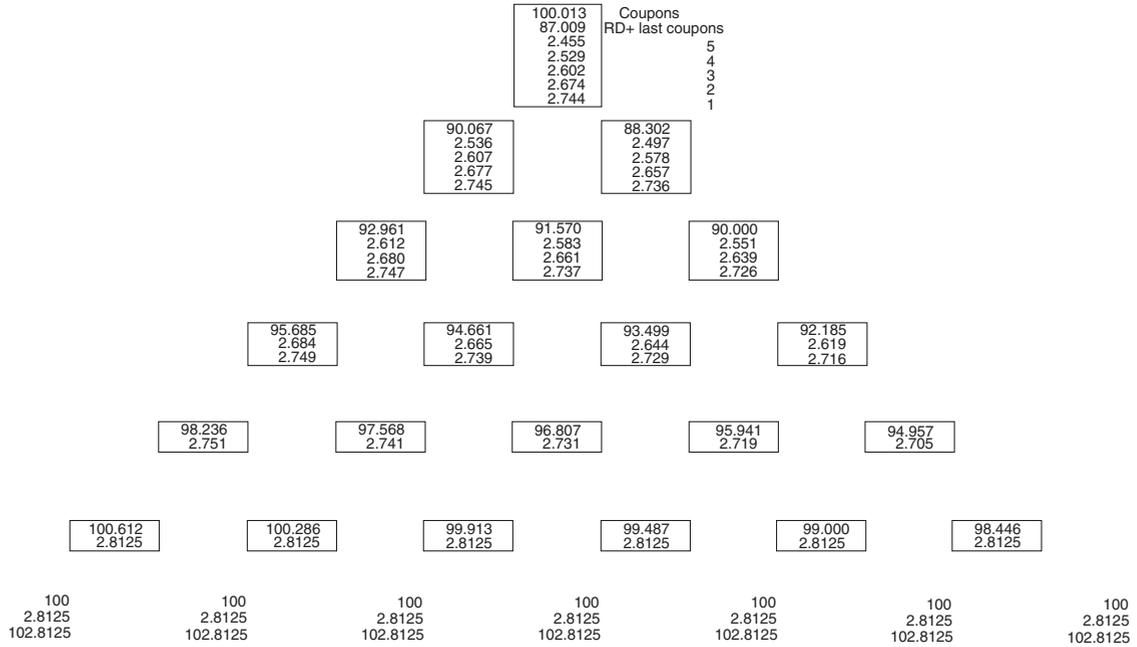


Figure 2 Calibration

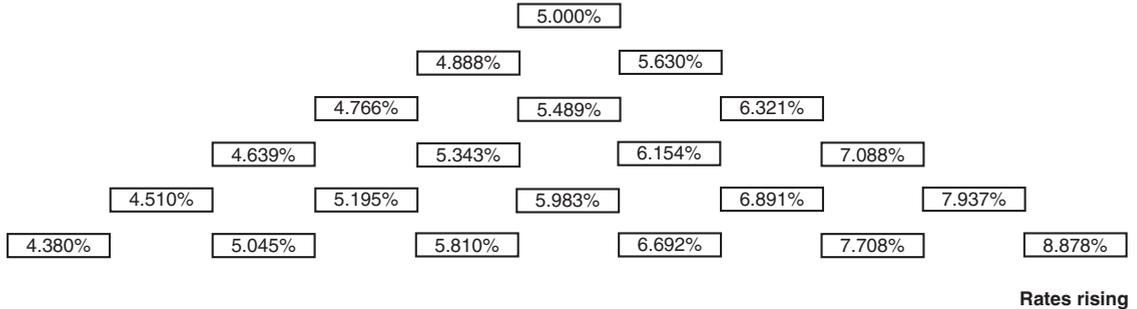


Figure 3 Adjusted Tree to Coincide with Current Market Price

ignores the observed market rates and centers the evolution of rates around zero and identifies the point at which the mean-reversion process takes effect. The second stage introduces the observed market rates into the framework established in stage one. The trinomial approach gives the tree a great deal more flexibility over its binomial counterpart, not least in relaxing the assumption that rates can either rise or fall with probability 0.5.

HJM METHODOLOGY

Heath, Jarrow, and Morton (1990a, 1990b, 1992) derived both one-factor and multifactor models for movements of the forward rates of interest. The models were complex enough to match the current observable term structure of forward rate and by equivalence the *spot rates*. Ritchken and Sankarasubramanian (1995) provide necessary and sufficient conditions for the HJM

models with one source of error and two state variables such that the *ex post* forward premium and the integrated variance factor are sufficient statistics for the construction of the entire term structure at any future point in time.

Under this methodology, the bond dynamics are described by an Ito process:

$$dp(t, T) = r(t)p(t, T)dt + \sigma(t, T)p(t, T)dW(t) \tag{29}$$

Then

$$d \ln p(t, T) = \left[r(t) - \frac{1}{2}\sigma^2(t, T) \right] dt + \sigma(t, T)dW(t) \tag{30}$$

The equation for the forward rate can be derived now:

$$\begin{aligned} df(t, T) &= -d \left(\frac{\partial}{\partial T} \ln p(t, T) \right) = - \left(\frac{\partial}{\partial T} d \ln p(t, T) \right) \\ &= - \frac{\partial}{\partial T} \left\{ \left[r(t) - \frac{1}{2}\sigma^2(t, T) \right] dt + \sigma(t, T)dW(t) \right\} \\ &= \sigma(t, T) \frac{\partial \sigma(t, T)}{\partial T} dt - \frac{\partial \sigma(t, T)}{\partial T} dW(t) \end{aligned} \tag{31}$$

The Wiener process $W = \{W(t)\}$ is symmetric, and therefore we can safely replace W with $-W$, so

$$df(t, T) = \sigma(t, T) \frac{\partial \sigma(t, T)}{\partial T} dt + \frac{\partial \sigma(t, T)}{\partial T} dW(t) \tag{32}$$

Applying the fundamental theorem of calculus for $\partial \sigma(t, T)/\partial T$ leads to

$$\sigma(t, T) - \sigma(t, t) = \int_t^T \frac{\partial \sigma(t, s)}{\partial s} ds \tag{33}$$

It is obvious that $\sigma(t, t) = 0$ and therefore the volatility of the forward rate determines the drift as well. In other words, all that is needed for the HJM methodology is the volatility of the bond prices. The short rates are easily calculated from the forward rates. Once a model for short rates is determined under the risk-neutral measure Q , the bond prices are calculated from

$$p(t, T) = E^Q \left[e^{-\int_t^T r(s)ds} | F_t \right] \tag{34}$$

Using (3) it follows that $\int_t^T f(t, s)ds = -\ln p(t, T) = g(r(t), t, T)$ where

$$g(x, t, T) = -\ln E^Q \left[e^{-\int_t^T r(s)ds} | r(t) = x \right] \tag{35}$$

The continuous variant of the Ho-Lee model can be obtained for

$$\begin{aligned} g(x, t, T) &= x(T-t) - \frac{1}{6}\sigma^2(T-t)^3 \\ &\quad + \int_t^T (T-s)\alpha(s)ds \end{aligned} \tag{36}$$

where $\sigma(t, T) = \sigma(T-t)$, which implies that $\frac{\partial f(t, T)}{\partial t} = \sigma^2(T-t)dt + \sigma dW(t)$ so the initial forward curve is

$$\begin{aligned} f(0, T) &= \frac{\partial g(r(0), 0, T)}{\partial T} = r(0) - \frac{1}{2}\sigma^2 T^2 \\ &\quad + \int_0^T \alpha(s)ds \end{aligned} \tag{37}$$

The short rate is given by

$$r(t) = f(0, t) + \sigma^2 \frac{t^2}{2} + \sigma W(t) \tag{38}$$

and the price of the pure discount bond with maturity T is

$$\begin{aligned} p(t, T) &= \exp \left[- \int_t^T f(t, s)ds - \sigma^2 t \right. \\ &\quad \left. \int_t^T \left(s - \frac{t}{2} \right) ds - \sigma(T-t)W(t) \right] \end{aligned} \tag{39}$$

Similarly, the Vasicek model is recovered for

$$\begin{aligned} \sigma(t, T) &= \sigma e^{-\beta(T-t)} \text{ and } f(0, T) \\ &= \frac{\alpha}{\beta} + e^{-\beta T} \left(r(0) - \frac{\alpha}{\beta} \right) - \frac{\sigma^2}{2\beta^2} (1 - e^{-\beta T})^2 \end{aligned} \tag{40}$$

and this leads to

$$r(t) = \frac{\alpha}{\beta} + e^{-\beta t} \left(r(0) - \frac{\alpha}{\beta} \right) + \sigma e^{-\beta t} \int_0^t e^{\beta s} dW(s) \tag{41}$$

BOND OPTION PRICING

Formulae for bond options were found by Cox, Ingersoll, and Ross using the CIR model (square root process) for short rates and by Jamshidian (1989), Rabinovitch (1989), and Chaplin (1987) using the Vasicek model for the short-rate process. Rabinovitch advocated the idea that the bond follows a lognormal process (similar to equity prices). Chen (1991) pointed out that this assumption is grossly misleading since the bond price is a contingent claim on the same interest rate, so the *bond option pricing* model cannot be a two-factor model as proposed by Rabinovitch and it rather collapses onto a one-factor model, in which case the formulas are the same with those proved respectively by Chaplin (1987) and by Jamshidian (1989).

Bonds are traded generally over the counter. Futures contracts on bonds may be more liquid and may remove some of the modeling difficulties generated by the known value at maturity of the bonds. Hedging may be more efficient in this context using the futures contracts on pure discount bonds (provided they are liquid) rather than the bonds themselves. Chen (1992) provides closed-form solutions for futures and European futures options on pure discount bonds, under the Vasicek model.

Hull and White used a two-factor version of the Vasicek model to price discount bond options. Turnbull and Milne (1991) proposed a general equilibrium model outside the *HJM framework*. They provide analytical solutions for European options on Treasury bills, interest rate forward and futures contracts, and Treasury bonds. In addition, a closed formula is identified for a call option written on an interest rate cap. A two-factor model is also investigated, and closed-form solutions are provided for a European call on a Treasury bill. Chen and Scott (1992) use a two-factor CIR model that is essentially the same as the model analyzed by Longstaff and Schwartz (1992), and derive solutions for bond and interest rate options. The two-factor model is used, with the first factor having a strong mean reversion, explaining the

variation in short-term rates, while the second factor has a very slow mean reversion, modeling long-term rates. The model is also used for calculating premiums for caps on floating interest rates and for European options on discount bonds, coupon bonds, coupon bond futures, and Eurodollar futures. These are not closed-form solutions, but they are expressed as multivariate integrals. However, the calculus can be reduced to univariate numerical integrations.

European Options on the Money Fund

In this section we consider the pricing of a European option on the money fund (this is the same as a bank account when the initial value $B(0)=1$). Thus, the payoff of a European call option with exercise price K is $\max[B(T) - K, 0]$. The continuous version of the Ho-Lee model is assumed for the short interest rate process. The risk-neutral valuation methodology provides the solution as

$$\begin{aligned} c_{B(0),T,K} &= E^Q \left[e^{-\int_0^T r(u)du} \max[B(T) - K, 0] \right] \\ &= B(0)N(d_+) - p(0, T)KN(d_-) \quad (42) \end{aligned}$$

where

$$\begin{aligned} d_+ &= \frac{\ln\left(\frac{B(0)}{p(0,T)K}\right) + \sigma^2 \frac{T^3}{6}}{\sigma \sqrt{\frac{T^3}{3}}} \quad \text{and} \\ d_- &= \frac{\ln\left(\frac{B(0)}{p(0,T)K}\right) - \sigma^2 \frac{T^3}{6}}{\sigma \sqrt{\frac{T^3}{3}}} = d_+ - \sigma \sqrt{\frac{T^3}{3}} \end{aligned}$$

A proof of this formula is described in Epps (2000) in Section 10.2.2.

Options on Discount Bonds

Discount bond options are not very liquid, but they form an elementary component for pricing other options. For example, a floating rate cap can be decomposed into a portfolio of European puts on discount bonds. Similarly, with the European option contingent to the bank account,

we can price European options contingent on discount bonds.

When the short rate process $r = \{r(t)\}$ follows the continuous time version of the Ho-Lee model given above by (10), the price at time 0 of a European call option with maturity T_0 with exercise price K on a discount bond maturing at T ($T_0 < T$) is

$$\begin{aligned}
 c_{p(0,T);T_0;K} &= E^Q \left[e^{-\int_0^T r(u)du} \max[p(T_0, T) - K, 0] \right] \\
 &= p(0, T_0) \frac{p(0, T)}{p(0, T_0)} N(d_+) \\
 &\quad - p(0, T_0) K N(d_-) \tag{43}
 \end{aligned}$$

where

$$\begin{aligned}
 d_+ &= \frac{\ln\left(\frac{p(0,T)}{p(0,T_0)K}\right) + \sigma^2 \frac{(T-T_0)T_0}{2}}{\sigma \sqrt{(T-T_0)T_0}} \quad \text{and} \\
 d_- &= d_+ - \sigma \sqrt{(T-T_0)T_0}
 \end{aligned}$$

A proof of this result is provided in Epps (2000). There is a similar put-call parity for European options contingent on a discount bond. If $p_{p(0,T);T_0;K}$ is the price at $t = 0$ of a European put option on the discount bond with maturity T , then for $B(0) = 1$,

$$\begin{aligned}
 c_{p(0,T);T_0;K} - p_{p(0,T);T_0;K} &= E^Q [e^{-\int_0^{T_0} r(s)ds} (\max[p(T_0, T) - K, 0] \\
 &\quad - \max[K - p(T_0, T), 0])] \\
 &= E^Q [e^{-\int_0^{T_0} r(s)ds} [p(T_0, T) - K]] \\
 &= E^Q [e^{-\int_0^T r(s)ds}] - p(0, T_0)K \\
 &= p(0, T) - p(0, T_0)K
 \end{aligned}$$

Put-call parity can be used to derive the price of a European put option:

$$\begin{aligned}
 p_{p(0,T);T_0;K} &= p(0, T_0) K N(-d_-) \\
 &\quad - p(0, T_0) \frac{p(0, T)}{p(0, T_0)} N(-d_+) \tag{44}
 \end{aligned}$$

Initially, the first formulas on pricing options on pure discount bonds used the Vasicek model for the term structure of interest rates. Thus, given that r follows (6), the price of a European

call option with maturity T_0 with exercise price K on a discount bond maturing at T ($T_0 < T$) is

$$c_{p(0,T);T_0;K} = p(0, T)N(d_+) - Kp(0, T_0)N(d_-) \tag{45}$$

where $d_+ = \frac{\ln\left(\frac{p(0,T)}{Kp(0,T_0)}\right) + \eta^2/2}{\eta}$ and $d_- = d_+ - \eta$, with $\eta = \frac{\sigma(1-e^{-\beta(T-T_0)})}{\beta} \sqrt{\frac{1-e^{-2\beta T_0}}{2\beta}}$

The put price can be obtained from put-call parity as

$$p_{p(0,T);T_0;K} = Kp(0, T_0)N(-d_-) - p(0, T)N(d_+) \tag{46}$$

Example: Valuing a Zero-Coupon Bond Call Option with the Vasicek Model

Let's consider this model for pricing a 3-year European call option on a 10-year zero-coupon bond with face value \$1 and exercise price K equal to \$0.5. As in Jackson and Staunton (2001), we use for the parameters of this model the values estimated by Chan et al. (1992) for U.S. 1-month Treasury bill yield from 1964 to 1989. Thus, $\alpha = 0.0154$, $\beta = 0.1779$, and $\sigma = 2\%$. In addition, the value of the short rate r at time $t = 0$ is needed, so we take $r_0 = 3.75\%$. Feeding this information into the above formulas, we get the output in Table 2. Thus, the value of the European call option is

$$\begin{aligned}
 c_{p(0,T);T_0;K} &= 0.5406 \times 0.9822 - 0.5 \times 0.8655 \\
 &\quad \times 0.9767 \cong 0.108
 \end{aligned}$$

A more general case is discussed by Shiryaev (1999) for single-factor Gaussian models modeling the short interest rate. These are single-factor affine models where the short rate r is also a Gauss-Markov process. The equation for this short rate process is

$$dr(t) = [\alpha(t) - \beta(t)r(t)]dt + \sigma(t)dW(t) \tag{47}$$

Table 2 Calculations of Elements for Pricing a European Call Option on a Zero-Coupon Bond When Short Rates Are Following the Vasicek Model

$p(0,T_0)$	$p(0,T)$	d_+	d_-	$N(d_+)$	$N(d_-)$
0.8655	0.5406	2.1013	1.9926	0.9822	0.9767

and we can easily recognize the first *Hull-White model*. The price of a European call option is also

$$c_{p(0,T);T_0;K} = p(0, T)N(d_+) - Kp(0, T_0)N(d_-) \quad (48)$$

but where

$$d_+ = \frac{\ln\left(\frac{p(0,T)}{Kp(0,T_0)}\right) + \frac{1}{2}\eta^2(T_0, T)B^2(T_0, T)}{\eta(T_0, T)B(T_0, T)} \quad \text{and}$$

$$d_- = d_+ - \eta$$

with

$$B(T_0, T) = \int_{T_0}^T \frac{\varphi(s)}{\varphi(T_0)} ds \quad \text{and} \quad \varphi(s) = e^{-\int_0^s \beta(u) du}$$

The price of the European put option is obviously again $p_{p(0,T);T_0;K} = Kp(0, T_0)N(-d_-) - p(0, T)N(d_+)$.

Example: Valuing a Zero-Coupon Bond Call Option with the Hull-White Model

When considering the pricing of a forward pure discount bond earlier in this entry, we used a numerical example. That example can now be expanded to demonstrate how, in practice, European calls and puts can be estimated in a Hull-White framework. Explicitly, the illustration will demonstrate the pricing of a one-year European call option on a four-year-to-maturity discount bond with a strike price set equal to the forward price of the bond (0.8858...).

Breaking down (d_+) into its component parts and evaluating each individually yields:

$$\begin{aligned} \ln\left(\frac{p(0, T)}{K(p(0, T_0))}\right) &= \ln\left(\frac{0.8521}{(0.8858)(0.9607)}\right) \\ &= 0, \quad B(T_0, T) = 2.5918 \\ \eta &= \frac{\sigma(1 - e^{-\beta(T-T_0)})}{\beta} \sqrt{\frac{1 - e^{-2\beta T_0}}{2\beta}} \\ &= \frac{0.02(1 - e^{-0.1(3)})}{0.1} \sqrt{\frac{1 - e^{-2(0.1)(1)}}{2(0.1)}} = 0.0493 \end{aligned}$$

The expression for (d_+) reduces to

$$\frac{\eta(T_0, T)B(T_0, T)}{2} = \frac{(0.0493)(2.5918)}{2} = 0.6395$$

The expression for d_- is (d_-) = (d_+) - η = 0.6395 - 0.0493 = 0.0146. $N(d_+)$ is found to be 0.5255 and $N(d_-)$ = 0.5058. Substituting these

results into the call option formula gives a premium of

$$\begin{aligned} c_{p(0,T);T_0;K} &= (0.8521)(0.5255) \\ &\quad - (0.8858)(0.9608)(0.5058) \\ &= 0.01730 \end{aligned}$$

or 1.73%.

One notable exception from this general class is the CIR model. There is a closed formula for this case, too. Following Clewlow and Strickland (1998), the price at time 0 of a European pure discount bond option is

$$\begin{aligned} c_{p(0,T);T_0;K} &= p(0, T)\chi^2\left(2\delta[\phi + \psi + B(T_0, T)]; 2\omega, \right. \\ &\quad \left. \frac{2\phi^2 r(0)e^{\theta T_0}}{\phi + \psi + B(T_0, T)}\right) \\ &\quad - Kp(0, T_0)\chi^2\left(2\delta[\phi + \psi]; 2\omega, \frac{2\phi^2 r(0)e^{\theta T_0}}{\phi + \psi}\right) \end{aligned} \quad (49)$$

where

$$\begin{aligned} \theta &= \sqrt{\beta^2 + 2\sigma^2}, \quad \phi = \frac{2\theta}{\sigma^2(e^{-\theta T} - 1)}, \\ \psi &= \frac{\beta + \theta}{\sigma^2}, \quad \lambda = \frac{\beta + \theta}{2}, \quad \omega = \frac{2\beta}{\sigma^2}, \\ B(t, s) &= \left(\frac{e^{\theta(s-t)} - 1}{\lambda(e^{\theta(s-t)} - 1) + \theta}\right), \\ \delta &= \frac{\omega(\lambda T + \ln \theta - \ln[\lambda(e^{\theta T} - 1) + \theta]) - \ln(K)}{B(T_0, T)} \end{aligned}$$

and $\chi^2(.; a, b)$ is the noncentral chi-squared density with a degrees of freedom and noncentral parameter b .

Example: Valuing a Zero-Coupon Bond Call Option with the CIR Model

Let's consider the same problem as described in the example using the Vasicek model above and price the 3-year European call option on a 10-year pure discount bond using the CIR model for the short interest rates. Recall that face value is \$1 and exercise price K is equal to \$0.5. As in the example with the Vasicek model, we consider that $\sigma = 2\%$ and $r_0 = 3.75\%$. The CIR model overcomes the problem of negative

interest rates known for the Vasicek model as long as $2\alpha \geq \sigma^2$. This is true, for example, if we take $\alpha = 0.0189$ and $\beta = 0.24$. Feeding this information into the above formulas is relatively tedious. A spreadsheet application is provided by Jackson and Staunton. After some work, we get that the price of the call is

$$c_{p(0,T);T_0;K} = 0.5324 \times 1 - 0.5 \times 0.8624 \times 1 \cong 0.1012$$

Options on Coupon-Paying Bonds

When short rates are modeled with single-factor models, Jamshidian (1989) proved that an option on a coupon bond can be priced by valuing a portfolio of options on discount bonds. This approach does not work in multifactor models as proved by El Karoui and Rochet (1995).

Consider a bond paying a periodic cash payment ρ at times T_1, T_2, \dots, T_m , and the principal at maturity $T = T_m$. A coupon bond can be mapped into a portfolio of discount bonds with corresponding maturities (under one source of uncertainty, that is, one factor model). The value of a coupon-bearing bond at time $t < T_m$ is

$$p(t, T_1, \dots, T_m; \rho) = \rho \sum_{i=i[t]}^m p(t, T_i) + p(t, T_m) \tag{50}$$

where $i[t] = \min\{j : t < T_j\}$.

Under the one-factor HJM model corresponding to the Ho-Lee model, a European option on a coupon bond can be valued as a portfolio of options contingent on zero discount bonds with maturities T_1, T_2, \dots, T_m . Let T_0 be the maturity of such a European option.

Epps (2000) shows that

$$p(T_0, T_i) = \frac{p(0, T_i)}{p(0, T_0)} e^{\left[-\sigma^2 \frac{(T_i - T_0)^2 T_0}{2} - (T_i - T_0)(r(T_0) - f(0, T_0)) \right]} \tag{51}$$

For any strike price K , there is a value r_K of $r(T_0)$ such that when replaced in (48) with $t = T_0$, implies $p(T_0, T_1, \dots, T_m) = K$. Let's de-

note by K_i the value of $p(T_0, T_i)$ as calculated from (49) with r_K instead of $r(T_0)$. Then

$$\rho \sum_{i=i[T_0]}^m K_i + K_m = K \tag{52}$$

Hence, the value at time 0 of a European call option with maturity T_0 and strike price K on the coupon-bearing bond, under the one-factor HJM model described above, is given by

$$\begin{aligned} c_{p(0,T_1,\dots,T_m;\rho)} &= E^Q \left\{ e^{-\int_0^{T_0} r(s)ds} \max[p(T_0, T_1, \dots, T_m; \rho) - K, 0] \right\} \\ &= \rho \sum_{i=i[T_0]}^K E^Q \left\{ e^{-\int_0^{T_0} r(s)ds} \max[p(T_0, T_i) - K_i, 0] \right\} \\ &\quad + E^Q \left\{ e^{-\int_0^{T_0} r(s)ds} \max[p(T_0, T_m) - K_m, 0] \right\} \\ &= \rho \sum_{i=i[T_0]}^m c_{p(0,T_i);T_0,K_i} + c_{p(0,T_m);T_0;K_m} \end{aligned} \tag{53}$$

Example: Valuing a Coupon-Bond Call Option with the Vasicek Model

The above example is reconsidered using the Vasicek model for the short-term interest rates. The bond is no longer a zero-bond but now pays an annual coupon at a 5% rate ($\rho = 0.05$), all the other characteristics being the same as before. In order to calculate the European call option price on the coupon bond, we need to calculate the interest rate r_K such that the present value at the maturity of the option of all later cash flows on the bond equals the strike price. This is done by trial and error using (48), and the value we get here is $r_K = 22.30\%$. Next, we map the strike price into a series of strike prices via (50) that are then associated with coupon payments considered as zero-coupon bonds and calculate the value of the European call options contingent on those zero-coupon bonds as in the preceding example. The calculations are described in Table 3.

Because we started with a one-factor model for the short interest rates, we can use the decomposition property emphasized by Jamshidian (1997) and calculate the required coupon-bond European call price as the sum

Table 3 Calculations Using the Vasicek Model for Separate Zero-Coupon European Call Options (bond prices shown are calculated with the estimated r_K)

Year	$p(T_0, T_i) r_K$	Face Value	ρK_i	Call Option
4.0	0.8094	0.05	0.0405	0.006
5.0	0.6688	0.05	0.0334	0.009
6.0	0.5624	0.05	0.0281	0.012
7.0	0.4800	0.05	0.0240	0.013
8.0	0.4148	0.05	0.0207	0.013
9.0	0.3622	0.05	0.0181	0.013
10.0	0.3192	1.05	0.3351	0.278

of all the elements in the last column in Table 3, which includes the coupon rate factor ρ . Thus, the value of this option is 0.344.

Example: Valuing a Coupon-Bond Call Option with the CIR Model

We repeat the calculation of the coupon-bond call option when the CIR model is employed for the short rates. The procedure is the same as in the case discussed previously for the Vasicek model. First, we calculate the interest rate r_K such that the present value at the maturity of the option of all later cash flows on the bond equals the strike price. This value is here $r_K = 25.05\%$. Next, we map the strike price into a series of strike prices via (50) that are then associated with coupon payments considered as zero-coupon bonds and calculate the value of the European call options contingent to those zero-coupon bonds. The calculations are described in Table 4.

Table 4 Calculations Using the CIR Model for Separate Zero-Coupon European Call Options (bond prices shown are calculated with the estimated r_K)

Year	$p(T_0, T_i) r_K$	Face Value	ρK_i	Call Option
4.0	0.7934	0.05	0.0397	0.006
5.0	0.6503	0.05	0.0325	0.010
6.0	0.5470	0.05	0.0273	0.012
7.0	0.4694	0.05	0.0235	0.013
8.0	0.4094	0.05	0.0205	0.013
9.0	0.3615	0.05	0.0181	0.013
10.0	0.3223	1.05	0.3385	0.267

The value of the call option is 0.334, that is, the sum of all zero-coupon bond call option prices in the last column.

Pricing Swaptions

Swaptions options allow the buyer to obtain at a future time one position in a swap contract. It is elementary that an interest rate swap, fixed for floating, can be understood as a portfolio of bonds. Let's consider here that the notional principal is 1. Then the claim on the fixed payments is the same as a bond paying coupons with the rate ρ and no principal. Let τ be the time when the swap is conceived. The claim on the fixed income stream is worth, at time τ , $\rho \sum_{i=1}^m p(\tau, T_i)$. The floating income stream is made up of cash returns on holding, over the period $[T_{i-1}, T_i]$ a discount bond with maturity T_i , which is worth $\frac{p(T_i, T_i)}{p(T_{i-1}, T_i)} - 1$. Thus, the value of the whole floating stream at time $t = \tau$ is

$$\begin{aligned} E_\tau \left(\sum_{i=1}^m e^{-\int_\tau^{T_i} r(s)ds} \frac{1 - p(T_{i-1}, T_i)}{p(T_{i-1}, T_i)} \right) \\ = E_\tau \left(\sum_{i=1}^m e^{-\int_\tau^{T_{i-1}} r(s)ds} e^{-\int_{T_{i-1}}^{T_i} r(s)ds} \frac{1 - p(T_{i-1}, T_i)}{p(T_{i-1}, T_i)} \right) \end{aligned} \quad (54)$$

Applying the properties of conditional expectations it follows that the above is equal to

$$\begin{aligned} E_\tau \left\{ \sum_{i=1}^m e^{-\int_\tau^{T_{i-1}} r(s)ds} \left(E_{T_{i-1}} e^{-\int_{T_{i-1}}^{T_i} r(s)ds} \left[\frac{1 - p(T_{i-1}, T_i)}{p(T_{i-1}, T_i)} \right] \right) \right\} \\ = E_\tau \left\{ \sum_{i=1}^m e^{-\int_\tau^{T_{i-1}} r(s)ds} (1 - p(T_{i-1}, T_i)) \right\} \\ = \sum_{i=1}^m [p(\tau, T_{i-1}) - p(\tau, T_i)] = 1 - p(\tau, T_m) \end{aligned} \quad (55)$$

Imposing the condition that the two streams have equal initial value leads to

$$\rho \sum_{i=1}^m p(\tau, T_i) = 1 - p(\tau, T_m)$$

which is equivalent to

$$\rho \sum_{i=1}^m p(\tau, T_i) + p(\tau, T_m) - 1 = 0$$

It follows then that the value of the swap at initialization is $p(\tau, T_1, \dots, T_m) - 1$. Thus, the option to get a long position in the fixed leg of the swap, with a fixed payment rate ρ , is worth at time 0

$$E_0^Q \left\{ e^{-\int_0^\tau r(s)ds} \max [p(\tau, T_1, \dots, T_m) - 1, 0] \right\} \quad (56)$$

It is clear now that this is the same as a European call option on a coupon-bearing bond when the exercise price is equal to 1.

PRACTICAL CONSIDERATIONS

As mentioned in the introduction, the 10-year U.S. Treasury note option traded on the CME is an extremely popular contract offering tight bid/ask spreads and transparent price quotes.

The Eurodollar futures option traded on the CME is the most actively traded short-term interest rate option in the world. If the option contracts are exercised, the buyer and the seller of the option take positions in the Eurodollar futures contract, which is cash-settled, and the final price at delivery is equal to 100 minus the three-month US dollar LIBOR.

Another liquid interest rate derivative market is the OTC in floating rate caps. The majority of caps are contingent on LIBOR (but can be also on a Treasury rate), and discounted payments are made at the beginning of each tenor. The payments can be made either at the beginning or the end of each reset period, and the life of a cap may be only a few years or as long as 10 years. The starting point in pricing these European options is a model for future changes in US dollar LIBOR.

Hull and White (1990) showed that the cap can be priced as a portfolio of European puts on discount bonds.

KEY POINTS

- One-factor short-rate models for interest rate derivatives are easy to work with since the majority of them lead to closed-form solutions for options pricing. However, some of them allow for negative interest rates, which may not be acceptable in a real trading environment.
- Two-factor models for interest rates provide improved calibration at the expense of computational simplicity.
- The two-factor Hull and White model, falling under the general Heath-Jarrow-Morton framework, is complex enough to calibrate market data easily while retaining computational simplicity through closed-form solutions for a wide range of interest rate derivatives.
- The need for improved calibration of forward curves led to the development of a different class of models called LIBOR models.
- The calibration of caps and floors, and also swaptions, is indicative of the success of an interest rate model.

REFERENCES

- Black, F., Derman, E., and Toy, W. (1990). A one-factor model of interest rate and its application to Treasury bond options. *Financial Analysts Journal* 46: 33–39.
- Black, F., and Karasinski, P. (1991). Bond and option pricing when short rates are lognormal. *Financial Analysts Journal* 47: 52–59.
- Brennan, M. J., and Schwartz, E. (1979). A continuous time approach to the pricing of bonds. *Journal of Banking and Finance* 3: 133–155.
- Brace, A., Gatarek, D., and Musiela, M. (1997). The market model of interest rate dynamics. *Mathematical Finance* 7: 127–155.
- Brennan, M. J., and Schwartz, E. (1982). An equilibrium model of bond prices and a test of market efficiency. *Journal of Financial and Quantitative Analysis* 17: 301–329.
- Brigo, D., and Mercurio, F. (2007). *Interest Rate Models: Theory and Practice*, 2nd edition. Berlin: Springer.

- Campbell, J. (1986). A defense of traditional hypotheses about the term structure of interest rates. *Journal of Finance* 41, 1: 183–194.
- Chen, L. (1995). A three-factor model of the term structure of interest rates. Preprint, Federal Reserve Board, Washington, July.
- Chen, R. R. (1991). Pricing stock and bond options when the default-free rate is stochastic: A comment. *Journal of Financial and Quantitative Analysis* 26, 3: 433–434.
- Chen, R. R. (1992). Exact solutions for futures and European futures options on pure discount bonds. *Journal of Financial and Quantitative Analysis* 27, 1: 97–107.
- Chen, R. R., and Scott, L. (1992). Pricing interest rate options in a two-factor Cox-Ingersoll-Ross model of the term structure. *Review of Financial Studies* 5, 4: 613–636.
- Chaplin, G. (1987). A formula for bond option values under an Ornstein-Uhlenbeck model for the spot. Actuarial Science working paper No. 87–16, University of Waterloo.
- Clelow, L., and Strickland, C. (1998). *Implementing Derivatives Models*. Chichester, UK: John Wiley & Sons.
- Constantinides, G. M. (1992). A theory of the nominal term structure of interest rates. *Review of Financial Studies* 5, 4: 531–552.
- Courtadon, G. (1982). The pricing of default-free bonds. *Journal of Financial and Quantitative Analysis* 17: 75–100.
- Cox, J. C., Ingersoll, J. E., and Ross, S. A. (1980). An analysis of variable rate loan contracts. *Journal of Finance* 35: 389–403.
- Cox, J. C., Ingersoll, J. E., and Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica* 53, 2: 385–407.
- Dothan, M. U. (1978). On the term structure of interest rates. *Journal of Financial Economics* 6: 9–69.
- Eales, B. A. (2000). *Financial Engineering*. New York: Palgrave Macmillan.
- Epps, T. W. (2000). *Pricing Derivative Securities*. Singapore: World Scientific.
- El Karoui, N., and Rochet, J. C. (1995). A price formula for options on coupon bonds. SEEDS Discussion Series, Instituto de Economica Publica, Spain.
- Flesaker, B., and Hughston, L. P. (1996). Positive interest. *Risk* 9, 1: 115–124.
- Fong, H. G., and Vasicek, O. A. (1992). Interest rate volatility and a stochastic factor. Working paper, Gifford Fong Associates.
- Heath, D., Jarrow, R., and Morton, A. (1990a). Bond pricing and the term structure of interest rates: A discrete time approximation. *Journal of Financial and Quantitative Analysis* 25: 419–440.
- Heath, D., Jarrow, R., and Morton, A. (1990b). Contingent claim valuation with a random evolution of interest rates. *Review of Futures Markets* 9: 54–76.
- Heath, D., Jarrow, R., and Morton, A. (1992). Bond pricing and the term structure of interest rates. *Econometrica* 60, 1: 77–105.
- Ho, T., and Lee, S. (1986). Term structure movements and pricing interest rates contingent claims. *Journal of Finance* 41: 1011–1029.
- Hull, J. C. (2003). *Options, Futures, and Other Derivatives*. Upper Saddle River, NJ: Prentice Hall.
- Hull, J. C., and White, A. (1990). Pricing interest rate derivative securities. *Review of Financial Studies* 3, 5: 573–592.
- Hull, J. C., and White, A. (1994). Numerical procedures for implementing term structure models I: Single-factor models. *Journal of Derivatives* 2, 1: 7–16.
- Hull, J. C., and White, A. (1996). Using Hull-White interest rate trees. *Journal of Derivatives*, Spring: 26–36.
- Jamshidian, F. (1989). An exact bond option formula. *Journal of Finance* 44, 1: 205–209.
- Jamshidian, F. (1997). LIBOR and swap market models and measures. *Finance and Stochastics* 1: 293–330.
- Longstaff, F. A., and Schwartz, E. S. (1992a). Interest rate volatility and the term structure: A two-factor general equilibrium model. *Journal of Finance* 47: 1259–1282.
- Longstaff, F. A., and Schwartz, E. S. (1992b). A two-factor interest rate model and contingent claim valuation. *Journal of Fixed Income* 3: 16–23.
- Mercurio, F., and Moraleda, J. M. (2000). An analytically tractable interest rate model with humped volatility. *European Journal of Operational Research* 120: 205–214.
- Merton, R. C. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4, Spring: 141–183.
- Musiela, M., and Rutkowski, M. (1997). *Martingale Methods in Financial Modelling*. Berlin: Springer.
- Rabinovitch, R. (1989). Pricing stock and bond options when the default-free rate is stochastic: A comment. *Journal of Financial and Quantitative Analysis* 24, 4: 447–457.
- Rebonato, R. (1998) *Interest-Rate Option Models*, 2nd Edition. Chichester, UK: John Wiley & Sons.

- Rendleman, R., and Bartter, B. (1980). The pricing of options on debt securities. *Journal of Financial and Quantitative Analysis* 15: 11–24.
- Rogers, L.C.G. (1997). The potential approach to the term structure of interest rates and foreign exchange rates. *Mathematical Finance* 7: 157–176.
- Ritchken, P., and Sankarasubramanian, L. (1995). Volatility structures of forward rates and the dynamics of the term structure. *Mathematical Finance* 5: 55–72.
- Rutkowski, M. (1997). Models of forward LIBOR and swap rates. Preprint, University of New South Wales.
- Sandmann, K., and Sondermann, D. (1993). A term structure model and the pricing of interest rate derivatives. *Review of Futures Markets* 12, 2: 391–423.
- Schmidt, W. M. (1997). On a general class of one-factor models for the term structure of interest rates. *Finance and Stochastics* 1: 3–24.
- Strickland, C. R. (1992). The delivery option in bond futures contracts: An empirical analysis of the LIFFE long gilt futures contract. *Review of Futures Markets* 11: 84–102.
- Shiryaev, A. N. (1999). *Essentials of Stochastic Finance: Facts, Models, Theory*. Singapore: World Scientific.
- Turnbull, S. M., and Milne, F. (1991). A simple approach to interest-rate option pricing. *Review of Financial Studies*. 4, 1: 87–120.
- Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics* 5: 177–188.

Basics of Currency Option Pricing Models

SHANI SHAMAH

Senior Consultant, RBC Capital Markets

Abstract: Historically, theorists have devoted a substantial amount of work developing a mathematical model for pricing options and, hence, a number of different models exist as a result. All make certain assumptions about market behavior, which are not totally accurate, but which give the best solution to the price of an option. Professionals use these models to price their own options and to give theoretical fair value; however, actual market rates will always be the overriding determinant. In other words, an option is worth as much as someone is prepared to pay for it. Although the formula for pricing options is complex, they are all based on the same principles.

Historically, option-pricing models have fallen into two categories:

- Ad hoc models, which generally rely only upon empirical observation or curve fitting and, therefore, need not reflect any of the price restrictions imposed by economic equilibrium.
- Equilibrium models, which deduce option prices as the result of maximizing behavior on the part of market participants.

The acknowledged basis of modern option pricing formulas is the often-quoted Black-Scholes formula, devised by Black and Scholes (1973) to produce a “fair value” for options on equities. Of course, currency options differ because there is no dividend and both elements of the exchange carry *interest rates* that can be fixed until maturity. Therefore, various adaptations to the original Black-Scholes formula have been made for use in currency option pricing. The

best known of these is the Garman-Kohlhagen adaptation, which adequately allows for the two interest rates and the fact that a currency can trade at a premium or at a discount forward depending on the interest rate differential.

American-style options cause further problems in the pricing due to the probability of early exercise. Cox, Ross, and Rubinstein (1979) introduced a pricing model to take account of American-style options. By using the same basics as Black-Scholes, they adopted what is now known as the “binomial” method for pricing such options. This same binomial model is now used alongside the Garman-Kohlhagen version to price currency options.

BASIC PROPERTIES

First, though, there are a few basic properties of options, especially when looking at option prices to consider:

- Options cannot have a negative value to their holders. Since options are rights and these rights will be exercised to benefit only the holder, the option cannot be a liability to its holder.
- Option prices should not allow simple arbitrage; that is, it should not be possible to buy an American call or put and immediately exercise the option for a profit greater than the price paid for the option. This need not be true for European options, since the option holder does not have the right to exercise until the maturity date.
- American-type options should be worth at least as much as European-type options. Since *American* options have all the rights a *European* option has plus the right of early exercise, an American option will be as valuable as a European option if the right to early exercise is worthless and more valuable than a European option if the right of early exercise is valuable.

In addition to the currency price, the exercise price, and the time to maturity, option values depend on the price volatility of the underlying currency, the risk-free rate of interest, and any cash distributions made by the currency during the life of the option. For a call option, a higher current currency price should imply a greater value to the option holder. This is because a higher present currency price makes it more likely that on the expiration date, the market price of the currency will be above the exercise price. As this is precisely the condition under which the option will be exercised, the value of a call option increases as the present currency price increases. For put options, however, the effects of changes in the current asset price go in the opposite direction, as it pays the holder of the put to exercise when the currency price is low; that is, the value of a put option decreases as the present currency price increases.

The effect of the exercise price, X , on the value of the call option is straightforward. Holding all

other factors constant, a higher exercise price diminishes the profit from the exercise of the option. An increase in the exercise price would, therefore, lead to a decrease in the price of the call option. In the case of put options, a higher exercise price increases the profit from exercise of the option. Thus, put option prices increase with an increase in their exercise price.

The effect of an increase in time to maturity on the value of an option depends on the nature and type of option. There is an asymmetric nature to option contracts that causes the holder to benefit from increased uncertainty. The option holder stands to gain by a rise in uncertainty, and therefore the value of the call option increases as its time to maturity increases. Also, the present value of the exercise price decreases as the time to maturity increases. Therefore, the time left to maturity has a way of influencing option values. An American put option cannot logically decrease in value with an increased time to maturity, but with a European put option, the net effect of these two influences is ambiguous; that is, increased uncertainty increases value, while the decreased present value of the exercise price decreases value.

An increase in the volatility of the currency price makes future currency prices more variable and increases the probability of large gains. Again, the asymmetry of the option contract allows the option holder to benefit from increased uncertainty since the option is effectively insured against downside risk.

THEORETICAL VALUATION

The price and subsequent value of an option are determined by a theoretical valuation based on several known and estimated factors. The time until maturity, the current foreign exchange spot and forward exchange prices, the strike, and the cost of funding the *option premium* are all readily available. Meanwhile, a market has developed that estimates the future volatility

or, in other terms, the activity of the underlying cash product. The greater the anticipated movement, the greater the value of the option for a given fixed set of parameters. Options also increase in value the smaller the distance between the *strike price* and the forward foreign exchange rate, and the greater the time to maturity. For European and American options, most market participants accept the valuation put forward by Black and Scholes, and, as such, option prices can be agreed once the factors are entered into the equation.

This theoretical model also calculates the risk associated with changes in any of the variables required for pricing the currency option. The *delta*, or hedge ratio, of the option is the degree to which the option value will change with a movement in the underlying currency. A dollar/Swiss franc option with a 20% delta would change in value by approximately 20 franc points for every 100-point spot move. While the delta is the first derivative of the price, *gamma* is the second one, or change in delta for every move in the spot foreign exchange rate. A 50-delta dollar call option with a 15% gamma would have a 65 delta if the dollar appreciated 1%.

It is this dynamic nature of the delta that allows an option to be a leveraged product with limited risk and unlimited profit potential. Profitable positions effectively grow in size, while unprofitable trades are impacted less by adverse changes in the market.

The *vega*, or volatility risk, of an option is the extent to which the valuation will change with varying estimates of volatility. The *theta*, or time decay, is the decrease in value of the option as it approaches maturity, as an option is a constantly diminishing asset. Finally, every option has forward foreign exchange risk equivalent to the delta and an interest rate exposure based on changes in funding costs. The delta and interest rate risks can be hedged easily in the relevant markets. The dynamic nature of the other risks is the essence of the options market.

BLACK-SCHOLES MODEL

In 1973, Black and Scholes published a paper describing an equilibrium model of stock option pricing that is based on arbitrage. This is made possible by their crucial insight that it is possible to replicate the payoff to options by following a prescribed investment strategy involving the underlying asset and lending/borrowing.

The mathematics employed in the *Black-Scholes model* is complex, but the principle is straightforward. The model states that the stock and the call option on the stock are two comparable investments. Therefore, it should be possible to create a riskless portfolio by buying the stock and hedging it by selling call options. The hedge is a dynamic one because the stock and the option will not necessarily move by the same amount, but by continuously adjusting the option hedge to compensate for movement in the underlying market, the overall position should be riskless. Therefore, the income received from investing in the call option premium will be offset exactly by the cost of replicating (hedging) the position. If the option premium is too high, the arbitrageur will make a riskless profit by writing call options and hedging the underlying stock. If too low, it should be possible to profit by buying the call option and selling sufficient stock.

Black and Scholes demonstrated that the option premium could be arrived at through an arbitrage process in a similar manner to that in which a currency forward rate can be derived through a formula linking the spot rate and the interest rate differential. Also, in the same way that a currency *forward rate* is not “what the market thinks the currency will be worth at a future date” but simply based on an arbitrage relationship, the Black-Scholes model is not influenced by such factors as market sentiment, direction, or apparent bias. In fact, an assumption of the model is that the market moves in a random fashion in that, while prices will change, the chances of an up move against a down move are

about even, and that future price movements cannot be predicted from the behavior of the past.

The Model:

$$C = SN(d_1) - Ke^{(-rt)}N(d_2)$$

C = theoretical call premium

S = current stock price

t = time until option expiration

K = option striking price

r = risk-free interest rate

N = cumulative standard normal distribution

e = exponential term (2.7183)

$$d_1 = \frac{\ln(S/K) + \left(r + \frac{s^2}{2}\right)t}{s\sqrt{t}}$$

$$d_2 = d_1 - s\sqrt{t}$$

s = standard deviation of stock returns

ln = natural logarithm

Plotted over a period of time, the distribution of prices takes on the characteristics of the “bell-shaped” curve. Such a distribution is a key assumption of the Black-Scholes model, yet with the foreign exchange markets in particular, it is a questionable one. Even with its economic liquidity and its global 24-hour structure, foreign exchange is by no means a perfect market. Frequently, there are times when prices do not behave in a normally distributed fashion. Such occurrences as wars, central bank intervention, and unexpected political or economic news are all factors, which can and do disrupt the day-to-day business of the market.

Furthermore, in order to simplify the calculation process, Black and Scholes made other assumptions about market behavior, which may vary from the real world. They assumed that volatility was known and constant, that interest rates were constant, that there were no transaction costs or taxation effects, that trading was continuous, that there were no dividends payable, and that options could only be exercised on the expiry date.

Interest rates will vary, of course, as will volatility, and even the foreign exchange markets have transaction cost in the bid-offer spread. Frequently, the market will become very thin or almost untradable during highly volatile periods. However, most of these assumptions can be relaxed without inordinately affecting the formulations of the pricing model, and where the assumptions are more critical, other models have been developed.

EXAMPLES OF OTHER MODELS

Theorists have devoted a substantial amount of time and effort developing mathematical models for pricing options, and a number of different models exist as a result. All make certain assumptions about market behavior, which are not totally accurate, but which give the best solutions to the price of an option. For example, the model formulated by *Merton* (1973) generalized the Black-Scholes formula, so it could price European options on stocks or stock indexes paying a known dividend yield.

Another example is the *Cox, Ross, and Rubinstein model* (1979), which could account for the early exercise provisions in American-style options. Using the same parameters as in the Black-Scholes model, they adopted what is known as a “binomial” method to evaluate the premium. Making the assumption that the option market behaves efficiently and therefore the holder of a call or put option will exercise if the benefit of holding the option is outweighed by the cost of carrying the hedge, the binomial process involves taking a series of trial estimates over the life of the option; each estimate (or iteration) is a probability analysis of the likelihood of early exercise on any given day.

Garman and Kohlhagen (1983) extended the Black-Scholes model to cover the foreign exchange market, where they allowed for the fact that currency pricing involves two interest rates, not one, and that a currency can trade

at a premium or discount forward, depending on the interest rate differential. Like the Merton formula, the *Garman and Kohlhagen* formula applies only to European options.

PRICING WITHOUT A COMPUTER MODEL

Against all the above theories, there is a way to price an option without a computer model. This can be obtained by the following equation, which will give a good approximation for a European option premium. The formula is:

$$\text{Price} = \text{BB} \times \text{forward outright rate}$$

This is where:

$$\text{AA} = \text{square root} (\text{days to expiry}/365) \\ \times \text{volatility} \times 0.19947$$

and

$$\text{BB} = ((\text{AA} + 0.5) \times 2) - 1$$

This formula is where price is the premium for an at-the-money European option quoted in units per base currency.

Educated Guess

Another calculation relies heavily on probability theory. The principal concepts are expected value and the lognormal distribution. Since the future is unknown, it is an “educated guess” about where the spot market might be in order to determine the value of that right today. Thus, rather than trying to predict the future spot rate, option pricing takes a systematic, mathematical approach to the educated guess.

In this case, expected value (EV) is the payoff of an event multiplied by the probability of it occurring. For example, the probability of rolling a six on one die is 1/6 or 16.67%. The EV of a game in which is paid \$100 for rolling a six and nothing for any other roll is:

$$(1/6 \times \$100) + (5/6 \times \$0) = \$16.67$$

where the expected value is the fair price for playing such a game.

An options premium can be thought of in the same way, although instead of six possible outcomes, there are hundreds. All the spot rates that might prevail are the options expirations. Each outcome will have a specific value. This will either be zero if the option is out-of-the-money or the difference between the closing spot and the strike price if the option is in-the-money. Each closing spot rate can also be thought of as having its own discrete probability. If, for each outcome, the value of that outcome is multiplied by its probability and then the results are added up, the sum would be the premium of the option. The expected value of an option (the probability minus the weighted sum of all its possible payoffs) is the fair price for buying the option.

THE PRICE OF AN OPTION

The price of an option is made up of two separate components:

$$\text{Option premium} = \text{Intrinsic value} + \text{Time value}$$

where *intrinsic value* is the value of an option relative to the outright forward market price, that is, it represents the difference between the strike price of the option and the forward rate at which one could transact today. Intrinsic value can be zero but never negative.

There are six factors that contribute to this pricing of an option:

- Prevailing spot price
- Interest rate differentials (forward rate)
- Strike price
- Time to expiry
- Volatility
- Intrinsic value

As described above, the best-known original closed-form solution to option pricing is the Black-Scholes model. Also, as was mentioned, in its simplest form, it offers a solution to

pricing European-style options on assets with interim cash payouts over the life of the option. The model calculates the theoretical, or fair value for the option by constructing an instantaneously riskless hedge that is one whose performance is the mirror image of the option payout. The portfolio of option and hedge can then be assumed to earn the risk-free rate of return.

Central to the model is the assumption that markets' returns are normally distributed (that is, have lognormal prices), that there are no transaction costs, that volatility and interest rates remain constant throughout the life of the option, and that the market follows a diffusion process. The model has these five major inputs:

- The risk-free interest rate
- The option's strike price
- The price of the underlying
- The option's maturity
- The volatility assumed

Since the first four are usually determined, markets tend to trade the implied volatility of the option. For example, a six-month European-style sterling put/dollar call with the spot rate at $\$/\text{£}1.7500$ and forward points of 515, giving an outright forward of 1.6985 ($1.7500 - 0.0515$), will have an intrinsic value of 4.15 cents per pound.

While the Black-Scholes pricing formula looks formidable, it is important to understand that the formula is nothing more than the simple two-state option-pricing model applied with an instantaneous trading interval.

If the strike price of the option is more favorable than the current forward price, the option is said to be in-the-money. If the strike price is equal to the forward rate, it is an at-the-money option, and if the strike price is less favorable than the outright, the option is termed out-of-the-money.

For American-style options, a similar definition applies except that the option's "money-ness" relative to the spot price also needs to be

considered. Clearly, in the example above, an American-style option would be in-the-money relative to the forward but not to the spot. Conversely, if the option had the same details except that it was a call on sterling, it would clearly be out-of-the-money under the European definition, but as an American style option it would be in-the-money relative to the spot price. Naturally, the cost of the option would need to be considered in order to achieve a profitable early exercise of an American option and this leads to a phenomenon peculiar to American-style options known as "optimal exercise." This is the point at which it becomes profitable to exercise an American-style option early, having taken account of the premium paid.

Option Premium Profile

Figure 1 shows premium against spot at a given point in time. It can be seen that the time value call position is greatest when the option is at-the-money. This is because it represents the highest level of asymmetric risk, which is optimum risk reward profile.

The time value tends to zero as spot goes deep out-of-the-money and thus converges with the maximum loss expiry line and also as it goes deep in-the-money, converging with the unlimited profit expiry line. The change in the premium is not parallel to the change in the underlying value. The premium will change more rapidly when the option is near at-the-money and less rapidly when the option is in-the-money or out-of-the-money.

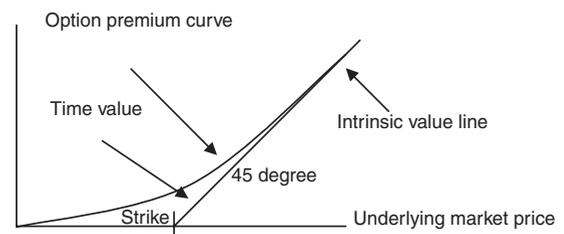


Figure 1 Option Premium Profile

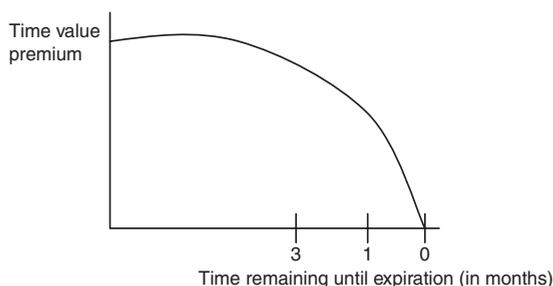


Figure 2 Time Value Premium Delay

Time Value and Intrinsic Value

The option premium can be split into two parts: intrinsic value and time value. The effect of an increase in time on an options premium is not linear. This is because the probability of a rise or fall in a currency's value does not increase on a straight-line basis. For example, all things being equal, the premium for an at-the-money three-month option is worth only about two-thirds more than for a one-month option (not three times its value). A one-year option is worth only about one-third more than a six-month option (instead of twice its value). As a consequence, the premium for at-the-money options declines at an accelerating rate towards expiry. Figure 2 demonstrates the time value premium delay.

Time value is affected by a number of factors:

- The time remaining to expiration.
- The volatility of the underlying spot market.
- The strike price of the option.
- The forward rate of the currency pair.
- The current interest rates.

Time to Expiry

The time decay of an option is related to the time remaining in the option; in fact, it is proportional to the square root of the time remaining. The reason for this phenomenon is twofold:

1. The longer the time to maturity, the greater is the chance that the exchange rate moves such that the option will be exercised. The rate at which the premium diminishes as the option approaches expiry is called the "time decay"

and the rate of decay is exponential, that is, the option loses time value more quickly approaching expiry than it does earlier in its life. At expiry, the option will have only intrinsic value and no time value.

2. The time value can be thought of as "risk premium" or the cost to the writer of hedging the uncertainty of exercise.

Volatility

In essence, *volatility* is a measure of the variability (but not the direction) of the price of the underlying instrument, essentially the chances of an option's being exercised. It is defined as the annualized standard deviation of the natural log of the ratio of two successive prices.

Historical volatility is a measure of the standard deviation of the underlying instrument over a past period and is calculated from actual price movements by looking at intraday price changes and comparing this with the average (the standard deviation). The calculation is not affected by the absolute exchange rates, merely the change in price involved. Thus, for example, the starting and finishing points for two separate calculations could be exactly the same but could give two very different levels of volatility depending on how the exchange rate traded in between. Thus, if the market has traded up and down erratically, the reading will be high, and if instead it has gradually moved from one point to the other in even steps, then the reading will be lower.

Implied volatility is the volatility implied in the price of an option, that is, the volatility that is used to calculate an option price. Implied volatilities rise and fall with market forces and tend to reflect the level of activity anticipated in the future although supply and demand can at times be dominant factors. In the professional interbank market, two-way volatility prices are traded according to market perception and these volatilities are converted into premium using option models. Implied volatility is the only variable affecting the price of an

option that cannot be directly observed in the markets, thus leading to the typical variations in price inherent in any marketplace.

Actual volatility is the actual volatility that occurs during the life of an option. It is the difference between the actual volatility experienced during delta hedging and the implied volatility used to price an option at the outset, which determines if a trader makes or loses money on that option.

In summary, implied volatility is a timely measure, in that it reflects the market's perceptions today. On the other hand, historical volatility is a retrospective measure of volatility. This implies that it reflects how volatile the variable has been in the recent past. But it has to be remembered that it is a highly objective measure. Implied volatilities can be biased, especially if they are based upon options that are traded in a market with very little liquidity. Also, historical volatility can be calculated for any variable for which historical data is tracked.

Volatility affects the time value or risk premium of an option, as an increase in volatility increases the time value and thus the price of the option. Likewise, a decrease in volatility lowers the price of the option. For example, consider the position of the writer of an option, whereby, say, a bank sells an option to a client, giving the client the right to purchase dollars and sell Swiss francs in three months' time. In order to correctly hedge the position, consider what will happen in three months' time.

If the spot is above the strike price of the option, the client will exercise the option and the bank will be obliged to sell dollars and buy francs. However, if the spot is below the strike price, the client will allow the option to lapse. Hence, the bank's initial hedge for the option will be to purchase a proportion of dollars in the spot market against this potential short dollar position. If the spot subsequently rises, the likelihood of the option's being exercised will increase and so the initial hedge will be too small. Therefore, the bank will need to buy some more dollars, which it does at a rate worse than the

original rate at which the option was priced, thereby losing money. Conversely, if the spot rate falls, this makes the option less likely to be exercised and the bank will then find itself holding too many dollars and will have to sell them out at a lower price than where they were purchased, thus losing more money. These losses are called "hedging costs," and each time the spot market moves, the rehedging required will lose the bank money. In essence, the premium received by the writer is effectively the best estimate of these hedging costs over the life of the option.

Strike Price and Forward Rates

An option's time value is greatest when the strike price is at-the-money and the further in or out-of-the-money the strike price is, the lower the time value is. This can be explained by again considering the hedging costs. If the option is originally at-the-money, it is said to have a 50 delta and therefore the initial hedge will be to buy or sell half the principle amount of the option. The delta of the option can be thought of as the probability of exercise and so a 50 delta gives a 1-in-2 chance of exercise, that is, maximum uncertainty. As the spot moves, the delta will change and require readjusting of the hedge in the spot market. The change in delta (or gamma) is greater for a 50-delta option than for an option with a much higher or lower delta, for example 80- or 20-delta. This is because the likelihood of exercise, and therefore the amount of hedge required, changes more rapidly. Thus, less readjustment is required for these high and low delta options, and consequently, fewer hedging costs are incurred for the low and high delta options. This leads to lower levels of risk premium or time value for in-the-money and out-of-the-money options.

Interest Rates

The currency interest rate is another factor that affects option premiums. As premium is usually paid up front, it must be discounted to take

account of the interest that would be earned by putting the premium on deposit. Thus, the higher the domestic interest rate, the greater the discounting effect on the premium.

The effect of interest rate differential on the option premium is not intuitively obvious, yet it is one of the most important components of the premium for a currency option. If the dollar interest rate rises in relation to the interest rate of the foreign currency, the premium of a currency call option will increase in value. This is because holding a foreign currency and buying a currency call option are alternative investments. On the one hand, the investor will sell (borrow) dollars and buy (invest in) a foreign currency in order to take advantage of a rise in that foreign currency. On the other hand, the trader could just simply buy a currency call option. If the dollar interest rate rises, the cost of borrowing dollars will increase, which will make the alternative of buying a currency call option more attractive. Consequently, the premium will rise.

This can equally be explained in terms of the forward value of a currency. If the dollar interest rate rises in relation to the foreign currency interest rates, and the spot rate remains the same (unchanged), covered interest rate arbitrage will ensure that the forward rate of the foreign currency will rise relative to the spot. Therefore, the call option on that currency will also rise in value. Of course, the dollar interest rate might remain the same, but the interest rate of the foreign currency might fall. The effect on the interest rate differential and therefore on the value of the currency call option will remain the same, but the premium will rise.

The converse is true for currency put options, because an increase in the dollar interest rate in relation to the foreign currency interest rate will, given no change in the spot price, result in a rise in the forward value of the currency. Thus, the holder of a put option on the currency will see the premium fall. Buying a currency put option is an alternative strategy to borrowing in that currency and investing in dollars. Hence, a

rise in the dollar interest rate or a fall in the foreign currency interest rate makes the put option strategy less attractive, and the put premium will fall.

The effect of interest rate differential changes on currency option premiums can be summarized as follows:

- Assuming the spot rate remains unchanged, a rise in dollar interest rates relative to the foreign currency interest rate, or a fall in the foreign currency interest rate relative to the dollar interest rate, will increase the premium for a currency call option and decrease the premium for a currency put option.
- Assuming the spot rate remains the same, a fall in the dollar interest rate relative to the foreign currency interest rate, or a rise in the foreign currency interest rate relative to the dollar interest rate, will decrease the premium for a currency call option and increase the premium for a currency put option.

American versus European

For European options, intrinsic value is the value of an option relative to the outright forward market price; that is, it represents the difference between the strike price of the option and the forward rate at which one could transact today. Intrinsic value can be zero but is never negative. If the strike price of the option is more favorable than the current forward price, the option is in-the-money. If the strike price is equal to the forward rate, the option is at-the-money and if the strike price is less favorable than the outright forward, the option is out-of-the-money.

A similar definition applies for American-style options, except that the option's "moneyness" relative to the spot price also needs to be considered. Naturally, the cost of the option needs to be considered in order to achieve a profitable early exercise and this leads to a phenomenon peculiar to American options known as optimal exercise. This is the point at which it becomes profitable to exercise an

American option early, having taken account of the premium paid.

In fact, there are several occasions when it would be better to pay extra premium and buy a more expensive American-style option. For example:

1. When a trader is buying an option where the call currency has the higher interest rate and there is an expectation that the interest rate differential will widen significantly.
2. When a trader is buying an option where the interest rates are close to each other and there is an expectation that the call interest rate will move above the put interest rate.
3. When a trader is buying an out-of-the-money option with interest rates as in both of the above and there is an expectation for it to move significantly into the money, then the American-style option is more highly leveraged and will hence produce higher profits.

THE GREEKS

Traders extensively use *the "Greeks,"* a set of factor sensitivities, to quantify the exposure of portfolios that contain options. Each measures how the portfolio's market value should respond to a change in some variable. For speculative purposes, the value of an option needs to be known on a continual basis, and more importantly, the factors that change an option's value need to be understood. In analyzing an option risk (or value), the market norm is to use letters of the Greek alphabet. Not surprisingly, they are often referred to as the "Greeks," and they include delta, vega/kappa, theta, gamma, and rho. However, vega is not in the Greek alphabet, but is named after a star in the constellation Lyra. Sometimes, vega has also been referred to as kappa. Also, four of the five are risk metrics. The exception here is theta, because the passage of time is certain and thus entails no risk.

These major Greeks, which measure these risks and need to be taken into account before taking any option positions, are:

Vega/Kappa	Theta	Delta	Gamma	Rho
Measures the impact of a change in volatility	Measures the impact of a change in time remaining	Measures the impact of a change in the price of the underlying	Measures the rate of change in delta	Measures the sensitivity to an applicable interest rate

Delta

When option traders sell or buy a currency option, they will use the foreign exchange market to hedge the exposure. The most common type of hedging is delta hedging.

Delta is the change in premium per change in the underlying. Technically, the underlying is the forward outright rate but as the option-pricing model assumes constant interest rates, this is often calculated using spot. For example, if an option has a delta of 25 and spot moved 100 basis points, then the option price gain/loss would be 25 basis points. For this reason, delta is sometimes thought of as representing the "spot-sensitive" amount of the option.

Also, delta can be thought of as the estimated probability of exercise of the option. As the option-pricing model assumes an outcome profile based around the forward outright rate, an at-the-money option has a delta of 50%. It falls for out-of-the-money options and increases for in-the-money options, but the change is non-linear, in that it changes much faster when the option is close-to-the-money.

Turning to calculus for the formal definition of delta, let t be the current time. Let 0p and 0s be current values for the portfolio and underlier. Delta is the first partial derivative of a portfolio's value with respect to the value of the underlier:

$$\text{delta} = \frac{\partial {}^0p}{\partial {}^0s}$$

This technical definition leads to an approximation for the behavior of a portfolio.

$$\Delta^0 p \approx \text{delta} \Delta^0 s$$

where $\Delta^0 s$ is a small change in the underlier's current value, and $\Delta^0 p$ is the corresponding change in the portfolio's current value. This is called the delta approximation.

An option is said to be delta hedged if a position has been taken in the underlying in proportion to its delta. For example, if one is short a call option on an underlying with a face value of \$1 million and a delta of 0.25, a long position of \$250,000 in the underlying will leave one delta neutral with no exposure to changes in the price of the underlying, but only if these are infinitesimally small.

As the underlying market moves throughout the life of the option, the delta will change, thus requiring the underlying hedge to be adjusted. Once the initial hedge has been transacted, calls and puts behave in precisely the same way, in terms of the hedging required.

For example, an at-the-money sterling call/dollar put option in £10 million, with a strike price of 1.75, has an initial delta of 50. The option writer, therefore, buys £5 million in the spot market to hedge the option position. If the spot rises to 1.77, the delta will increase to, say, 60. Now, the writer needs to purchase an extra £1 million to attain delta neutrality. If the exchange rate then falls back again to the original rate, the option writer is overhedged and requires selling back £1 million in order to remain delta neutral. Clearly, as the option writer rehedges, losses will be incurred, which will increase as volatility increases.

Another example could be where a trader sells a dollar call/Swiss franc put at 1.5500 for six months for \$10 million. The trader's risk is that in six months, the option will be exercised and there will be a payout of dollars and a receipt of francs. The trader's hedge against this risk would therefore be to buy dollars and sell francs, thus hedging the delta amount because this represents the likelihood of exercise. If spot

is 1.5300 and the forward outright is 1.5345, then the trader's hedging, ignoring time movement, would look like that shown in the following table, as the forward rate changes:

Forward	Delta	Hedge	Total
1.5345	35	Buy \$3.5 million	+\$3.5 million
1.5500	50	Buy \$1.5 million	+\$5.0 million
1.5600	57	Buy \$0.7 million	+\$5.7 million
1.5200	30	Sell \$2.7 million	+\$3.0 million

Whether or not the trader loses money will depend on volatility. From the preceding table, it can be seen that hedging a short option position loses money, as the trader would be continually buying high and selling low. However, when the option was first sold, the trader received a premium for it, representing the estimated cost of hedging to the trader. If the volatility of the market is higher than the trader expected and then has to hedge more frequently, then the trader may lose more money hedging than originally gained on the premium. If, however, the market is less volatile than the assumption of the option price, the trader should lose less money hedging than received in premium and therefore make a profit overall.

If the trader had bought the option rather than sold it, the trader would then hope for increased volatility because the hedging exercise would be making money.

For example, the trader buys exactly the same options, a dollar call/Swiss franc put at 1.5500 in \$10 million. The trader's risk is now that there will be a long dollar position in six months, so the hedge will be to sell dollars and buy francs. As the forward outright rate moves, however, the delta of the option will move in exactly the same way as before. This follows as the option is the same and the delta does not depend on who owns the option. In this case, therefore, the trader will be buying low and selling high and making money on the hedging. Just as before, this makes sense, as the trader originally paid out a premium to buy the option, so the hedging is making back that premium. This time,

the trader has bought volatility and hopes that volatility will in fact be higher than the rate at which the option was bought for. If it is, the trader will make more money hedging than was paid out in premium.

Hence, buying and selling volatility is like any other product in that there is a wish to buy at a low rate and sell at a higher rate to make a profit.

As another example, consider a short sterling call at a 1.8100 position at 342 points. The loss profile corresponds to the loss profile on a short sterling cash position. Thus, a hedge on a short sterling call position would be to buy sterling cash. The value of the option will go up with sterling going up, but it is not a one-to-one relationship.

The delta ratio indicates the increase in value of the option for every increase in value of one point on the cash market. Thus, the following rules on delta can be established. On a call option, delta will range from 0% when out-of-the-money to 50% at-the-money to 100% when deep in-the-money. Conversely, the delta of a put option goes from 0% when out-of-the-money to -50% at-the-money to -100% when deep in-the-money.

In the preceding example, the delta of the option is, say, 45%, which means that to hedge the position, an amount of sterling of 45% of the face value of the option will have to be bought. Therefore, if the option is for £1 million, a move up of 50 points on the rate will result in a loss of:

$$£1 \text{ million} \times 0.0050 \times 45\% = \$2,250$$

This will be offset by the long cash position of:

$$450,000 \times 0.0050 = \$2,250$$

The delta of an option does not remain constant and the new delta of this position is, say, 47%. In order to maintain a delta-neutral position, the trader will have to buy another £20,000. Such a hedging strategy will enable the trader

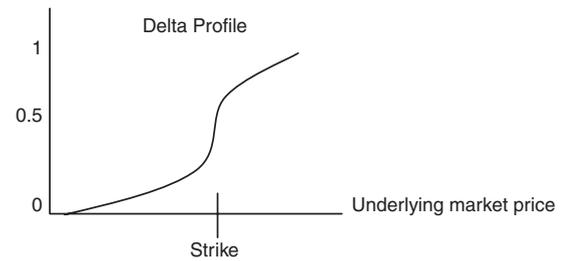


Figure 3 Delta Profile

to keep the premium received initially when selling the option.

Figure 3 shows that delta is the gradient of the tangent of the curve of the premium in relation to the cash prices. This will also reveal that delta will move more rapidly for an option with a short remaining life than for an option with a long remaining life.

In conclusion, basically, the delta of an option will change if any factor which influences the potential probability of exercise changes. These include spot price, volatility, time, and interest rates. Option trades use the delta as a guide to hedging. Taken simply, if a bank is short one option with a delta of 50%, the bank will hedge only half of the nominal amount of the option as it only has a 50% chance of being exercised. This is known as “delta hedging.” This is a simplistic example, and, in reality, banks have large option books, which they hedge on a daily basis, but the principal applies no matter what the size of the portfolio.

Also, there are three points to keep in mind with delta:

1. Delta tends to increase as it gets closer to expiration for near or at-the-money options.
2. Delta is not a constant.
3. Delta is subject to change given changes in implied volatility.

Gamma

The rate of change of delta is called gamma, and it will give a measure of the amount of change in the delta for a given change in the cash price.

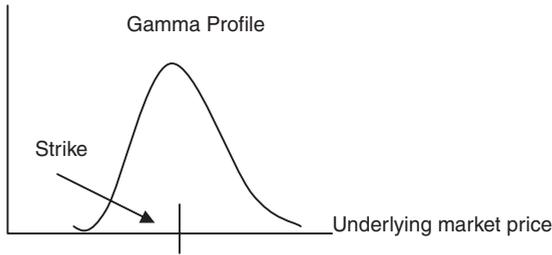


Figure 4 Gamma Profile

Therefore, it will provide an estimate of how much it will cost to delta hedge.

The cost of rebalancing the hedge is a consequence of the curvature of the premium curve against cash prices. The curvature is greatest at-the-money and reduces when in-the-money or out-of-the-money. This is shown in Figure 4.

A short option position is called gamma negative. The higher the gamma, the less stable is the delta hedge. A first conclusion is that it is more costly to hedge a short long-dated option position than a short position of short-dated options.

Thus, gamma is the change in delta per change in the underlying and is important because the option model assumes that delta hedging is performed on a continuous basis. In practice, however, this is not possible, as the market gaps and the net amounts requiring further hedging would be too small to make it worthwhile. The gapping effect that has to be dealt with in hedging an option gives the risk proportional to the gamma of the option.

For a formal definition of gamma, again turn to calculus. Gamma is the second partial derivative of a portfolio's value 0p with respect to the value 0s of the underlier:

$$\text{gamma} = \frac{\partial^2 p}{\partial^2 s}$$

By incorporating gamma, there can be an improvement to the approximation for how the portfolio's value should change in response to small changes in the underlier's value:

$$\Delta^0 p \approx \frac{\text{gamma}}{2} \Delta^0 s^2 + \text{delta} \Delta^0 s$$

This is called the delta-gamma approximation.

An option's gamma is at its greatest when an option is at-the-money and decreases as the price of the underlying moves further away from the strike price. Therefore, gamma is U-shaped and is also greater for short-term options than for long-term options.

By convention, gamma can be expressed in two ways:

1. A gamma of, say, 5.23 will mean that for 1% change in the underlying price the delta will change by 5.23 units. That is, from 50% to 55.23%.
2. A gamma of 3% will mean that for a one unit change in the underlying price, the delta will change by 3%, for example from 50% to 51.5%.

As an example of gamma hedging, as the forward outright rate moves from 1.5600 to 1.5200, the delta of the option moves from 57 to 30. The size of movement of the delta given this movement of the underlying is the gamma of the option by the definition "gamma is the change in delta per change in the underlying." The hedging the trader was required to do was to sell \$1.7 million. In practice, the trader sold the full amount at a rate of 1.5200. If the trader were able to hedge continuously as the model assumes, the trader would have sold the same amount, that is, \$1.7 million, but at an average rate of 1.5450. This would obviously have been more profitable. From this example, it can be seen that the gapping effect works against the trader when there is a short options position (and therefore short gamma), and a repetition of the exercise would show that the gapping is in the trader's favor if a long options position were being held (and gamma).

The value of gamma is, therefore, very important in determining sensitivity to spot movement and this gapping effect.

However, gamma is not the same for all options. Gamma is greater for short-term options

than for long-term options. For example, assume a dollar call/Swiss franc put option with a strike of 1.5500 and that there is one second to get to expiry. If the spot at the time is 1.5501, the option is extremely likely to be exercised and the delta will be 100. If, in that second, the spot moved to 1.5499, the option would not, in fact, be exercised and the delta would move to 0. Here, it can be seen that a 0.0002 move in spot produced a change in delta from 100 to 0. If it were the same option but there was one year to maturity, a movement of 0.0002 in spot would not significantly alter the likelihood that the option would be exercised; that is, the delta would not change noticeably.

Gamma is greater for at-the-money options than for options with deltas above or below 50. Assume an extreme example to see this effect, using the same option of a dollar call/Swiss franc put with a strike of 1.5500, and there is a second to go before expiry. If the spot is at 1.5500 and thus the option has a delta of 50, there would be the same situation as before when a 0.0001 movement in spot created a movement of 50 in the delta. If, however, the spot were at 1.5200, the delta of the option would be 0, and a movement even as large as 0.0200 would not increase that delta.

In conclusion, gamma is seen as a second-generation derivative, where the others considered are regarded as first-generation derivatives in the pricing of an option, in that the others all consider the change that an external effect has on an option's value, such as change in spot. However, gamma measures the rate of change of the delta itself. Therefore, it is literally the delta of the delta. Since the delta is the key pricing tool used by market participants in controlling the portfolio risk, to be able to work out the rate of change of this risk is very useful. Hence, gamma is a very important part of any option portfolio and is affected by three different factors: spot movement, time to maturity, and volatility.

Also, the three points to keep in mind with gamma are:

1. Gamma is smallest for deep out-of-the-money and deep in-the-money options.
2. Gamma is highest when the option gets near-the-money.
3. Gamma is positive for long options and negative for short options.

Theta

Theta is the depreciation of the time value element of the premium, that is, it measures the effect on an option's price of a one-day decrease in the time to expiration. The more the market and strike prices diverge, the less effect theta has on an option's price. Obviously, if you are the holder of an option, this effect will diminish the value of the option over time, but if you are the seller (the writer) of the option, the effect will be in your favor, as the option will cost less to purchase. Theta is nonlinear, meaning that its value accelerates as the option approaches maturity. Positive gamma is generally associated with negative theta and vice versa.

The rate at which the time value decays with respect to time is expressed as hundredths of a percent per unit of time (day/week). Obviously, the theta factor plays in favor of a short option position. Shorter-dated options have larger thetas as do those at-the-money. This effect will give rise to trading strategies referred to as a calendar spread.

To determine theta, assume t denotes time, and let ${}^t p$ denote the portfolio's value at time t . Formally, theta is the partial derivative of the portfolio's value with respect to time:

$$\text{theta} = \frac{\partial {}^t p}{\partial t}$$

where the derivative is evaluated at time $t = 0$. This technical definition leads to an approximation for the behavior of a portfolio.

$$\Delta {}^t p \approx \text{theta} \Delta t$$

where Δt is a small interval of time, and $\Delta {}^t p$ is the change in the portfolio's value that will

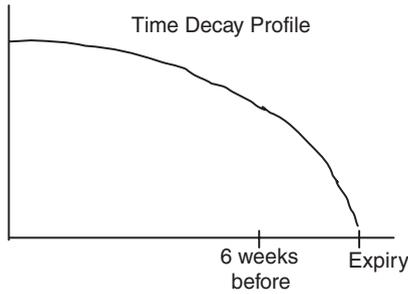


Figure 5 Time Decay Profile

occur during that interval, assuming all other market variables remain the same.

The delta of an option does have an influence on the time decay of an option because the time value element of an option total value is maximum for at-the-money options. As the delta increases or decreases, the time value of the option decreases. Obviously, for options where there is very little time value, there will be very little time decay. If there is any doubt about which date to choose for an option maturity, as can be seen in Figure 5, there is little increase in time value for days at the far end of the option. To buy a slightly longer option, therefore, will not cost much more. However, if a trader waits until the option expires and then has to buy another option to cover the final period, the additional cost could be substantially more. For this reason, buying an option for the longest period needed is recommended.

In actual practice, traders do not use theta, but it is an important conceptual dimension. However, some additional points of note are:

1. Theta can be very high for out-of-the-money options if they contain a lot of implied volatility.
2. Theta is typically highest for at-the-money options.
3. Theta will increase sharply in the last few weeks of trading and can severely undermine a long option holder's position, especially if implied volatility is on the decline at the same time.

Vega

Vega, sometimes also called kappa, quantifies risk exposure to implied volatility changes. Vega tells us approximately how much an option price will increase or decrease given an increase or decrease in the level of implied volatility. Option sellers benefit from a fall in implied volatility, while option buyers benefit from an increase in implied volatility. Vega is greatest for at-the-money options and increases with the time to maturity. This is the case because the longer the time to maturity, the greater the possibility of exchange rate movements and, therefore, the greater the sensitivity of the option price to a change in volatility.

Vega is the first partial derivative of a portfolio's value 0p with respect to the value ${}^0\sigma$ of implied volatility. This technical definition leads to an approximation for the behavior of a portfolio.

$$\Delta^0 p \approx \text{vega} \Delta^0 \sigma$$

where, here, $\Delta^0 \sigma$ is a small change in the implied volatility from its current value, and Δp is the corresponding change in the portfolio's value.

Thus, the more volatile the underlying price the more expensive the option will become because of the uncertainty element. The ratio of how much the value of the premium changes for a 1% change in volatility is vega. Longer-dated options have higher vegas and at-the-money options have higher vegas. It is expressed as a percentage change of dollars for a 1% change of volatility. For example, a vega of 1.0 means the option premium will appreciate by 1% in dollar or sterling terms.

Rho

It is generally considered to be the least important of the Greeks, but nevertheless any option, be it a single position or a large portfolio, will be exposed to such a risk. This is because with over-the-counter European-style

options, the price (in part) is derived from the forward rate. Therefore, if either of the two interest rates of the currency pair in the option should change, so the forward and hence the price will change. This can happen without a move in the spot price.

In formulating rho, let 0p and 0r be current values for the portfolio and underlier. Formally, rho is the partial derivative of the portfolio's value with respect to the risk-free rate:

$$\text{rho} = \frac{\partial {}^0p}{\partial {}^0r}$$

This technical definition leads to an approximation for the behavior of a portfolio.

$$\Delta {}^0p \approx \text{rho} \Delta {}^0r$$

where $\Delta {}^0r$ is a small change in the risk-free rate, and $\Delta {}^0p$ is the corresponding change in the portfolio's value.

In summary, rho is the general term used for interest rate risk, but it is broken down further. Rho usually refers to the base currency interest rate (usually dollars), and phi relates to the traded currency interest rates (e.g., Swiss francs or Japanese yen).

Beta and Omega

Some other Greek letters that are used do not actually measure an option's value but are more geared to looking at the use of options or risks associated with valuation methods. Briefly, they include beta and omega.

Beta represents the risk involved in hedging one currency pair against another, especially when sometimes currency pairs have a high correlation, for example, within the old European Monetary System (EMS) with the deutsch mark and the French franc. Some traders that had a dollar against the franc position would have been happier hedging this exposure in the more liquid dollar against the mark market because it fairly closely correlated to the franc. The risk here would have been if the mark against the franc correlation had started to weaken.

Omega measures the translation profit/loss risk assumed by trading in currency pairs (which result in profits/losses in those two currencies) that are not the same as the reporting base currency for accounting purposes. An example would be an American bank that gets profits for its sterling against Swiss franc trades in either sterling or francs, yet has to convert these to dollars for the balance sheet.

KEY POINTS

- The generally accepted pricing basis for options today is the Black-Scholes formula, which was devised in the early 1970s to provide a "fair value" for equity options. However, the foreign exchange markets needed something to take account of interest rates and the fact that there are no dividends due on currencies.
- Various adaptations of the Black-Scholes model emerged, of which the most popular one used today is the Garman-Kohlhagen system. This method makes allowances for the interest rates of the respective currencies and the fact that a currency can trade at a discount or premium forward relative to the other currency.
- American-style options differ due to the possibility of early exercise. The Cox-Ross-Rubenstein model is the generally accepted method for these, but they do not feature heavily in the over-the-counter market.
- Overall, the industry norm is to use the Black-Scholes formula adapted by Garman-Kohlhagen for valuing over-the-counter European-style currency options.
- The factors required to price an option include: (1) currency pair; (2) call or put; (3) strike rate; (4) amount; (5) style (European or American); (6) expiration date and time (New York expiry or Tokyo expiry); (7) prevailing spot rate; (8) interest rates for both currencies; (9) foreign exchange swap rate (calculated from the information in the

previous factor); and (10) volatility of the currency pair.

- The six factors chosen by the potential buyer/seller of the option are the currency pair, call or put, strike rate, amount, style, and expiration date and time. The prevailing spot rate, interest rates for both currencies, and foreign exchange swap rate are given by the market. The volatility of the currency pair is the only unknown factor, representing the anticipated market volatility expected for the life of the option, and is determined using the option pricing models discussed in this entry.

REFERENCES

- Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 3: 637–654.
- Cox, J., Ross, S., and Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of Financial Economics* 7: 229–263.
- Garman, M. B., and Kohlhagen, S. W. (1983). Foreign currency option values. *Journal of International Money and Finance* 2: 231–237.
- Merton, R. C. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4, 1: 141–183.

Credit Default Swap Valuation

REN-RAW CHEN, PhD

Professor of Finance, Graduate School of Business, Fordham University

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

DOMINIC O'KANE, PhD

Affiliated Professor of Finance, EDHEC Business School, Nice, France

Abstract: Credit default swaps are the most popular of all the credit derivative contracts traded. Their purpose is to provide financial protection against losses incurred following a credit event of a corporate or sovereign reference entity. Replication arguments attempting to link credit default swaps to the price of the underlying credits are generally used by the market as a first estimate for determining the price at which a credit default swap should trade. The replication argument, however, is dependent on the existence of same maturity and same seniority floating rate bonds. Even if such securities do exist, contractual differences between CDS and bonds can weaken the replication relationship. Over the past decade, the increased liquidity of the CDS market has meant that in some cases, it, and not the bond market, is the place where credit price discovery occurs. Despite this it still necessary to have a CDS valuation model for the valuation and risk-management of existing positions.

Credit default swaps (CDSs), or simply default swaps, provide an efficient credit-risk transferring financial instrument. Their over-the-counter nature also makes them infinitely customizable, thereby overcoming many of the limitations of the traditional credit market instruments such as lack of availability of instruments with the required maturity or seniority. Increasing standardization and familiarization with the legal framework has made capital market participants more willing to enter into default swap transactions as have developments in credit modeling and pricing that have made

it possible to mark-to-market and hedge default swap positions.

Bonds are the main source of liquidity in the credit markets, especially in the United States. In the early years of the CDS market, replication arguments that attempted to link CDSs to bonds were therefore generally used by the market as a first estimate for determining the price at which CDSs should trade. Nowadays, the greater liquidity of the CDS market means that it is often the place where price discovery occurs and can at times lead the cash credit bond market. So while the replication relationship is still

important, it is now a two-way process with bond traders looking at CDS prices and CDS traders looking at bond prices, all watching to see if the replication relationship breaks down to the extent that any dislocation becomes arbitrageable, at which point they will step in and enter into positions to profit from the dislocation. If done in a material size, the effect of such an action should be to realign the two markets. However, the replication argument is not exact, as it is based on a number of assumptions that often break down in practice. Market participants who wish to price CDSs and examine relative value opportunities need to understand replication and its assumptions. We discuss the replication approach in this entry.

However, replication only provides a starting point for quoting CDS spreads. It does not allow traders to actually mark to market their existing CDS positions. By definition, marking a CDS position to market must involve pricing it off the current market CDS spread curve—a set of CDS spreads quoted for different maturities. The main objective of this chapter will be to explain how to determine the CDS spread, what factors affect its pricing, and how to mark-to-market CDSs. We show that this requires a model and set out the standard model that is used by the market.

DEFAULT SWAPS

In a standard CDS contract one party pays a regular fee to another to purchase credit protection to cover the loss of the face value of an asset following a credit event. The company (or sovereign) to which the triggering of the credit event is linked is known as the reference entity.

This protection lasts until some specified maturity date which falls on the 20th of either March, June, September or December, typically five years from the trade date. To pay for this protection, the protection buyer makes a regular stream of payments. These are quoted in terms of an annualized percentage known as the CDS spread. These payments are typically paid quarterly according to an Actual 360 ba-

sis convention and are collectively known as the premium leg. Payments occur until maturity of the contract or a credit event occurs, whichever happens first. The protection buyer will also pay the protection seller the fraction of the coupon which has accrued since the previous premium payment date.

If a credit event does occur before the maturity date of the contract, there is a payment by the protection seller, known as the protection leg. There are two ways to settle the payment of the protection leg: physical settlement and cash settlement. The form of settlement is specified at the time of the ISDA organised auction used to determine the final recovery price of the deliverable obligations. This can take the form of physical or cash settlement and one of the purposes of the auction is to ensure that both have the same economic value.

- **Physical settlement:** Following the ISDA auction, a protection buyer who elects for physical settlement will submit a facevalue amount of bonds into the auction and receive a payment of 100 on the same facevalue. A protection seller who elects for physical settlement will end up receiving a deliverable obligation and paying par. In general there is a choice of deliverable obligations from which the protection buyer can choose. These deliverable obligations will satisfy a certain number of characteristics that typically include restrictions on the maturity of the deliverable obligations and the requirement that they be pari passu—most default swaps are linked to senior unsecured debt. Typically, they include both bonds and loans. If deliverable obligations trade with different prices following a credit event, which they are most likely to do if the credit event is a restructuring, the protection buyer can take advantage of this situation by buying and delivering the cheapest deliverable. The protection buyer is therefore long a cheapest to deliver (CTD) option.
- **Cash settlement:** A protection buyer who opts for cash settlement receives par minus the recovery price on his face value. The

recovery price is the one determined by the ISDA auction. The protection seller pays par minus the same recovery price.

CDS spreads are typically quoted for a variety of maturities with most liquidity at the five-year maturity followed by the three-year and seven-year maturities. The bid is the spread at which the dealer is willing to buy protection, while the offer is the spread at which the dealer is willing to sell protection. Clearly, the bid spread will be less than the offer spread. Note that this is opposite to the convention for bonds where the bid spread is the spread at which the dealer is willing to buy the bond and this will be higher than the offer spread. This is because the buyer of a bond is selling protection, while the buyer of a CDS is buying protection.

Illustration

Suppose a protection buyer purchases 5-year protection on a company at a default swap spread of 200bp. The face value of the protection is \$10 million. The protection buyer therefore makes quarterly payments approximately equal to $\$10 \text{ million} \times 0.02 \times 0.25 = \$50,000$. (The exact payment amount is a function of the calendar and basis convention used.) After a short period the reference entity suffers a credit event. Assuming that the subsequent ISDA auction which takes place within 2 months of the credit event determines a recovery price of \$35 per \$100 of face value, the payments are as follows:

- The protection seller compensates the protection seller for the loss on the face value of the asset received by the protection buyer. This is equal to $\$10 \text{ million} \times (100\% - 35\%) = \6.5 million .
- The protection buyer pays the accrued premium from the previous premium payment date to time of the credit event. For example, if the credit event occurs after a month then the protection buyer pays approximately $\$10 \text{ million} \times 0.02 \times 1/12 = \$16,666$ of premium accrued. Note that this is the standard for corporate reference entity linked default swaps.

The Mechanics of Settlement

The timeline around the physical settlement of a CDS following a *credit event* consists of three steps:

1. A CDS market participant who has previously signed up to the ISDA protocols submits a request to the ISDA determinations committee asking whether or not a credit event has occurred on a specified reference entity. The event must be evidenced by at least two sources of publicly available information (e.g., a news article on Reuters, the *Wall Street Journal*, the *Financial Times* or some other recognized publication or electronic information service). The determinations committee, which consists of both buy and sell-side representation then has to decide whether or not the credit event has occurred. An 80% supermajority is needed to approve any decision. If it is determined that a credit event has occurred, the process leading to the ISDA auction is then begun.
2. The ISDA then begins compiling a list of the deliverable obligations and publishes the details of the auction which will take place in order to determine the recovery price. If the credit event is a bankruptcy or a failure to pay then CDS contracts are automatically triggered. However if the event is a restructuring, CDS protection buyers can decide whether to trigger their contract or not – if they decide not to trigger then the contract can be used later if a bankruptcy or failure to pay occurs. In Europe, the settlement of a restructuring event is also complicated by the fact that standard CDS contracts with different maturities can have different baskets of deliverable obligations and separate auctions will be needed to determine their final recovery price for each basket.
3. The auction takes place. CDS market participants who have positions in the triggered contracts need to decide whether or not to settle physically or in cash. Buyers and sellers of CDS protection can choose physical settlement even if their trade counterparty

chose cash settlement, and vice-versa. The various dealers through whom market participants trade then bring all of these positions plus their own positions into an auction at the end of which only the net position – the net open interest – will be transferred, thereby averting any short squeeze which may be caused if the gross notional of CDS positions exceeds the outstanding notional of deliverable obligations. Dealers can then submit bids or offers on the net open interest of physical obligations, which may be long or short. At the end of this auction procedure, a recovery price is determined. All CDS contracts are then automatically settled at this recovery price.

As a result, the maximum delay between notice of a credit event and the actual payment of the protection is approximately 72 calendar days.

CREDIT EVENTS

The most important section of the documentation for a default swap is what the parties to the contract agree constitutes a credit event that will trigger a payment by the protection seller to the protection buyer. Definitions for credit events are provided by the International Swap and Derivatives Association (ISDA). First published in 1999, there have been periodic updates and revisions of these definitions. The most recent, and one of the most important updates of the ISDA documentation for credit default swaps was the introduction of the Big Bang protocol in 2009. These were a response to the Financial Crisis of 2008 and were intended to streamline the process of determining and settling a credit event. They were also intended to enable the migration of CDS trades to centralised counterparties by increasing fungibility.

ISDA Credit Event Definitions

Of the eight possible credit events referred to in the *1999 ISDA Credit Derivative Definitions*, the ones typically used within most contracts are listed in Table 1. In terms of which are used,

Table 1 Credit Events Typically Used within Most CDS Contracts

Credit Event	Description
Bankruptcy	Corporate becomes insolvent or is unable to pay its debts. The bankruptcy event is of course not relevant for sovereign issuers.
Failure to pay	Failure of the reference entity to make due payments, taking into account some grace period to prevent accidental, triggering due to administrative error.
Restructuring	Changes in the debt obligations of the reference creditor but excluding those that are not associated with credit deterioration such as a renegotiation of more favorable terms.
Obligation acceleration/ obligation default	Obligations have become due and payable earlier than they would have been due to default or similar condition. Obligations have become capable of being defined due and payable earlier than they would have been due to default or similar condition. This is the more encompassing definition and so is preferred by the protection buyer.
Repudiation/ moratorium	A reference entity or government authority rejects or challenges the validity of the obligations.

Source: ISDA.

the market distinguishes between corporate- and sovereign-linked CDSs. For corporate-linked CDSs the market standard is to use just three credit events—bankruptcy, failure to pay, and restructuring. For sovereign-linked CDSs, obligation acceleration/default and repudiation/moratorium are also included.

Restructuring Controversy

Restructuring means a waiver, deferral, restructuring, rescheduling, standstill, moratorium, exchange of obligations, or other adjustment with respect to any obligation of the reference entity such that the holders of those obligations are materially worse off from either an economic, credit, or risk perspective. It has been the most controversial credit event that may be included in a default swap.

In bankruptcy or failure to pay, *pari passu* assets trade at or close to the same recovery value. But restructuring is different. Following a restructuring, debt continues to trade. Short-dated bonds trade at higher prices than longer-dated bonds, bonds with large coupons trade at a higher price than bonds with low coupon. Loans, which are typically also deliverable, tend to trade at higher prices than bonds due to their additional covenants.

This makes the delivery option that is embedded in a default swap potentially valuable. A protection buyer hedging a short-dated high coupon asset may find that following a restructuring credit event it is trading at, say, \$80 while another longer-dated deliverable may be trading at \$65. By selling the \$80 asset, purchasing the \$65 asset, and delivering it into the CDS, the protection buyer may make a \$15 windfall gain out of the delivery option. However, this gain is made at the expense of the protection seller who has to take ownership of the \$65 asset in return for a payment of par.

Such a situation arose in the summer of 2000 when the U.S. insurer Consecro restructured its debt. At that time, the range of deliverable obligations following a restructuring event was the same as those used for bankruptcy or failure to pay. This meant that bonds or loans with a maximum maturity of 30 years could be delivered. Protection sellers were displeased at being delivered long-dated low-priced bonds in the price range 65 to 80 by banks who held much higher-priced short-term loans. In addition, it was believed that there was a conflict of interest—banks who exercised their default swaps had also been party to the restructuring of Consecro's debt.

The results of this experience led to the market discussing a restructuring supplement to the standard ISDA documentation. This was completed on May 11, 2001, and introduced a new restructuring definition called modified restructuring (*mod-re*). The essence of this was to reduce the range of deliverable obligations following a restructuring event and so limit the value of the delivery option.

Although adopted by the North American market between 2002 and 2009, this standard has now become redundant for the standard North American contract (SNAC) since restructuring is no longer one of the standard triggering credit events. Europe has retained the restructuring credit event. However the basket of allowed deliverable obligations is determined by the Modified Modified Restructuring clause which effectively limits the maturity of these obligations to the greater of the maturity of the CDS contract and 60 months. Credit default swaps linked to Asian corporate credits continue to include restructuring as a credit event. They also retain the old style rules about what can be delivered, allowing all bonds and loans of the appropriate seniority and with a maximum maturity of 30 years. A summary description of the different standard market contracts by geographical region is shown in Table 2.

Where the same credit trades with different restructuring conventions, these different contract standards should be reflected in the quoted market spreads. For example,

Table 2 Different Restructuring Standards by geographic region.

Region	Description
North America	The standard North American contract (SNAC) now trades without restructuring as a credit event.
Europe	Both CDS and CDS indices trade with bankruptcy, failure to pay and restructuring. In the case of restructuring, the deliverable obligations are determined according to the Modified Modified Restructuring clause which limits the maturity of deliverables to the maximum of the maturity of the CDS and 60 months.
Asia	Both CDS and CDS indices trade with bankruptcy, failure to pay and restructuring. Following a restructuring event the only limit on deliverables is the old-style limit of a maximum maturity of 30 years.

Source: ISDA.

modified-modified restructuring allows the protection buyer to have a broader range of deliverables than modified restructuring. This means that the value of the delivery option is greater for mod-mod-re than for mod-re and so the protection should trade at a wider spread for the more valuable delivery option. More generally, there should be a strict theoretical relationship between these spread levels of¹

$$\text{Spread}_{\text{Old-Re}} > \text{Spread}_{\text{Mod-Mod-Re}} > \text{Spread}_{\text{Mod-Re}} > \text{Spread}_{\text{No-Re}}$$

In this entry, the aim is not to determine what the spread differences should be, but to price contracts of a given type given the corresponding curve of market spreads.

Credit Events and Implementation of Default Swap Pricing Models

In the pricing model presented in this entry, we refer to “default.” By this we mean any of the credit events included in the CDS contract. This means that the value of a contract will depend on which credit events are included in a particular trade.

While the model presented handles any of the credit events that may be selected by the parties to a trade, the data required are typically drawn from databases that collect defaults defined in a different way than those set forth by ISDA credit event definitions. For example, major studies regarding default rates and recovery rates, as well as default times, define default in terms of the legal definition of default. In contrast, consider restructuring. Suppose that full restructuring is included in a trade as a credit event. Then a reduction in a reference obligation’s interest rate that is material is a credit event. In fact, actions by lenders to modify the terms of a reference obligation without a bankruptcy proceeding are not uncommon. Yet, they are not included (or even known) to researchers who compile data on defaults.

The key point is that in the implementation stage, the inputs must be modified based on the credit events included in a trade.

PRICING CREDIT DEFAULT SWAPS BY STATIC REPLICATION

There is a fundamental relationship between the *default swap* market and the cash market in the sense that a default swap can be shown as being economically equivalent to a combination of cash bonds. This cash-CDS relationship means that determination of the appropriate default swap spread for a particular credit usually begins by observing the London Interbank Offered Rate (LIBOR) spread at which bonds of that issuer trade. The usual comparison is to look at what is called the *par asset swap spread* of a bond of a similar maturity to the default swap contract. This is the spread over LIBOR paid by a package containing a fixed-rate bond and interest rate swap purchased at par. This spread can easily be calculated.²

Since 2009, CDS contracts have traded with fixed premiums. Prior to this, any new CDS contract would have its premium set at initiation so that the contract would have zero initial value. In order to facilitate moves towards a centralised counterparty for CDS, in 2009 the market decided that all contracts on a specific reference entity, regardless of their maturity and when they were traded will trade with the same fixed premium. The value of this fixed premium is different for different reference entities. In the US it is 100bp for investment grade credits and 500bp for high-yield credits. A similar convention exists in Europe with additional spreads levels. The effect of this is that CDS contracts no longer have zero value at initiation. This is actually not a radical change – it simply means that new contracts have to be valued in the same way that seasoned CDS contracts were valued in the past. However, it does mean that the old CDS-bond static replication argument becomes less realisable.

Since it is a fixed number through time, the premium spread on the CDS linked to some reference entity no longer reflects the market implied credit risk of the reference entity at the time of the trade. That information is now

embedded in the upfront cost of the CDS. But this cost is not a spread measure and is difficult to use to compare the market implied credit risk across different credits and different maturities. Instead, the market has created a new spread measure known as the par CDS spread. This is defined as the coupon on a fictional CDS which would give it a zero initial value today. It is the old CDS premium now reborn as a spread measure. The following static replication argument is therefore based on such a fictional CDS contract where the spread S is set so that the contract has zero initial value. The reason for doing this is that we wish to understand the relationship between this par spread and the par asset swap spread. Note also that the standard model which we will describe later is the mechanism used to convert the upfront cost of a CDS contract to a par spread and vice-versa.

The premium payments in a default swap contract are defined in terms of a default swap spread, S , which is paid periodically on the protected notional until maturity or a credit event. It is possible to show that the default swap spread can, to a first approximation, be proxied by a par floater bond spread (the spread to LIBOR at which the reference entity can issue a floating rate note of the same maturity at a price of par) or the asset swap spread of an asset of the same maturity provided it trades close to par.

To see this, consider a strategy in which an investor buys a par floater issued by the reference entity with maturity T . The investor can hedge the credit risk of the par floater by purchasing protection to the same maturity date. Suppose this par floater (or asset swap on a par asset) pays a coupon of LIBOR plus F . Default of the par floater triggers the default swap, as both contracts are written on the same reference entity. With this portfolio the investor is effectively holding a default-free investment, ignoring counterparty risk.

The purchase of the asset for par may be funded on balance sheet or on repo—in which case we make the assumption that the repo rate can be locked in to the bond's maturity. The resulting funding cost of the asset is LIBOR plus

B , assumed to be paid on the same dates as the default swap spread S . Consider what happens in the following scenarios:

No credit event—The hedge is unwound at the bond maturity at no cost since the protection buyer receives the par redemption from the asset and uses it to repay the borrowed par amount.

Credit event—The protection buyer delivers the reference asset to the protection seller in return for par. If we assume that the credit event occurs immediately following a coupon payment date, then the cost of closing out the funding is par, which is repaid with this principal. The position is closed out with no net cost.

Both scenarios are shown in Figure 1. As the hedged investor has no credit risk within this strategy they should not earn (or lose) any excess spread. This implies that $S = F - B$; that is, the default swap spread should be equal, to the par floater spread minus the funding cost of the cash bond. For example, suppose the par floater pays LIBOR plus 25 basis points and the protection buyer funds the asset on balance sheet at LIBOR plus 10 basis points. For the protection buyer the breakeven default swap spread equals $F - B = 25 - 10 = 15$ basis points.

This analysis certainly shows that there should be a close relationship between cash and default swap spreads. However, the argument is not exact as it relies on several assumptions that could result in small but observable differences. Some are listed below:

1. We have assumed the existence of a par floater with the same maturity date as the default swap and that the coupon on the default swap contract has been set so that it has zero initial value.
2. We have assumed a common market-wide funding level of LIBOR + B . In practice, different market participants have different funding costs which therefore imply different spread levels.
3. We have assumed repo funding to term. Repo funding cannot usually be locked in

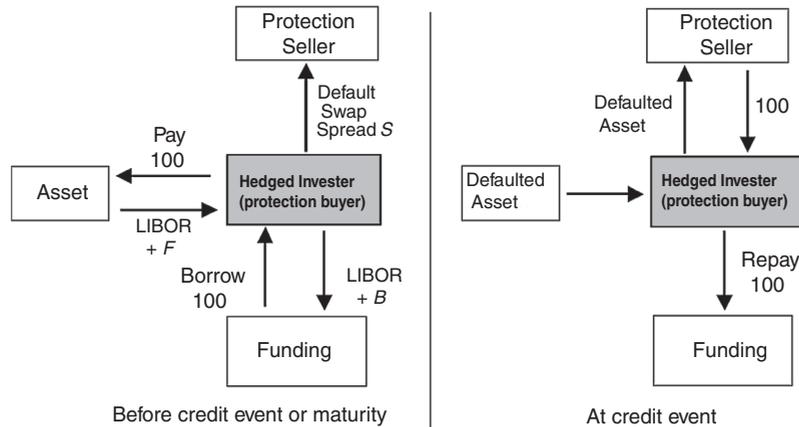


Figure 1 Theoretical Default Risk-Free Hedge for an Investor Who Buys Protection

to term but only for short periods of a couple of months only. One attraction of CDS is that unlike cash, they effectively lock in funding at LIBOR flat to maturity.

4. We have ignored accrued coupons. If the credit event occurs just before a coupon payment on the funding leg, the protection does not cover the loss of par plus coupon on the funding leg. We have also ignored the effect of the accrued CDS premium payment from the previous payment date.
5. We have assumed that the par floater is the cheapest-to-deliver asset.
6. We have ignored counterparty risk on the CDS. This is usually mitigated through the use of collateral.
7. Due to the difficulty of shorting cash bonds, any widespread market demand to go short a particular credit will first impact CDS, causing spreads to widen before cash.
8. For asset swaps the initial price of the asset must be close to par. This is because the loss on an asset swap of a bond trading with a full price P is about $P - R$. The credit risk is then only comparable to a default swap when the asset trades close to par.
9. We have ignored transaction costs.

Despite these assumptions, cash market spreads usually provide the starting point for where the default swap spreads should trade.

Empirically, there is a high correlation between the two spread levels. The difference between where and cash LIBOR spreads trade is known as the *default swap basis*, defined as

$$\text{Default swap basis} = S - F$$

There are now a significant number of market participants who actively trade the default swap basis, viewing it as a new relative value opportunity.³

PRICING OF A SINGLE-NAME CREDIT DEFAULT SWAP

Reduced versus Structural Models

To value credit derivatives it is necessary to be able to model the default risk, the recovery rate risk and the effect of interest rates. The two most commonly used approaches to model credit risk are structural models and reduced form models. The first structural model for credit-risky bonds was proposed by Black and Scholes (1973) who explained how equity owners hold a call option on the firm. After that Merton (1973 and 1974) extended the framework and analyzed the behavior of risky debt using the model.⁴

The second type of credit models, known as reduced-form models, are more recent.⁵ These models, most notably the Jarrow-Turnbull model and Duffie-Singleton model, do not look

inside the firm. Instead, they model directly the likelihood of a default occurring. Not only is the current probability of default modeled, they also attempt to model a “forward curve” of default probabilities that can be used to price instruments of varying maturities. Characterizing default as an event that occurs with a modeled probability has the effect of making default a surprise—the default event is a random event, which can suddenly occur at any time. All we know is its probability.

Reduced-form models are easy to calibrate to the term structure of CDS prices observed in the marketplace. This is known as working in an “arbitrage-free” framework. It is only by ensuring that a pricing model fits the market that a trader can be sure that he does not quote prices that expose him to any price arbitrages. The ability to quickly and easily calibrate to the entire CDS market is the major reason why reduced-form models are strongly favored by real-world practitioners in the credit derivatives markets for pricing. Structural-based models are used more for default prediction and credit risk management.

Increasingly, investors are seeking consistency between the markets that use different modeling approaches, as the interests in seeking arbitrage opportunities across various markets grows. Chen (2003) has demonstrated that all the reduced-form models described above can be regarded in a nonparametric framework. This nonparametric format makes the comparison of various models possible. Furthermore, as Chen contends, the nonparametric framework focuses the difference of various models on recovery.

The basic framework that underlies the reduced-form model is a binomial default process. There are two branches at each time point on the tree: default and survival. The branches that lead to default will terminate the contract and incur a recovery payment. The branches that lead to survival will continue the contract that will then face future defaults. This is a very general framework to describe how default oc-

curs and contract terminates. Various models differ in how the default probabilities are defined and the recovery is modeled.

Reduced form models use risk-neutral pricing to be able to calibrate to the market. In practice, we need to determine the risk-neutral probabilities in order to reprice the market and price other instruments not currently priced. In doing so, we do not need to know or even care about the real-world default probabilities.

Since in reality, a default can occur any time, to accurately value a default swap, we need a consistent methodology that describes the following: (1) when defaults occur, (2) how recovery is paid, and (3) how discounting is handled.

Survival Probability

Assume the risk-neutral probabilities exist. Then we can identify a series of risk-neutral *default probabilities* so that the weighted average of default and no-default payoffs can be discounted at the risk-free rate. The risk-free rate used in the pricing of CDS is LIBOR. This is because within a derivatives framework, the risk-free rate is close to the rate at which market dealers fund their hedges.

Assume $Q(t)$ to be the *survival probability* from now till some future time t . Then $Q(t) - Q(t + \tau)$ is the default probability between t and $t + \tau$ (that is, survive till t but default before $t + \tau$). Assume defaults can only be observed at multiples of τ . Then the total probability of default over the life of the CDS is the sum of all the per period default probabilities:

$$\sum_{j=1}^n Q[(j-1)\tau] - Q(j\tau) = 1 - Q(n\tau) = 1 - Q(T)$$

where $Q(0) = 1.0$ and $n\tau = T$, the maturity time of the CDS. It is no coincidence that the sum of the all the per-period default probabilities should equal one minus the total survival probability.

The survival probabilities have a useful application. A \$1 “risky” cash flow received at time t has a risk-neutral expected value of $Q(t)$ and a

present value of $P(t)Q(t)$ where P is the risk-free discount factor.

The value of the protection leg of a CDS is the present value of the payment of $(1 - R)$ at default. To take into account the timing of the default payment $(1 - R)$, we break the time to maturity into n intervals which correspond to the premium payment dates on the premium leg. This is a simple numerical approximation which works well given the quarterly payment convention of CDS. However a more exact model would break the time to maturity into monthly or even weekly time steps. For each time period we consider the probability of defaulting in each. The probability of defaulting in a forward interval $[(j - 1)\tau, j\tau]$ is given by

$$Q[(j - 1)\tau] - Q(j\tau) \quad (1)$$

We then discount the payment of $(1 - R)$ back to today by multiplying it by the risk-free discount factor $P(t)$. We then consider the likelihood of default occurring in all of the intervals by summing over all intervals. We therefore have

$$V = (1 - R) \sum_{j=1}^n P(j\tau) \{Q[(j - 1)\tau] - Q(j\tau)\} \quad (2)$$

where $R(\cdot)$ is the expected recovery rate determined by a CDS auction which takes place soon after a credit event rate.

In the above equation, it is implicitly assumed that the discount factor is independent of the survival probability. In reality, these two may be correlated—usually higher interest rates lead to more defaults because businesses suffer more from higher interest rates. To account for this we would need to introduce a stochastic probability and interest rate model. However, the effect of this correlation is almost negligible on the valuation of CDS and is further reduced by calibration. Equation (2) has no easy solution.⁶

Premium payments on the premium leg of a CDS terminate as soon as a credit event occurs. As a result the expected present value of the premium leg of the default swap is given by discounting each of the expected spread payments

by the risk-neutral discount factor weighted by the probability of surviving to each payment date. This is given by

$$S \sum_{j=1}^N \Delta_j P(j\tau) Q(j\tau)$$

where Δ_j is the corresponding year fraction in the appropriate basis convention (typically actual 360). By definition the value of the default swap spread is the value at which the premium and protection legs have the same present value. Hence, we have

$$V = S \sum_{j=1}^n \Delta_j P(j\tau) Q(j\tau) \quad (3)$$

giving

$$S = \frac{V}{\sum_{j=1}^n \Delta_j P(j\tau) Q(j\tau)}$$

Figure 2 depicts the general default and recovery structure. The payoff upon default of a default swap is par minus the recovery value as determined by any future ISDA auction which takes place after a credit event. As of today, the value of this recovery is unknown, we do not even know if a credit event will occur. As our model is based on the expected value of the protection leg, the recovery rate used has to be the expected value of the recovery rate conditional on a default and for this, market practitioners refer to historical recovery rates. Market convention is to use a 40% recovery rate as this is close to the average historical recovery

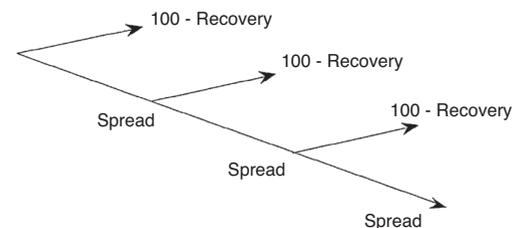


Figure 2 Payoff and Payment Structure of a CDS where as a simple approximation we assume that a credit event can only occur on a CDS premium leg payment date. In practice the credit event can occur at any time and the market standard model would take this into account.

rate for senior unsecured US corporate bonds – most CDS are linked to bonds which are senior unsecured.

In practice the portion of the premium payment that has accrued from the previous coupon payment date is paid by the protection buyer following the credit event. We have ignored it in our analysis since its effect on the calculated spread is small.⁷

Valuation of a Credit Default Swap

The valuation of CDS can be broken down into two separate tasks. The first is the determination of the default swap spread, which should be paid by a protection buyer at the initiation of a trade. This has already been discussed. The second is to determine the value of an existing CDS position, which we call the *mark-to-market* (MTM) or the upfront value. They are the same.

Since the recouping of CDS contracts in 2009, we can no longer state that the MTM or upfront of a new trade is always zero. The effect of fixing the premium leg coupon means that the risk of the reference entity must now be embedded in the initial cost of protection.

Once a CDS position has been established, changes in the current market CDS spread will mean that the MTM begins to deviate from its initial value and must be determined by observing the current level of default swap spreads in the market. To see how this is done, consider the following example.

An investor sells protection on a high yield reference entity for five years at an agreed contractual spread of 500 basis points. By selling protection the investor is assuming the credit risk of the reference entity as though he was buying one of the reference entity's issued bonds. A year later the reference entity's credit rating has improved and the market quoted 4-year par CDS spread is at 100 basis points. What is the MTM or upfront value of the position?

To begin with, the MTM value of the contract to the investor is given by the difference between what the investor is expecting to receive

minus what they are expected to pay. As a result we can write

$$\text{MTM} = + \text{Present value of four years of risky premium payments of 500 basis points} - \text{Present value of protection for the remaining four years}$$

We can also write that the current four-year par CDS spread of 100 basis points is the current break-even spread. By definition, the current value of a new four-year "par" CDS contract with a coupon equal to the par CDS spread is zero so we can write

$$\text{Present value of four years of risky premium payments of 100 basis points} = \text{Present value of protection for the remaining four years}$$

Substituting, we write

$$\text{MTM} = + \text{Present value of four years of risky premium payments of 500 basis points} - \text{Present value of four years of risky premium payments at 100 basis points}$$

which can be rewritten as

$$\text{MTM} = + \text{Present value of four years of risky premium payments of 400 basis points}$$

To go any further we have to compute the expected present value of these 400-basis points payments. However these payments are only made until the maturity of the CDS or to the time of a credit event, whichever occurs first. To compute the MTM we therefore need to weight each premium payment by the probability that there is no credit event up until that payment date. We therefore write

$$\text{MTM} = 400 \text{ basis points} \times \text{RPV01}$$

where the RPV01 is the "risky" price value of a basis point (PV01). This is defined as the present value of a 1 basis points payment made until the contractual maturity date of the position or to

the date of a credit event, whichever is sooner. Mathematically, we can write the RPV01 as

$$\text{RPV01} = \sum_{j=1}^n \Delta_j p(j\tau) Q(j\tau)$$

where Δ_j is the year fraction for the payment j in the appropriate basis (typically Actual 360). For quarterly paying CDS, Δ_j is usually close to or equal to 0.25. Bringing this all together, we can write the MTM value of a long protection position as

$$\text{MTM} = +[S(t, T) - S(0, T)] \times \text{RPV01}[S(t, T), R]$$

and that of a short protection position as

$$\text{MTM} = -[S(t, T) - S(0, T)] \times \text{RPV01}[S(t, T), R]$$

where $S(0, T)$ is the contractual spread of the contract, T is the contractual maturity date and $S(t, T)$ is the current par CDS spread to the contractual maturity date. It is essential to note that the RPV01 is a function of the market spread $S(t, T)$ and the assumed recovery rate R since both are used to imply out the risk-neutral survival probabilities.

To crystallize all of this theory, we present in Table 3 the valuation of the trade introduced at the beginning of this section in which an investor sells \$10 million of five-year protection at 500 basis points and then wishes to mark it to market one year later when the market has a flat term structure at 100 basis points. For simplicity we have assumed a flat LIBOR term structure at 2.5%. We assume a recovery rate of 40%. In particular we show the quarterly coupon payment dates (we have ignored holidays and weekends for simplicity) and the corresponding values of P and Q , calibrated to reprice the term structure of default swap spreads.

We see that the current par CDS spread is 100 basis points, and that the risky PV01 of the position is 3.7247—the present value of four years of risky 1 basis points payments is 3.7247 basis points. The resulting MTM value is \$1,489,892. This makes sense. The market has valued the risk of four year protection on the reference entity at 100bp in spread terms, but the fixed

Table 3 An Illustration of Calculation of the MTM Value

Long or short protection				Short
Notional (\$)				10,000,000
Contractual Spread (bp)				500
Settlement Date				20-Mar-13
Maturity Date				20-Mar-17
Flat LIBOR				2.50%
Par CDS Spread (bp)				100
Recovery Rate				40%
Payment Dates	YearFrac	Premium Leg Flows	Q(t)	P(t)
20-Mar-13			1.00000	1.00000
20-Jun-13	0.25556	127,778	0.99575	0.99372
20-Sep-13	0.25556	127,778	0.99152	0.98748
20-Dec-13	0.25278	126,389	0.98735	0.98135
20-Mar-14	0.25000	125,000	0.98324	0.97533
20-Jun-14	0.25556	127,778	0.97906	0.96920
20-Sep-14	0.25556	127,778	0.97490	0.96312
20-Dec-14	0.25278	126,389	0.97080	0.95714
20-Mar-15	0.25000	125,000	0.96677	0.95126
20-Jun-15	0.25556	127,778	0.96266	0.94529
20-Sep-15	0.25556	127,778	0.95857	0.93936
20-Dec-15	0.25278	126,389	0.95454	0.93352
20-Mar-16	0.25278	126,389	0.95053	0.92773
20-Jun-16	0.25556	127,778	0.94649	0.92190
20-Sep-16	0.25556	127,778	0.94246	0.91612
20-Dec-16	0.25278	126,389	0.93850	0.91043
20-Mar-17	0.25000	125,000	0.93460	0.90484
20-Jun-17	0.25556	127,778	0.93063	0.89916
Prot Leg PV	372,473			
Risky PV01	3.7247			
Replication Spread (bp)	100.00			
Contract MTM	1,489,892			

coupon is 500bp. A new investor wanting to sell four year protection is therefore being over-compensated and to correct for this, has to pay a large upfront cost.

CDS Risk and Sensitivities

Market practitioners using CDS usually consider two risk measures. First is the Credit01 or Spread01. This is the change in the MTM value of a CDS position for a 1 basis points parallel shift in the CDS curve. Then there is the Interest Rate 01 which is the change in the MTM value of a CDS position for a 1 basis points change in LIBOR. In practice the LIBOR sensitivity of a CDS is small, usually at least an order of magnitude less than that of the Credit01. This reflects the fact that a CDS is almost a pure credit play.

It is actually possible to make some simple approximations that make clear the dependence of the MTM on these inputs. First, we can approximate the CDS spread in terms of the risk-neutral annualized default probability p , and assumed recovery rate R , using the equation $S = p(1 - R)$. The interpretation is that the annualized spread received for assuming a credit risk should equal the annualized default probability times the loss on default, which in a CDS equals $(100\% - R)$. This approximation works very well in practice. If we assume a flat term structure of CDS spreads, approximate Δ with $1/4$, then we can approximate the MTM of a long protection position as

$$\text{MTM} = \frac{[S(t, T) - S(0, T)]}{4} \sum_{j=1}^N P(j\tau) \left[1 - \frac{S(t, T)}{1 - R} \right]^{j/4}$$

We can immediately draw a number of conclusions from this mathematical expression for the MTM value. First, the MTM value is not a linear function of the market spread $S(t, T)$. In fact the MTM value of a short protection position is convex in the market spread, just as the price of a corporate bond is convex in the yield. Furthermore, it is also clear that the recovery rate sensitivity of the MTM value is large when the market spread is large. This means that where the market spread is below, say, 300 basis points, one does not have to be so precise about the recovery rate assumption. However, if spreads become large (say, 300 basis points and above) the recovery rate sensitivity becomes increasingly significant and care must be taken in making a recovery rate assumption.

Calibrating the Recovery Rate Assumption

To be precise, the recovery rate assumption, R , is the assumed price of the cheapest-to-deliver asset into the CDS contract within 72 calendar days of the notification of the credit event. This is not known today. Nor can it be extracted from any market prices. In theory, this would be pos-

sible given the existence of an active and liquid digital default swap market. A digital default swap is a contract that pays the face value in the event of default—it is like a standard default swap but instead assumes a fixed recovery rate of zero. The ratio of the normal CDS spread and the digital default swap spread would equal $(1 - R)$. However, the lack of liquidity of the digital market makes this calibration approach impractical.

The usual starting point for calibrating recovery rates is to observe rating agency statistics. Both Moody's and S&P maintain significant databases of U.S. corporate bond defaults. Care must be taken to adjust any average recovery rates for country and sector effects. Recovery rates also have a link to the economic cycle. In recent years, average recovery rates have fallen well below the long-term averages computed by rating agencies. One reason why this is so is that Moody's, for example, defines the recovery rate of a bond as the price of that bond within some short period following the default. It is not the final value received by holders of the bond after going through the workout process. This means that the recovery rate is driven by the size of the bid for the bond in the distressed debt market. In periods of credit weakness, the distressed debt market is unable to absorb the oversupply of defaulted assets and the bid consequently falls.

Another consideration when marking recovery rate assumptions is to take into account that following a restructuring event, which is not a full default, the deliverable obligations may trade at higher prices than in a full default. Since rating agencies do not consider restructuring as a full default, this effect is not accounted for in their statistics. Typical recovery rates being quoted in the market for good quality credits vary between 30% and 45%.

When spreads are trading at very high levels of 1,000 basis points and above, it is important to look to the bond market to see if bond prices are revealing any information about the expected recovery rate in the event

of a default. For example, a recovery rate assumption of 40% would make no sense if one of the deliverable bonds into the CDS is trading at 30 cents on the dollar. In this case, the recovery rate assumption should clearly be moved below 30%.

The Practicalities of Unwinding a Credit Default Swap

A CDS is an over-the-counter (OTC) derivative contract. This means that unlike some other derivatives contracts it is not exchange traded. Instead it involves an agreement between two counterparties. As almost all CDS are traded within the framework of the ISDA Master Agreement, there is widespread standardization of the documentation of CDS and many counterparties are happy to trade these bilateral contracts in what is effectively a secondary market. To unwind a CDS before its maturity date, an investor may consider one of three courses of action:

1. Negotiate a cash upfront price with the original counterparty. The price should be the same as the MTM value calculated according to the model. In practice a bid-offer spread will have to be crossed. Part of this negotiation may involve some exchange of information as to the recovery rate assumptions used by both counterparties.
2. If the investor is shown a better upfront price by a counterparty different to the one with whom the initial trade was executed, they can ask to have the contract reassigned to this other counterparty and then close it out for a cash unwind value.
3. They may choose to enter into an offsetting position. For example, an investor who has sold protection for five years may decide a year later to close out the contract by selling protection for four years. The value of this combined position should exactly equal the model market to market.

Which one of these choices is made is usually determined by which is showing the best price.

Prior 2009 we would have said that option 3 is different from the others because it leaves the CDS holder with an ongoing position consisting of a future stream of risky cashflows equal to the difference between the spread of the initial contract and that of the new unwind contract. However now that CDS contracts on the same reference entity all trade with the same coupon, option 3 actually now leaves the parties with no net cashflows as both coupon streams will cancel each other. Instead the CDS unwind value is realised through the upfront cost of the offsetting position and will be the same as options 1 and 2.

The matching of coupons means there is no economic value in retaining both positions and both positions can be cancelled. Indeed this effect was the purpose of fixing CDS coupons since it means that in future, major dealers in CDS will no longer be left with many tens of thousands of legacy partially offsetting positions and their associated counterparty risk. This reduces the gross notional of the CDS market and should reduce fears, unfounded or not, about systemic risk. It may also help to facilitate any future plans to migrate CDS contracts from the OTC market to an exchange traded market.

KEY POINTS

- There is a fundamental no-arbitrage relationship that links the pricing of credit default swaps and the bonds which they reference. Various market and contractual differences mean that this relationship is not strictly obeyed at all times. However material deviations from this relationship should not persist.
- Since the recouping of CDS contracts in 2009, CDS contracts no longer trade with zero initial value. The valuation of a CDS contract has become the process of determining the upfront value of a contract.
- A pricing model for CDS contracts needs to take into account the different factors that drive the pricing of CDS. These include the market implied term structure for the

probability of survival/default, the expected recovery price if there is a credit event, and the level of interest rates used to discount future cashflows.

- The role of the standard valuation model set out in this chapter is to determine this upfront value. As market prices are actually quoted in the form of a term structure of CDS par spreads, the model must be able to exactly refit these par spreads and to then use the implied survival curve plus assumptions about the expected recovery price to determine the upfront value of any given CDS contract.
- An implementation of the standard pricing model has been produced by the ISDA and is available from www.cdsmodel.com.

NOTES

1. See O’Kane, Pedersen, and Turnbull (2003).
2. See O’Kane (2001).
3. For a discussion of the driving factors behind the basis, see O’Kane and McAdie (2001).
4. Geske (1977) extended the Black-Scholes-Merton model to include multiple debts. See also Geske and Johnson (1984). Many barrier models appear as an easy solution for analyzing the risky debt problem.
5. The name “reduced-form” was first given by Darrell Duffie to differentiate from the structural form models of the Black-Scholes-Merton type.
6. A continuous-time version of the equation can be found in the appendix of Chen, Fabozzi, and O’Kane (2003).
7. See O’Kane and Turnbull (2003).

REFERENCES

- Black, F., and Scholes, M. (1873). The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 3: 637–654.
- Chen, R-R. (2003). Credit risk modeling: A general framework. Working paper, Rutgers University.
- Chen, R-R., Fabozzi, F. J., and O’Kane, D. (2003). The valuation of credit default swaps. *Professional Perspectives on Fixed Income Portfolio Management* 4: 255–280.
- Duffie, D., and Singleton, K. J. (1999). Modeling the term structures of defaultable bonds. *Review of Financial Studies* 12, 4: 687–720.
- Geske, R. (1977). The valuation of corporate liabilities as compound options. *Journal of Financial and Quantitative Analysis* 12, 4: 541–552.
- Geske, R., and Johnson, H. (1984). The valuation of corporate liabilities as compound options: A correction. *Journal of Financial and Quantitative Analysis* 19, 2: 231–232.
- Jarrow, R., and Turnbull, S. (1995). Pricing derivatives on financial securities subject to credit risk. *Journal of Finance* 50, 1: 53–86.
- Merton, R. (1973). Theory of rational option pricing. *Bell Journal of Economics* 4, Spring: 141–183.
- Merton, R. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29, 2: 449–470.
- O’Kane, D. (2001). Credit derivatives explained. *Lehman Brothers Fixed Income Research*, March.
- O’Kane, D., and McAdie, R. (2003). Explaining the basis: Cash versus default swaps. *Lehman Brothers*, March.
- O’Kane, D., Pedersen, C., and Turnbull, S. (2003). The restructuring clause in credit default swap contracts. *Lehman Brothers Quantitative Credit Research Quarterly*.
- O’Kane, D., and Turnbull, S. (2003). Valuation of credit default swaps. *Lehman Brothers Fixed Income Research*, April.

Valuation of Fixed Income Total Return Swaps

REN-RAW CHEN, PhD

Professor of Finance, Graduate School of Business, Fordham University

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: A total return swap is a swap in which one party makes periodic floating rate payments to a counterparty in exchange for the total return realized on a reference asset (or underlying asset). The reference asset could be a credit-risky bond, a loan, a reference portfolio consisting of bonds or loans, an index representing a sector of the bond market, or an equity index. A total return swap can be used by asset managers for leveraging purposes and/or a transactionally efficient means for implementing a portfolio strategy. Bank managers use a total return swap as an efficient vehicle for transferring credit risk and as a means for reducing credit risk exposures. The Duffie-Singleton model can be used to value total return swaps.

In this entry we explain the valuation of total return swaps.¹ We begin with an intuitive approach.

AN INTUITIVE APPROACH

A typical *total return swap* is to swap the return on a reference asset for a risk-free return, usually the London Interbank Offered Rate (LIBOR). The cash flows for the swap buyer (that is, the total return receiver) are shown in Figure 1. In the figure, L_t is LIBOR at time t , s is the spread to LIBOR, and R_t is the total return at time t . The cash outlay at time t per \$1 of notional amount that must be made by the swap

buyer is $L_t + s$; the cash inflow at time t per \$1 of notional amount is R_t .

As a result, the pricing of a total return swap is to decide the right spread, s , to pay on the funding (that is, LIBOR) leg. Formally,

$$\hat{E}_0 \left\{ \sum_{j=1}^n \exp \left(- \int_0^{T_j} r(t) dt \right) [R_j - (L_j + s)] \right\} = 0$$

where r is the risk-free discount rate.

In words, the spread should be set so that the expected payoff of the total return swap is equal to zero. (We employ the standard risk-neutral pricing and discounting at the risk-free rate.) To make the matter simple (we shall discuss more rigorous cases later), we view r , R , and

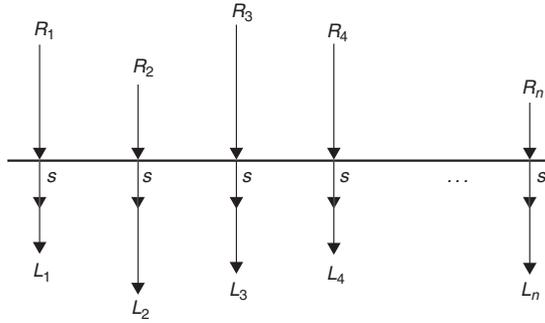


Figure 1 Cash Flows for the Total Return Receiver

L as three separate random variables. We then rearrange the above equation as

$$\begin{aligned} & \hat{E}_0 \left\{ \sum_{j=1}^n \exp \left(- \int_0^{T_j} r(t) dt \right) (R_j - L_j) \right\} \\ &= \hat{E}_0 \left\{ \sum_{j=1}^n \exp \left(- \int_0^{T_j} r(t) dt \right) \right\} s \end{aligned}$$

Exchanging expectation and summation of the right-hand side gives

$$\begin{aligned} & \hat{E}_0 \left\{ \sum_{j=1}^n \exp \left(- \int_0^{T_j} r(t) dt \right) \right\} \\ &= \sum_{j=1}^n \hat{E}_0 \left[\exp \left(- \int_0^{T_j} r(t) dt \right) \right] \\ &= \sum_{j=1}^n P(0, T_j) \end{aligned}$$

as the sum of risk-free pure discount bond prices. This implies

$$\begin{aligned} & \sum_{j=1}^n \hat{E}_0 \left[\exp \left(- \int_0^{T_j} r(t) dt \right) R_j - L_j \right] \\ &= \sum_{j=1}^n P(0, T_j) s \end{aligned}$$

The next step is to use the forward measure to simplify the left-hand side of the above equation:

$$\sum_{j=1}^n P(0, T_j) E_0^{F(j)} [R_j - L_j] = \sum_{j=1}^n P(0, T_j) s$$

Later, we show that the forward measure expectation of an asset gives the forward price of the asset. Hence, the left-hand side of the above equation gives two forward curves, one on the asset return, R , and the other on LIBOR, L :

$$\sum_{j=1}^n P(0, T_j) [f_j^R - f_j^L] = \sum_{j=1}^n P(0, T_j) s$$

where f_j^i is the forward rate of i ($i = R$ or L) for j periods ahead. Therefore, the spread can be solved easily as

$$s = \frac{\sum_{j=1}^n P(0, T_j) [f_j^R - f_j^L]}{\sum_{j=1}^n P(0, T_j)}$$

The result is intuitive: the spread is a weighted average of the expected difference between two floating-rate indexes. The weight is

$$\frac{P(0, T_j)}{\sum_{j=1}^n P(0, T_j)}$$

Note that all the weights should sum to one.

USING THE DUFFIE-SINGLETON MODEL

The difference in two floating rates is mainly due to their credit risk, otherwise they should both offer identical rates and give identical forward curves. As a consequence, to be rigorous about getting the correct result, we need to incorporate the credit risk in one of the indexes.

Among various choices, the model by Duffie and Singleton (1999) suits the best for this

situation. The *Duffie-Singleton model* is a popular reduced-form model that is used in credit risk modeling. In the model, the present value of any risky cash flow is defined as

$$C(t) = \left[\frac{S_{t+1} - S_t}{S_t} - L_{t+1} \right] N$$

where N is the notional, L is LIBOR, and S is the index level. As noted earlier, since both cash flows are random, it is a floating-floating swap. Also since the index is always higher than LIBOR because of credit risk, this swap requires a premium. As a result, the premium is computed as the sum of all future values, discounted and expected:

$$V = \sum_{j=1}^n \hat{E}_t \left[\exp \left(- \int_t^{T_j} [r(u) + q(u)] du \right) C(T_j) \right]$$

where q is the “spread” in the Duffie-Singleton model that incorporates the recovery rate and default probability.

THE FORWARD MEASURE

In this section, we show how the *forward measure* works and why a forward-adjusted expectation gives the forward value. We first state the separation principle that leads to the forward measure. Based on the no-arbitrage principle, the current value of any asset is the risk-neutral expected value of the discounted future payoff:

$$C(t) = \hat{E}_t \left[\exp \left(- \int_t^T r(u) du \right) C(T) \right]$$

The separation principle states that if we adopt the forward measure, then the above equation can be written as

$$C(t) = \hat{E}_t \left[\exp \left(- \int_t^T r(u) du \right) \right] E_t^{F(T)} [C(T)]$$

where $E_t^{F(T)} [\cdot]$ is the forward measure.² Note that the first term is nothing but the zero-coupon bond price:

$$P(t, T) = \hat{E}_t \left[\exp \left(- \int_t^T r(u) du \right) \right]$$

and hence

$$C(t) = P(t, T) E_t^{F(T)} [C(T)]$$

While we do not prove this result, we should note the intuition behind it. Let C be a zero-coupon bond expiring at time u . Then the above result can be applied directly and gives

$$P(t, s) = P(t, T) E_t^{F(T)} [P(T, u)]$$

or equivalently

$$E_t^{F(T)} [P(T, s)] = \frac{P(t, s)}{P(t, T)}$$

This is an indirect proof that the forward-adjusted expectation gives a forward value. The instantaneous forward rate can be shown to be the forward-adjusted expectation of the future instantaneous spot rate:

$$\begin{aligned} f(t, T) &= - \frac{d \ln P(t, T)}{dT} \\ &= - \frac{1}{P(t, T)} \hat{E}_t \left[\frac{d}{dT} \exp \left(- \int_t^T r(u) du \right) \right] \\ &= \frac{1}{P(t, T)} \hat{E}_t \left[\exp \left(- \int_t^T r(u) du \right) r(T) \right] \\ &= E_t^{F(T)} [r(T)] \end{aligned}$$

The discrete forward rates, $f_D(t, w, T)$ for all w and T , can also be shown to be the

forward-adjusted expectations of future discrete spot rates:

$$\begin{aligned} f_D(t, w, T) &= \frac{1}{\Psi(t, w, T)} - 1 \\ &= \frac{P(t, w)}{P(t, T)} - 1 \\ &= \frac{1}{P(t, T)} \hat{E}_t \left[\exp \left(- \int_t^T r(u) du \right) \frac{1}{P(w, T)} \right] - 1 \\ &= E_t^{F(T)} \left[\frac{1}{P(w, T)} - 1 \right] \end{aligned}$$

where $t < w < T$.

KEY POINTS

- A total return swap is a swap in which one party makes periodic floating rate payments to a counterparty in exchange for the total return realized on a reference asset such as a credit-risky bond.
- The pricing of a total return swap is to decide the right spread to pay on the funding leg.
- Using the standard risk-neutral pricing and discounting at the risk-free rate, the spread should be set so that the expected payoff of the total return swap is equal to zero.

- A reduced form model used in valuing credit derivatives, the Duffie-Singleton model, is employed to value total return swaps.
- The forward measure expectation of an asset gives the forward price of the asset that is the underlying for a total return swap.

NOTES

1. For a discussion of total return swaps and their applications, see Anson et al. (2004).
2. The derivation of this result can be found in a number of places. See, for example, Jamshidian (1987) and Chen (1996).

REFERENCES

- Anson, M. J. P., Fabozzi, F. J., Choudhry, M., and Chen, R-R. (2004). *Credit Derivatives: Instruments, Pricing, and Applications*. Hoboken, NJ: John Wiley & Sons.
- Chen, R-R. (1996). *Understanding and Managing Interest Rate Risks*. Singapore: World Scientific Publishing Company.
- Duffie, D., and Singleton, K. J. (1999). Modeling term structures of defaultable bonds. *Review of Financial Studies* 12, 4: 687–720.
- Jamshidian, F. (1987). Pricing of contingent claims in the one factor term structure model. New York: Merrill Lynch Capital Market.

Pricing of Variance, Volatility, Covariance, and Correlation Swaps

ANATOLIY SWISHCHUK, PhD, DSc

Professor of Mathematics and Statistics, University of Calgary

Abstract: Swaps are useful for volatility hedging and speculation. Volatility swaps are forward contracts on future realized stock volatility, and variance swaps are similar contracts on variance, the square of future volatility. Covariance and correlation swaps are covariance and correlation forward contracts, respectively, of the underlying two assets. Using change of time method, one can model and price variance, volatility, covariance, and correlation swaps.

Variance, volatility, covariance, and correlation swaps are relatively recent financial products that market participants can use for volatility hedging and speculation. The market for these types of swaps has been growing, with many investment banks and other financial institutions now actively quoting volatility swaps on various assets: stock indexes, currencies, and commodities.

A stock's volatility is the simplest measure of its riskiness or uncertainty. In this entry we describe, model, and price variance, volatility, covariance, and correlation swaps.

DESCRIPTION OF SWAPS

We begin with a description of the different kinds of swaps that we will be discussing in this entry: variance swaps, volatility swaps, covariance swaps, and correlation swaps. Table 1 provides a summary of studies dealing with these swaps.

Variance and Volatility Swaps

A stock's volatility is the simplest measure of its riskiness or uncertainty. Formally, the volatility σ_R is the annualized standard deviation of the stock's returns during the period of interest, where the subscript R denotes the observed or "realized" volatility.

Why trade volatility or variance swaps? As mentioned in Demeterfi et al. (1999, p. 9), "just as stock investors think they know something about the direction of the stock market so we may think we have insight into the level of future volatility. If we think current volatility is low, for the right price we might want to take a position that profits if volatility increases."

The easiest way to trade volatility is to use *volatility swaps*, sometimes called realized volatility forward contracts, because they provide only exposure to volatility and not other risk. *Variance swaps* are similar contracts on variance, the square of the future volatility. As noted by Carr and Madan (1998), both types of swaps

Table 1 Summary of Studies Dealing with Variance, Volatility, Covariance, and Correlation Swaps

Demeter et al. (1999)	<ul style="list-style-type: none"> • Explained properties and theory of both variance and volatility swaps. • Derived an analytical formula for theoretical fair value in the presence of realistic volatility skew. • Pointed out that volatility swaps can be replicated by dynamically trading the more straightforward variance swap.
Javaheri et al. (2002)	<ul style="list-style-type: none"> • Discussed the valuation and hedging of a GARCH(1,1) stochastic volatility model. • Used a general and flexible PDE approach to determine first two moments of the realized variance in a continuous or discrete context.
Brockhaus et al. (2000)	<ul style="list-style-type: none"> • Approximated the expected realized volatility via a convexity adjustment. • Provided an analytical approximation for the valuation of volatility swaps. • Analyzed other options with volatility exposure.
Swishchuk (2004)	<ul style="list-style-type: none"> • Priced covariance and correlation swaps in continuous time (Heston models for two stock prices)
Cheng et al. (2002)	<ul style="list-style-type: none"> • Priced covariance and correlation swaps in discrete time (Heston models for two stock prices)
Elliott and Swishchuk (2007)	<ul style="list-style-type: none"> • Studied option pricing formulae and pricing swaps for Markov-modulated Brownian with jumps.
Carr and Lee (2009)	<ul style="list-style-type: none"> • Provide an overview of the market of volatility derivatives and survey the early literature.
Swishchuk (2009a)	<ul style="list-style-type: none"> • Considered a semi-Markov modulated market consisting of a riskless asset or bond, B; and a risky asset or stock, S; whose dynamics depend on a semi-Markov process x: • Using the martingale characterization of semi-Markov processes, noted the incompleteness of semi-Markov modulated markets and found the minimal martingale measure. • Priced variance and volatility swaps for stochastic volatilities driven by the semi-Markov processes.
Swishchuk et al. (2010)	<ul style="list-style-type: none"> • Generalized results in Swishchuk (2009a) for the cases of the local current semi-Markov and local semi-Markov volatilities.
Kallsen et al. (2009)	<ul style="list-style-type: none"> • Priced variance and volatility swaps and options on variance in affine stochastic volatility models.
Swishchuk et al. (2010)	<ul style="list-style-type: none"> • Volatility and variance swaps for COGARCH(1,1).
Swishchuk (2005, 2006, 2007), Swishchuk et al. (2007), Swishchuk (2009a, 2010b), Swishchuk et al. (2010)	<ul style="list-style-type: none"> • Priced and modeled variance swaps for many stochastic volatility models with delay and jumps.
Swishchuk (2011)	<ul style="list-style-type: none"> • Priced variance and volatility swaps in energy markets
Howison et al. (2004)	<ul style="list-style-type: none"> • Considered the pricing of a range of volatility derivatives, including volatility and variance swaps and swaptions.

provide an easy way for investors to gain exposure to the future level of volatility.

A stock volatility swap's payoff at expiration is equal to

$$N(\sigma_R(S) - K_{vol})$$

where $\sigma_R(S)$ is the realized stock volatility (quoted in annual terms) over the life of contract,

$$\sigma_R(S) = \sqrt{\frac{1}{T} \int_0^T \sigma_s^2 ds}$$

σ_t is a stochastic stock volatility, K_{vol} is the annualized volatility delivery price, and N is the notional amount of the swap in dollar per annualized volatility point.

Although options market participants talk of volatility, it is variance, or volatility squared, that has more fundamental significance.¹ A variance swap is a forward contract on annualized variance, the square of the realized volatility. Its payoff at expiration is equal to

$$N(\sigma_R^2(S) - K_{var})$$

where $\sigma_R^2(S)$ is the realized stock variance (quoted in annual terms) over the life of the contract; that is,

$$\sigma_R^2(S) = \frac{1}{T} \int_0^T \sigma_s^2 ds$$

K_{var} is the delivery price for variance, and N is the notional amount of the swap in dollars per annualized volatility point squared. The holder of variance swap at expiration receives N dollars for every point by which the stock's realized variance $\sigma_R^2(S)$ has exceeded the variance delivery price K_{var} . Therefore, pricing the variance swap reduces to calculating the square of the realized volatility.

Valuing a variance forward contract or swap is no different from valuing any other derivative security. The value of a forward contract P on future realized variance with strike price K_{var} is the expected present value of the future payoff in the risk-neutral world:

$$P_{var} = E\{e^{-rT}(\sigma_R^2(S) - K_{var})\}$$

where r is the risk-free interest rate corresponding to the expiration date T , and E denotes the expectation. Thus, for calculating variance swaps we need to know only $E\{\sigma_R^2(S)\}$, namely the mean value of the underlying variance.

To calculate volatility swaps we need more. Using the Brockhaus and Long (2000) approximation (which is the second-order Taylor expansion for function \sqrt{x}) we have²

$$E\{\sqrt{\sigma_R^2(S)}\} \approx \sqrt{E\{V\}} - \frac{Var\{V\}}{8E\{V\}^{3/2}}$$

where $V = \sigma_R^2(S)$ and $\frac{Var\{V\}}{8E\{V\}^{3/2}}$ is the convexity adjustment.

Thus, to calculate the value of volatility swaps

$$P_{vol} = \{e^{-rT}(E\{\sigma_R(S)\} - K_{vol})\}$$

we need both $E\{V\}$ and $Var\{V\}$.

Later we explicitly solve the Cox-Ingersoll-Ross³ equation for the *Heston model* for *stochastic volatility*⁴ using the change of time method and present the formulas for price variance and volatility swaps for this model.

Covariance and Correlation Swaps

Options dependent on exchange rate movements, such as those paying in a currency different from the underlying currency, have an exposure to movements of the correlation be-

tween the asset and the exchange rate. This risk can be eliminated by using a covariance swap.

A *covariance swap* is a covariance forward contract of the underlying rates S^1 and S^2 , which have a payoff at expiration that is equal to

$$N(Cov_R(S^1, S^2) - K_{cov})$$

where K_{cov} is a strike price, N is the notional amount, and $Cov_R(S^1, S^2)$ is a covariance between two assets S^1 and S^2 .

Logically, a *correlation swap* is a correlation forward contract of two underlying rates S^1 and S^2 whose payoff at expiration is the following

$$N(Corr_R(S^1, S^2) - K_{corr})$$

where $Corr(S^1, S^2)$ is a realized correlation of two underlying assets S^1 and S^2 , K_{corr} is a strike price, and N is the notional amount.

Pricing covariance swaps, from a theoretical point of view, is similar to pricing variance swaps, since

$$Cov_R(S^1, S^2) = 1/4\{\sigma_R^2(S^1 S^2) - \sigma_R^2(S^1/S^2)\}$$

where S^1 and S^2 are two underlying assets, $\sigma_R^2(S)$ is a variance swap for the underlying assets, and $Cov_R(S^1, S^2)$ is a realized covariance of the two underlying assets S^1 and S^2 .

Thus, we need to know the variances for $S^1 S^2$ and for S^1/S^2 . Correlation $Corr_R(S^1, S^2)$ is defined as follows:

$$Corr_R(S^1, S^2) = \frac{Cov_R(S^1, S^2)}{\sqrt{\sigma_R^2(S^1)}\sqrt{\sigma_R^2(S^2)}}$$

where $Cov_R(S^1, S^2)$ is defined as above and $\sigma_R^2(S^1)$ is the realized variance for S^1 .

Given two assets S_t^1 and S_t^2 with $t \in [0, T]$, sampled on days $t_0 = 0 < t_1 < t_2 < \dots < t_n = T$ between today and maturity T , the log-return of each asset is

$$R_i^j = \log\left(\frac{S_{t_i}^j}{S_{t_{i-1}}^j}\right), \quad i = 1, 2, \dots, n, \quad j = 1, 2$$

Covariance and correlation can be approximated by

$$\text{Cov}_n(S^1, S^2) = \frac{n}{(n-1)T} \sum_{i=1}^n R_i^1 R_i^2$$

and

$$\text{Corr}_n(S^1, S^2) = \frac{\text{Cov}_n(S^1, S^2)}{\sqrt{\text{Var}_n(S^1)}\sqrt{\text{Var}_n(S^2)}}$$

respectively.

MODELING AND PRICING OF VARIANCE, VOLATILITY, COVARIANCE, AND CORRELATION SWAPS WITH STOCHASTIC VOLATILITY

In this section, we explicitly solve the Cox-Ingersoll-Ross equation for the stochastic volatility Heston model, using the change of time method, and present the formulas for price variance, volatility, covariance, and correlation swaps for this model.

Stochastic Volatility: Heston Model

Let $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ be a probability space with filtration \mathcal{F}_t , $t \in [0, T]$. Assume that the underlying asset S_t in the risk-neutral world and variance follow the following model (see Heston, 1993):

$$\begin{cases} \frac{dS_t}{S_t} = r_t dt + \sigma_t dw_t^1 \\ d\sigma_t^2 = k(\theta^2 - \sigma_t^2)dt + \gamma\sigma_t dw_t^2 \end{cases} \quad (1)$$

where r_t is the deterministic interest rate, σ_0 and θ are short and long volatility, $k > 0$ is the reversion speed, $\gamma > 0$ is the volatility (of volatility) parameter, and w_t^1 and w_t^2 are independent standard Wiener processes.

The Heston asset process has a variance σ_t^2 that follows a Cox-Ingersoll-Ross process, described by the second equation in (1). If the volatility σ_t follows the Ornstein-Uhlenbeck process (see, for example, Øksendal, 1998), then

Itô's lemma shows that the variance σ_t^2 follows the process described exactly by the second equation in (1). Note that if $2k\theta^2 > \gamma^2$, then $\sigma_t^2 > 0$ with $P = 1$ (see Heston, 1993).

Solving the equation for variance σ_t^2 in (1) explicitly using the change of time method gives

$$d\sigma_t^2 = k(\theta^2 - \sigma_t^2)dt + \gamma\sigma_t dw_t^2 \quad (2)$$

and takes the following form:

$$\sigma_t^2 = e^{-kt}(\sigma_0^2 - \theta^2 + \tilde{w}^2(\phi_t^{-1})) + \theta^2 \quad (3)$$

where $\tilde{w}^2(t)$ is an \mathcal{F}_t -measurable one-dimensional Wiener process, and ϕ_t^{-1} is an inverse function to ϕ_t :

$$\phi_t = \gamma^{-2} \int_0^t \{e^{k\phi_s}(\sigma_0^2 - \theta^2 + \tilde{w}^2(s)) + \theta^2 e^{2k\phi_s}\}^{-1} ds \quad (4)$$

This result simply follows from the following substitution

$$v_t = e^{kt}(\sigma_t^2 - \theta^2) \quad (5)$$

into the equation (2) instead of σ_t^2 .

Note that if $2k\theta^2 > \gamma^2$, then $\sigma_t^2 > 0$ with $P = 1$ (see, for example, Heston, 1993). From (5) it follows that

$$v_t e^{-kt} + \theta^2$$

is strictly positive too. If we take the integrand in the last integral we obtain

$$\begin{aligned} & [e^{k\phi_s}(\sigma_0^2 - \theta^2 + \tilde{w}^2(t)) + \theta^2 e^{2k\phi_s}]^{-1} \\ &= [e^{2k\phi_s}(e^{-kt}(\sigma_0^2 - \theta^2 + \tilde{w}^2(t))) + \theta^2]^{-1} \\ &= [e^{k\phi_s} \sqrt{e^{-kt}(\sigma_0^2 - \theta^2 + \tilde{w}^2(t)) + \theta^2}]^{-2} \\ &= [e^{k\phi_s} \sqrt{e^{-kt}v_t + \theta^2}]^{-2} \end{aligned}$$

since $v_t = \sigma_0^2 - \theta^2 + \tilde{w}^2(t)$. In the above integrals, the expression under the square root sign is positive and the square root is well defined. Hence, the last expression and therefore, the integrand in the integral in (4), are strictly positive. It means that ϕ_t is a monotone function and there exists an inverse function ϕ_t^{-1} in (3).

Valuing of Variance and Volatility Swaps

From previous results we get the following expression for the price of a variance and volatility swap.

The value (or price) P_{var} of a variance swap is

$$P_{var} = e^{-rT} \left[\frac{1 - e^{-kT}}{kT} (\sigma_0^2 - \theta^2) + \theta^2 - K_{var} \right] \tag{6}$$

and the value (or price) P_{vol} of volatility swap is approximately

$$P_{vol} \approx e^{-rT} \left\{ \left(\frac{1 - e^{-kT}}{kT} (\sigma_0^2 - \theta^2) + \theta^2 \right)^{1/2} - \left(\frac{\gamma^2 e^{-2kT}}{2k^3 T^2} [(2e^{2kT} - 4e^{kT} kT - 2)(\sigma_0^2 - \theta^2) + (2e^{2kT} kT - 3e^{2kT} + 4e^{kT} - 1)\theta^2] \right) / \left[8 \left(\frac{1 - e^{-kT}}{kT} (\sigma_0^2 - \theta^2) + \theta^2 \right)^{3/2} \right] - K_{vol} \right\} \tag{7}$$

The same expressions for $E[V]$ and for $Var[V]$ also may be found in Brockhaus and Long (2000).

Valuing of Covariance and Correlation Swaps

To value a covariance swap the following must be calculated

$$P = e^{-rT} (ECov(S^1, S^2) - K_{cov}) \tag{8}$$

To calculate $ECov(S^1, S^2)$ we need to calculate $E\{\sigma_R^2(S^1 S^2) - \sigma_R^2(S^1/S^2)\}$ for the two underlying assets S^1 and S^2 .

Let $S_t^i, i = 1, 2$, be two strictly positive Ito's processes given by the following model

$$\begin{cases} \frac{dS_t^i}{S_t^i} = \mu_t^i dt + \sigma_t^i dw_t^i \\ d(\sigma_t^i)^2 = k^i (\theta_t^i - (\sigma_t^i)^2) dt + \gamma^i \sigma_t^i dw_t^j, \quad i = 1, 2, j = 3, 4 \end{cases}$$

where $\mu_t^i, i = 1, 2$, are deterministic functions, $k^i, \theta^i, \gamma^i, i = 1, 2$, are defined in a similar way as in (1), standard Wiener processes $w_t^j, j = 3, 4$, are independent, $[w_t^1, w_t^2] = \rho_t dt, \rho_t$ is deterministic function of time, $[,]$ means the quadratic covariance, and standard Wiener processes $w_t^i, i = 1, 2$, and $w_t^j, j = 3, 4$, are independent.

We note that

$$d \ln S_t^i = m_t^i dt + \sigma_t^i dw_t^i$$

where

$$m_t^i := \left(\mu_t^i - \frac{(\sigma_t^i)^2}{2} \right)$$

and

$$\begin{aligned} Cov_R(S_T^1, S_T^2) &= \frac{1}{T} [\ln S_T^1, \ln S_T^2] \\ &= \frac{1}{T} \left[\int_0^T \sigma_t^1 dw_t^1, \int_0^T \sigma_t^2 dw_t^2 \right] \\ &= \frac{1}{T} \int_0^T \rho_t \sigma_t^1 \sigma_t^2 dt \end{aligned}$$

Let us show that

$$[\ln S_T^1, \ln S_T^2] = \frac{1}{4} ([\ln(S_T^1 S_T^2)] - [\ln(S_T^1/S_T^2)]) \tag{9}$$

First, note that

$$d \ln(S_t^1 S_t^2) = (m_t^1 + m_t^2) dt + \sigma_t^+ dw_t^+$$

and

$$d \ln(S_t^1/S_t^2) = (m_t^1 - m_t^2) dt + \sigma_t^- dw_t^-$$

where

$$(\sigma_t^\pm)^2 := (\sigma_t^1)^2 \pm 2\rho_t \sigma_t^1 \sigma_t^2 + (\sigma_t^2)^2$$

and

$$dw_t^\pm := \frac{1}{\sigma_t^\pm} (\sigma_t^1 dw_t^1 \pm \sigma_t^2 dw_t^2)$$

Processes w_t^\pm above are standard Wiener processes by the Levi-Kunita-Watanabe theorem and σ_t^\pm are defined above.

In this way, we obtain that

$$\begin{aligned} [\ln(S_t^1 S_t^2)] &= \int_0^t (\sigma_s^+)^2 ds = \int_0^t ((\sigma_s^1)^2 + 2\rho_t \sigma_s^1 \sigma_s^2 \\ &\quad + (\sigma_s^2)^2) ds \end{aligned} \tag{10}$$

and

$$\begin{aligned} [\ln(S_t^1/S_t^2)] &= \int_0^t (\sigma_s^-)^2 ds = \int_0^t ((\sigma_s^1)^2 - 2\rho_t \sigma_s^1 \sigma_s^2 \\ &+ (\sigma_s^2)^2) ds \end{aligned} \quad (11)$$

From (9)–(11) we have directly formula (8):

$$[\ln S_T^1, \ln S_T^2] = \frac{1}{4}([\ln(S_T^1 S_T^2)] - [\ln(S_T^1/S_T^2)]) \quad (12)$$

Thus, from (12) we obtain that

$$Cov_R(S^1, S^2) = 1/4(\sigma_R^2(S^1 S^2) - \sigma_R^2(S^1/S^2))$$

Returning to the valuation of the covariance swap in (8) we have

$$P = E\{e^{-rT}(Cov(S^1, S^2) - K_{cov})\} = \frac{1}{4}e^{-rT}(E\sigma_R^2(S^1 S^2) - E\sigma_R^2(S^1/S^2) - 4K_{cov})$$

The problem now has reduced to the same problem as above, but instead of σ_t^2 we need to take $(\sigma_t^+)^2$ for $S^1 S^2$ and $(\sigma_t^-)^2$ for S^1/S^2 (with $(\sigma_t^\pm)^2 = (\sigma_t^1)^2 \pm 2\rho_t \sigma_t^1 \sigma_t^2 + (\sigma_t^2)^2$), and proceed with similar calculations as for the variance and volatility swaps.

NUMERICAL EXAMPLE: VOLATILITY SWAP FOR S&P60 CANADA INDEX

In this section, we apply the analytical solutions provided above to price a swap on the volatility of the S&P60 Canada Index for five years (January 1997–February 2002).⁵

Suppose that at the end of February 2002 we wanted to price the fixed leg of a volatility swap based on the volatility of the S&P60 Canada Index. The statistics on log returns for the S&P60 Canada Index for the five years covering January 1997–February 2002 are presented in Table 2.

From the statistical data for the S&P60 Canada Index log returns for the 5-year historical period (1,300 observations from January 1997 to February 2002) it may be seen that the data

Table 2 Statistics on Log Returns S&P60 Canada Index

Series:	LOG RETURNS S&P60 CANADA INDEX
Sample:	1 1300
Observations:	1300
Mean	0.000235
Median	0.000593
Maximum	0.051983
Minimum	-0.101108
Std. Dev.	0.013567
Skewness	-0.665741
Kurtosis	7.787327

exhibit leptokurtosis. If we take a look at the S&P60 Canada Index log returns for the 5-year historical period, we observe volatility clustering in the return series. These facts indicate the presence of conditional heteroscedasticity. A GARCH(1,1) regression is applied to the series and the results are obtained as in Table 3. This table allows one to generate different input variables for the volatility swap model.

We use the following relationships: $\theta = \frac{V}{dt}$, $k = \frac{1-\alpha-\beta}{dt}$, $\gamma = \alpha\sqrt{\frac{\xi-1}{dt}}$, to calculate the following discrete GARCH(1,1) parameters:

- ARCH(1,1) coefficient $\alpha = 0.060445$
- GARCH(1,1) coefficient $\beta = 0.927264$
- The Pearson kurtosis (fourth moment of the drift-adjusted stock return) $\xi = 7.787327$
- Long volatility $\theta = 0.05289724$; $k = 3.09733$
- $\gamma = 2.499827486$
- Short volatility $\sigma_0 = 0.01$

Parameter V may be found from the expression $V = \frac{C}{1-\alpha-\beta}$, where $C = 2.58 \times 10^{-6}$ is defined in Table 3. Thus, $V = 0.00020991$; $dt = 1/252 = 0.003968254$.

Applying the analytical solutions (6) and (7) for a swap maturity T of 0.91 years, we find the following values:

$$E\{V\} = \frac{1 - e^{-kT}}{kT}(\sigma_0^2 - \theta^2) + \theta^2 = 0.3364100835$$

Table 3 Estimation of the GARCH(1,1) Process

Dependent Variable: Log returns of S&P60 Canada Index Prices
 Method: ML-ARCH
 Included Observations: 1,300
 Convergence achieved after 28 observations

	Coefficient	Std. error	z-statistic	Prob.
C	0.000617	0.000338	1.824378	0.0681
Variance Equation				
C	2.58E-06	3.91E-07	6.597337	0
ARCH(1)	0.060445	0.007336	8.238968	0
GARCH(1)	0.927264	0.006554	141.4812	0
R-squared	-0.000791	Mean dependent var	-	0.000235
Adjusted R-squared	-0.003108	S.D. dependent var	-	0.013567
S.E. of regression	0.013588	Akaike info criterion	-	-5.928474
Sum squared resid	0.239283	Schwartz criterion	-	-5.912566
Log likelihood	3857.508	Durbin-Watson stat	-	1.886028

and

$$\begin{aligned}
 Var(V) &= \frac{\gamma^2 e^{-2kT}}{2k^3 T^2} [(2e^{2kT} - 4e^{kT} kT - 2)(\sigma_0^2 - \theta^2) \\
 &\quad + (2e^{2kT} kT - 3e^{2kT} + 4e^{kT} - 1)\theta^2] \\
 &= 0.0005516049969
 \end{aligned}$$

The convexity adjustment $\frac{Var(V)}{8E\{V\}^{3/2}}$ is equal to 0.0003533740855.

If the nonadjusted strike is equal to 18.7751%, then the adjusted strike is equal to

$$18.7751\% - 0.03533740855\% = 18.73976259\%$$

This is the fixed leg of the volatility swap for a maturity $T = 0.91$.

Repeating this approach for a series of maturities up to 10 years, we obtain the result shown in Figure 2 for the S&P60 Canada Index Volatility Swap. Figure 1 illustrates the nonadjusted

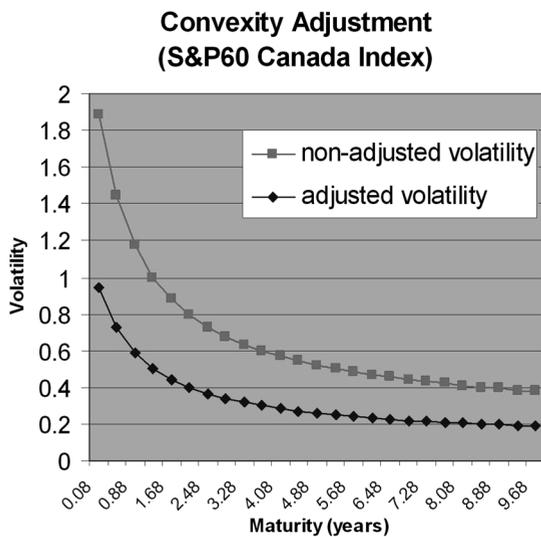


Figure 1 Convexity Adjustment

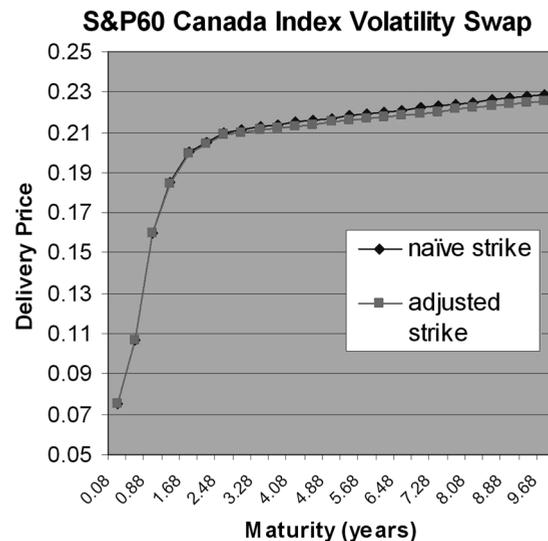


Figure 2 S&P60 Canada Index Volatility Swap

and adjusted volatility for the same series of maturities (see formula (7)).

KEY POINTS

- Variance, volatility, covariance, and correlation swaps are useful for volatility hedging and speculation.
- Volatility swaps are forward contracts on future realized stock volatility.
- Variance swaps are similar contracts on variance, the square of the future volatility.
- Covariance and correlation swaps are covariance and correlation forward contracts, respectively, of the underlying two assets.
- Using change of time one can model and price variance, volatility, covariance, and correlation swaps for the stochastic volatility Heston model.

NOTES

1. See Demeterfi, Derman, Kamal, and Zou (1999).
2. See also Javaheri et al. (2002, p. 16).
3. See Cox, Ingersoll, and Ross (1985).
4. See Heston (1993).
5. These data were supplied by Raymond Théoret (Université du Québec à Montréal, Montréal, Québec, Canada) and Pierre Rostan (Analyst at the R&D Department of Bourse de Montréal and Université du Québec à Montréal, Montréal, Québec, Canada). They calibrated the GARCH parameters from five years of daily historic S&P60 Canada Index from January 1997 to February 2002. See Théoret, Zabré, and Rostan (2002).

REFERENCES

- Brockhaus, O., and Long, D. (2000). Volatility swaps made simple. *Risk*, January: 92–96.
- Carr P., and Lee R. (2009) Volatility derivatives. *Annual Review of Financial Economics*, 1: 1–21.
- Carr, P., and Madan, D. (1998). Towards a Theory of Volatility Trading. *Volatility*, Risk Book Publications.
- Cheng, R., Lawi, S., Swishchuk, A., Badescu, A., Ben Mekki, H., Gashaw, A., Hua, Y., Molyboga, M., Neocleous, T., and Petrachenko, Y. (2002). Price Pseudo-Variance, Pseudo-Covariance, Pseudo-Volatility, and Pseudo-Correlation Swaps in Analytical Closed-Forms, *Proceedings of the Sixth PIMS Industrial Problems Solving Workshop*, PIMS IPSW 6, University of British Columbia, Vancouver, Canada, May 27–31, 2002, pp. 45–55.
- Chernov, R., Gallant, E., Ghysels, E., and Tauchen, G. (2003). Alternative models for stock price dynamics. *Journal of Econometrics*, 116: 225–257.
- Demeterfi, K., Derman, E., Kamal, M., and Zou, J. (1999). A guide to volatility and variance swaps. *The Journal of Derivatives*, Summer: 9–32.
- Elliott, R., and Swishchuk, A. (2007). Pricing options and variance swaps in Markov-modulated Brownian markets. *Hidden Markov Models in Finance*. New York: Springer.
- Howison, S., Rafailidis, A., and Rasmussen, H. (2004). On the pricing and hedging of volatility derivatives. *Applied Mathematical Finance*, 11: 317–346.
- Javaheri, A., Wilmott, P., and Haug, E. (2002). GARCH and volatility swaps. *Wilmott Magazine*, January: 17p.
- Kallsen, J., and Shiryaev, A. (2002). Time change representation of stochastic integrals. *Theory Probability and Its Applications*, 46, 3: 522–528.
- Swishchuk, A. (2011). Variance and volatility swaps in energy markets. *Journal of Energy Markets*, forthcoming.
- Swishchuk, A., and Malenfant, K. (2011). Pricing of variance swaps for Lévy-based stochastic volatility with delay. *International Review of Applied Financial Issues and Economics*, forthcoming.
- Swishchuk, A., and Li, X. (2011). Variance swaps for stochastic volatility with delay and jumps. *International Journal of Stochastic Analysis*, Article ID 435145.
- Swishchuk, A., and Couch, M. (2010). Volatility and variance swaps for COGARCH(1,1) model. *Wilmott Magazine*, 2, 5: 231–246.
- Swishchuk, A., and Manca, R. (2010). Modeling and pricing of variance swaps for local Semi-Markov volatility in financial engineering. *Mathematical Models in Engineering*, New York: Hindawi Publications, 1–17.

- Swishchuk, A. (2009a). Pricing of Variance and Volatility Swaps with Semi-Markov Volatilities, *Canadian Applied Mathematics Quarterly*, 18, 4.
- Swishchuk, A. (2009b). Variance swaps for local stochastic volatility with delay and jumps. Working paper, Calgary: University of Calgary.
- Swishchuk, A. (2007). Change of time method in mathematical finance. *Canadian Applied Mathematics Quarterly*, 15, 3: 299–336.
- Swishchuk, A. (2006). Modeling and Pricing of Variance Swaps for Multi-Factor Stochastic Volatilities with Delay. *Canadian Applied Mathematics Quarterly*, 14, 4.
- Swishchuk, A. (2005). Modeling and Pricing of Variance Swaps for Stochastic Volatilities with Delay. *Wilmott Magazine*, 19, September: 63–73.
- Swishchuk, A. (2004). Modelling and valuing of variance and volatility swaps for financial markets with stochastic volatilities. *Wilmott Magazine*, 2, September: 64–72.
- Théoret, R., Zabré, L., and Rostan, P. (2002). Pricing volatility swaps: Empirical testing with Canadian data. Working paper. Centre de Recherche en Gestion, Université du Québec à Montréal, Document 17-2002.

Modeling, Pricing, and Risk Management of Assets and Derivatives in Energy and Shipping

PAUL D. SCLAVOUNOS, PhD

Professor of Mechanical Engineering, Massachusetts Institute of Technology

Abstract: Derivatives are financial contingent claims designed for the pricing, transfer, and management of risk embedded in underlying securities in the fixed income, equity, and foreign exchange markets. Their rapid growth spurred their introduction to the energy commodity and shipping markets where the underlying assets are real commodities, crude oil, refined products, natural gas, electricity, and shipping tonnage. Risk-neutral pricing and stochastic models developed for financial derivatives have been extended to energy derivatives for the modeling of correlated commodity and shipping forward curves and for the pricing of their contingent claims. This has enabled the valuation and risk management of a wide range of assets and derivatives in the energy and shipping markets. They include storage for natural gas, floating storage of crude oil, products and liquefied natural gas in tankers, refineries, power plants and utility scale wind farms, shipping structured securities, cargo vessels, and shipping derivatives portfolios.

Investments in energy and *shipping* assets are exposed to interest rate, commodity price, and freight rate risks. The management of these risks has led to the introduction and widespread use of *derivatives*, which have experienced explosive growth over the past several decades. In the fixed income market, interest rate *futures* and *futures options* emerged in the 1980s in response to the need to hedge interest rate swap risk. This led to the development of financial models for the arbitrage-free evolution of the term structure of interest rates and the pricing of a wide range of fixed income derivatives, laying the foundation for the development of analogous models for the

arbitrage-free evolution of the *forward curves* of physical commodities including crude oil, its refined products, natural gas, and recently shipping freight rates.

Commodity futures settle physically against the price of a spot commodity that must be delivered at the contract expiration or in cash against a spot commodity index. The latter is the case in shipping where *forward freight agreements* (FFAs) and freight rate futures settle against shipping indexes composed of a basket of freight rates. The first generation of commodity futures models was based on the development of stochastic models for the spot index and the use of the principles of

risk-neutral pricing for the valuation of derivatives written on the spot. Recent models are based on the insight that in the absence of arbitrage futures prices with daily credits and debits into a margin account are martingales. This has enabled the modeling of the evolution of futures prices of any tenor as lognormal diffusions with zero drift in a Gaussian setting. The primary unknown in this model of futures prices is the volatility term structure, which may be estimated from market prices of liquid futures and futures options. The arbitrage-free price process for the spot commodity or underlying index follows from this martingale representation of the entire commodity forward curve in the limit of small tenors.

The martingale representation of the commodity forward curve lends itself to a parsimonious modeling using the powerful statistical techniques of *principal components analysis* in the case of a single futures curve and of *canonical correlation analysis* in the case of multiple correlated forward curves. In both cases a small number of statistical factors may be derived, which are shown to follow mean reverting log-Ornstein-Uhlenbeck diffusions. This financial modeling and statistical inference framework of cross-correlated energy commodity and shipping forward curves allows the pricing of a wide range of vanilla, spread, and exotic derivatives written on futures contracts. Moreover, the model of the forward curve in terms of a small number of statistical factors enables the explicit valuation and *hedging* of a wide range of energy and shipping assets with cash flow exposures that may be replicated by the prices of traded futures contracts.

This entry reviews the fundamental developments that led to the introduction of financial and commodity derivatives, their stochastic modeling, and risk-neutral pricing. Pricing in Gaussian and non-Gaussian settings is addressed and shown in most cases to lead to closed-form results even when the underlying is represented by an advanced stochastic model. The martingale modeling of forward

curves and their estimation by a principal components analysis is discussed for the crude oil market, its refined products, natural gas, and the shipping market.

The valuation of real assets and financial claims of energy and shipping entities is discussed. The parsimonious form of the arbitrage-free factor models of the pertinent forward curves enable the pricing of these assets and securities by risk-neutral pricing. Their *risk management* is also addressed, drawing upon the explicit form of the underlying factor model and the techniques of *stochastic optimal control*, which have found widespread use for the management of portfolios of financial securities.

ENERGY COMMODITY PRICE MODELS

The past several decades have witnessed the emergence and rapid development of the fields of financial engineering and derivatives. Grown out of Paul Samuelson's foundational insights on the relationship between informationally efficient markets and the random walk and his introduction of the lognormal diffusion model of security prices, a wide range of stochastic models of security prices and arbitrage-free valuation methods were developed for the pricing of derivatives written on financial securities, real assets, and other variables (see Samuelson, 1965). The use of these models and pricing methods in the fixed income, equity, foreign exchange, and credit markets is growing as is the complexity of the mathematical, econometric, and filtering methods necessary for their implementation. More recently, these methods have been adapted to the energy and shipping sectors in order to control the high volatility of energy prices and freight rates and spur new investment.

Spot Price Models

Energy commodity prices are characterized by idiosyncrasies not encountered in the

financial markets. The volatility of the price of oil, natural gas, and especially electricity is a lot larger than that of currencies, interest rates, and equities. Energy prices often exhibit mean reversion, seasonality, and sharp and asymmetric spikes, which require the development of advanced price models and derivative valuation methods, extensions of the Black-Scholes-Merton stock option pricing formula. Moreover, a complex interaction often exists between the attributes of the spot physical commodity and its forward contracts and other derivatives that is not present for financial securities and their derivatives, which settle electronically and do not require the delivery of a physical asset. This requires the use of the extended *risk-neutral valuation* method of derivatives written on real assets and other variables that are not tradable (Ross, 1976).

Standard reduced form stochastic models for the spot price of crude oil and natural gas are diffusions that account for mean reversion and seasonality and depend on hidden economic factors. They include a stochastic convenience yield—the implied dividend received by the owner of the commodity held in storage—and a long-term stochastic equilibrium price to which the spot price mean reverts. The two-factor spot price models of Gibson and Schwartz (1990), Schwartz (1997), and Schwartz and Smith (2000) model the spot price and its factors as diffusions and permit the explicit valuation of futures and forward contracts and their options written on the spot commodity using the extended risk-neutral valuation of derivatives written on real assets (Hull, 2003). More general spot price models that may include stochastic volatility and jumps are discussed in Clewlow and Strickland (2000) and London (2007). In the study of Cortazar and Naranjo (2006) the entire oil futures curve and its volatility term structure are shown to be very well modeled by a four-factor spot price Gaussian model, which was estimated by Kalman filtering.

Stochastic models for the evolution of the electricity prices must account for sharp and asym-

metric spikes, strong mean reversion, jumps, and a dependence on structural factors affecting the electricity market. Reduced form stochastic models of electricity prices are usually jump-diffusions and Levy processes. An example is the jump-diffusion model of Kou (2002), which permits the independent parametric adjustment of the tail thicknesses of its probability distribution and allows the explicit pricing of electricity derivatives. Other models are discussed in Eydeland and Wolyniec (2003), London (2007), and Bength, Bength, and Koekebakker (2008). Analogous models apply to the modeling of the spot price process of shipping freight rates.

Forward Curve Models

Crude oil and natural gas have liquid futures contracts trading on the New York Mercantile Exchange (NYMEX) and the Intercontinental Exchange (ICE) with tenors of several years. For these commodities arbitrage-free forward curve models have been developed by Mil-tersen and Schwartz (1998), which accept as input the market prices of liquid futures and lead to the pricing of a number of other derivatives. The arbitrage-free evolution of the spot price follows from futures contracts of small tenors.

The modeling of the oil and natural gas futures curve is based on the Heath-Jarrow-Morton (HJM) framework developed for the arbitrage-free modeling of the term structure of interest rates. A principal task of the HJM framework is the parameterization of the volatility and correlation structure of the futures curve by a small number of independent factors using a principal components analysis. This was carried out by Scлавounos and Ellefsen (2009), where it was shown that three principal components capture most of the fluctuations of the forward curve. In the same paper the arbitrage-free evolution of the spot price was derived as implied in equilibrium by the forward curve and was shown to be driven by three independent factors that follow mean

reverting logarithmic Ornstein-Uhlenbeck (log-OU) processes with stochastic drifts. Calls, puts, swaps, caps, and their options written on futures contracts may then be valued explicitly as in the interest rate markets for use in energy risk management applications (Hull, 2003; Musiela and Rutkowski, 2005).

Energy Derivatives

In addition to the standard derivatives discussed above, more complex derivatives have been introduced in the energy markets reflecting the economics of energy assets. In particular, power plants are exposed to the spot/futures price difference of two *energy commodities*, for example, natural gas/electricity, coal/electricity; refineries are exposed to the price differentials of two fuels—crude oil/gasoline, crude oil/jet fuel; and oil and natural gas pipelines and electricity transmission lines are exposed to the price differentials of the same spot commodity at two different geographical locations.

A partial list of exotic derivatives used for the valuation, hedging, and risk management of energy assets include options on the spread between two futures contracts with different expirations written on the same commodity, options on the price difference of two futures contracts with the same expiration written on two separate commodities, options to exchange two spot commodities or their futures, average-price and average-strike Asian options, barrier options—which are exercised when the commodity price crosses a threshold—and American swing options for the delivery of an uncertain amount of the commodity. A discussion of these and other exotic energy derivatives is presented in Clewlow and Strickland (2000), Eydeland and Wolyniec (2003), and Geman (2005).

Exotic energy derivatives are complex to price and hedge for advanced commodity price models. Furthermore, spread derivatives depend not only on the volatility but also on the correlation between various spot/futures contracts,

which may be challenging to model and calibrate to market prices. Consequently, the development of accurate stochastic price models and pricing methods for exotic derivatives and spread options may be particularly helpful for the valuation and hedging of energy assets. Accurate analytical approximations of spread options prices and their hedge ratios are derived by Li, Deng, and Zhou (2008) for two assets that follow correlated log-OU diffusions. Extensions to multiasset spread option pricing and hedging are presented in Li, Zhou, and Deng (2010).

Shipping Derivatives

The success and rapid growth of derivatives in the energy commodity markets has spurred their introduction in the shipping markets. Shipping derivatives—forward freight agreements (FFAs) and freight futures—were introduced in 1985 and are widely used by the dry bulk and tanker shipping markets as discussed by Alizadeh and Nomikos (2009). Freight rate swaps were also recently introduced in the container ship markets. The growth of shipping derivatives is also motivated by the correlation of the supply and demand for shipping ton-miles with that of the bulk commodities transported by cargo vessels—crude oil, refined products, iron ore, and coal. An example is the recent introduction of over-the-counter iron ore swaps following the initiation of quarterly pricing of that bulk commodity. Therefore the need arises for the robust statistical modeling of the correlated forward curves of shipping and commodity markets and the pricing of shipping derivatives for use in risk management.

VALUATION AND HEDGING OF DERIVATIVES

The pricing of derivatives written on a financial security, a spot commodity, or another variable—the underlying—may be carried out by using the fundamental principles of risk-neutral valuation. When the underlying is a

nontradable—for example, temperature—an associated market price of risk process enters in the derivative price, which must be estimated from the prices of traded instruments. Otherwise, the fundamental economic insight of risk-neutral pricing and the associated mathematical techniques apply over a wide range of assets and stochastic models used for the modeling of the underlying process.

Vanilla Derivatives for Jump-Diffusions

A standard derivative pricing method for the wide class of jump-diffusion processes is based on the derivation of a risk-neutral probability measure under which European derivative prices may be expressed as conditional expectations of a payoff at a specified horizon (Duffie, 2001; Hull, 2003; Shreve, 2004; Musiela and Rutkowski, 2005). Derivative prices expressed as conditional expectations may be evaluated explicitly in the form of Fourier integrals of the complex characteristic function of the jump-diffusion by using the methods developed by Heston (1993), Carr and Madan (1998), Duffie, Pan, and Singleton (2000), and Lewis (2005). The use of this derivative pricing method in practice for the modeling of the equity-implied volatility surface and the calibration of a wide range of jump-diffusion models are discussed in Gatheral (2006).

Derivative prices expressed in the form of Fourier integrals allow the explicit evaluation of the derivative sensitivities known as the Greeks. They permit the analytical derivation of the stochastic process followed by the derivative price itself by using the Ito-Doebelin formula and often allow the explicit pricing of European derivatives with more general payoffs. The evaluation of Fourier integrals may be carried out efficiently by complex contour integration, numerical integration, or fast Fourier transform techniques.

The valuation of American options for jump-diffusions and the optimal stopping problems

that arise when early exercise is permitted is discussed in Oksendal and Sulem (2005). When the use of analytical techniques is not possible for the evaluation of American options and the determination of the early exercise boundary, the approximate method of Longstaff and Schwartz (2001), the quasi-analytical method described in Albanese and Campolieti (2006), and Monte Carlo simulation methods described in Glasserman (2004) may be used.

Exotic Derivatives for Jump-Diffusions

The valuation of a number of exotic derivatives is considerably more complex than their vanilla counterparts because their price depends on the path of the underlying process. Typical examples are barrier and Asian options. Therefore, the price of exotic derivatives is more sensitive on the structure of the underlying stochastic process than is the price of vanilla calls and puts. Consequently, the choice of the underlying process and the subsequent pricing and hedging of exotic derivatives may be a task of considerable complexity, a topic discussed for equities by Gatheral (2006).

For the geometric Brownian motion with constant drift and volatility, explicit prices of a number of exotic derivatives are derived in Shreve (2004). When the underlying process follows a jump-diffusion, the pricing of exotic derivatives by Fourier methods leads to Wiener-Hopf problems in the complex plane, the factorization of which is often possible analytically. This is the case for the jump-diffusion model of Kou (2002), which leads to the explicit valuation of barrier options. These analytical results are developed in Cont and Tankov (2004), where the class of Levy stochastic processes is also studied.

The extension of these Fourier methods to the valuation of options on spread contracts and other complex energy derivatives is discussed in London (2007). In the same reference the derivation of the characteristic functions of

a number of jump diffusion models of energy prices is presented along with the valuation of weather derivatives.

Statistical Inference of Asset Price Models

Asset price models usually contain a number of parameters that need to be estimated upon calibration of the model against market prices. This may be carried out by using the econometric techniques presented in Campbell, Lo, and MacKinlay (1997), Greene (2000), and Singleton (2006).

Stochastic models of commodity prices often contain hidden factors—stochastic trends, volatilities, and the convenience yield—which are usually modeled as diffusions. The estimation of the models may be carried out by casting the time series obtained upon discretization in state space form by using the methods presented by Durbin and Koopman (2001). The simultaneous inference of the model parameters and hidden factors may then be carried out by using dual Kalman filters and the expectations maximization algorithm presented in Haykin (2001). These statistical inference techniques may also be used for the estimation of nonlinear structural form models of power prices and shipping freight rates, which are known to depend on nonlinearities in the supply and demand schedules of the underlying markets.

Stochastic Optimal Control Methods

The availability of analytical models governing the evolution of spot commodity prices and their derivatives allows the formulation and solution of a wide range of valuation and hedging problems involving energy assets and their derivatives. The resulting stochastic dynamic programming problems are often possible to treat analytically by using the stochastic optimal control methods presented in Yong and Zhou (1999) for diffusions with

time-dependent deterministic coefficients. These results follow from the solution of the Hamilton-Jacobi-Belman (HJB) partial differential equation or the Pontryagin stochastic maximum principle and its connection to backwards stochastic differential equations. Extensions of these stochastic optimal control methods for underlying processes that follow diffusions with stochastic coefficients are discussed in Lewis (2005). Stochastic control methods for jump-diffusions and the treatment of the associated integro-differential equations are discussed in Oksendal and Sulem (2005).

APPLICATIONS

The stochastic price models, derivative valuation methods, and stochastic optimal control algorithms presented above have found widespread use in the securities markets. A number of applications drawn from the energy and shipping sectors are discussed below.

Valuation of Natural Gas and Oil Storage

Storage facilities for natural gas and oil are assets that enable the transfer of power generation capacity between two time periods in response to supply and demand fluctuations. Such fluctuations are affected by the different seasonal variations of the natural gas and electricity prices, the former usually being higher and more volatile during the winter and the latter often being a lot higher during the summer.

The availability of inexpensive gas storage facilities and the need to invest in new capacity allows the low-cost shifting of cheap summer production and storage of gas into the winter season. Moreover, the availability of gas storage facilities allows the quick delivery of natural gas when demand peaks, circumventing the need for expensive new production.

These economic drivers call for the valuation and optimal operation of storage facilities for natural gas and other fuels, in the face of stochastic gas prices, which are assumed to be unaffected by the availability of storage.

The storage valuation problem may be cast in a stochastic dynamic programming framework that relies on the analytical modeling of the commodity spot prices, futures curve, and their derivatives as outlined above. In its generality, this valuation problem reduces to the determination of optimal storage in/out-flows given the commodity seasonal price dynamics. The analytical framework for this valuation problem is presented in Eydeland and Wolyniec (2003) and discussed below in the context of the valuation of *crude oil floating storage* using a principal components factor model for the forward curve.

Valuation of Flexible Hydrocarbon Reservoirs

The optimal dynamic management of proven but undeveloped hydrocarbon reservoirs and flexible oil fields leads to a sequence of decisions analogous to those described above for above-ground storage facilities. When significant irreversible investments with option-like value are necessary for the development of flexible hydrocarbon fields, the extended valuation framework of real options is needed. Its development is presented in Dixit and Pindyck (1994), and a number of applications are discussed in Brennan and Trigeorgis (2000) and Copeland and Antikarov (2001). Given an HJM model for the oil and natural gas futures curve and its derivatives, the operation of flexible hydrocarbon fields may be reduced to a stochastic dynamic programming problem leading to the determination of optimal investment and hydrocarbon extraction flows. A number of real projects where these valuation methods are applicable are presented in Ronn (2002).

Hedging of Fuel Costs

The risk management of fuel costs in the transportation and energy sectors entails the hedging of commitments to purchase or deliver energy commodities—crude oil, natural gas, aviation jet fuel, gasoline, heating oil, and shipping bunker fuels by various entities—refineries, utilities, airlines, and shipping companies. An objective of such hedging programs is the minimization of the variance of the commodity price exposures over a given horizon. Variance-minimizing quadratic hedges of complex derivative exposures using simpler securities is common in the financial markets and may be reduced to the solution of a stochastic dynamic programming problem (Yong and Zhou, 1999; Jouini, Cvitanic, and Musiela, 2001).

A fuel cost hedging program may be implemented by using a combination of physical storage and the futures market. Such a hedging task faces a number of challenges, including commodity price and volume uncertainty, a decreasing liquidity of futures contracts of increasing tenor, an increasing volatility of futures contracts of decreasing tenor that need to be rolled over, and exposure to basis risk when liquid futures contracts for the fuel of interest do not exist. The solution of the resulting dynamic optimization problem may be carried out by taking advantage of the analytical modeling, pricing, and optimal control techniques outlined above. The complexity of such hedging programs is considerable as is highlighted by the collapse of the stacked hedges of Mettalgeshellschaft studied in Culp and Miller (1999).

Valuation of Seaborne Energy Cargoes

Crude oil and other liquid energy cargoes transported in tanker fleets may be traded while the cargo is in transit. This is akin to the optimal financial management of energy commodities in movable storage. Here, the location and speed

of the tankers enter as controls in a stochastic dynamic programming framework, which may be treated with the analytical techniques described above. The timing, sales price, and port of delivery of the energy cargo are variables that may be selected in a value-maximizing manner while the commodity is in transit. These decisions must take into consideration the shape of the oil futures curve, which may be trading in contango, backwardation, or in a composite formation, as well as the tanker freight rate forward curve. Moreover, since a large portion of the above-ground crude oil is in transit, the aggregate tonnage and average speed of crude oil tanker fleets may have a material impact upon the crude oil convenience yield, the shape of its futures term structure, and its impact on the valuation of seaborne oil.

The principal components model of the crude oil forward curve developed by Sclavounos and Ellefsen (2009) was applied by Ellefsen (2010) to the valuation of crude oil floating storage. The value of a crude oil cargo carried by a very large crude carrier (VLCC) is shown to be that of an American option with an embedded early exercise premium. The valuation of this option is carried out in a semianalytical form by virtue of the explicit form of the Ornstein-Uhlenbeck diffusions and their transition densities that govern the independent factors that drive the crude oil forward curve, using the method presented in Albanese and Campolieti (2006). It is shown that the value of the early exercise premium can be significant, particularly in volatile markets and even if the forward curve is not trading in extreme contango. The returns of crude oil floating storage investments are also studied and shown to be significant. Their hedging using crude oil futures is also addressed.

The valuation methodology developed for crude oil floating storage extends with minor modifications to land-based storage of crude oil, products, bunker fuels, natural gas, and other commodities. The necessary analytical machinery lies in the development of the principal component analysis of the forward curve

of the commodity under consideration and the analytical derivation of the diffusions governing a small number of independent factors that drive the evolution of the respective forward curves.

Analogous considerations apply to the transportation of liquefied natural gas in LNG carriers. The LNG market is not as liquid or global as the oil market, yet it is likely to mature in the future in light of the growing demand for natural gas for the generation of electricity.

Fuel-Efficient Navigation and Optimal Chartering of Shipping Fleets

The shipping industry consumes approximately 5% of the world oil production in the form of bunker fuels. Assuming a daily world oil production of 87 million barrels and a long-term price of oil of \$100 per barrel, the daily bunker fuel costs for the shipping industry are estimated at \$400 million. The long-term daily average freight rate revenue is harder to estimate and is assumed to be over twice the daily bunker fuel costs.

The selection of the optimal speed and route of cargo vessels exposed to stochastic freight rates and subject to the constraints imposed by the charter contract, cargo loading schedules, and port and other fees leads to a stochastic dynamic programming problem. The ship resistance and propulsion characteristics may be supplied by the shipowner, estimated from models or inferred from real-time measurements of the ship speed, propeller revolutions, engine performance, and the weather using the inference methods described in Haykin (2001). Using a reduced form or structural stochastic price model for the shipping freight rate forward curve, optimal routing and chartering strategies may be derived analytically aiming to minimize the fuel consumption and maximize freight rate revenue over single or consecutive voyages. A cumulative 5% reduction in bunker fuel costs and increase in freight rate revenue

would translate into a \$50 million increase in the daily net income of the shipping industry. The promise of these advanced dynamic optimization algorithms is underscored by their adoption by the aviation industry for the optimal routing of commercial jets (“Calculating Costs in the Clouds,” *The Wall Street Journal*, March 6, 2007).

Valuation and Hedging of Power Plants and Refineries

The optimal economic dispatch of power plants presents a challenging problem that depends in part on the price differential of two energy commodities. The input commodity is usually a fuel—natural gas or oil—which may be traded in the spot and forward markets. The output commodity is electricity, which cannot be stored. It trades into a spot cash market and may not have liquid forward contracts, as discussed by Joskow (2006).

In simple cases, the valuation of power generating units may be reduced to the pricing of a strip of options written on the price differentials of electricity and the input fuel, for example natural gas. Given analytical price models for the price of the input fuel and electricity, the power plant valuation and hedging problem may be based on the pricing of these spark-spread options, which may be available explicitly. In more general settings where operational constraints apply, the valuation problem may be cast in a stochastic dynamic programming framework, which may benefit from the use of the analytical modeling and hedging methods outlined above. A similar set of issues arise in the valuation and hedging of refineries that process crude oil, which has a well-developed spot and futures market, into products—gasoline, heating oil, jet fuel, bunker fuel—which often do not have actively traded forward contracts. The use of this general valuation and hedging methodology in practice is presented in Eydeland and Wolyniec (2003).

Valuation of Wind Farms and Electricity Storage Facilities

Wind is an ample, clean, renewable energy source, yet its availability is variable. Consequently the electricity generated from a wind farm varies stochastically and is a function of the statistical properties of the wind speed averaged over a certain time interval. The volatility of the annual mean wind speed is typically about 10%. The development of onshore wind farms is growing at a 25–35% rate worldwide. Offshore wind energy is the next frontier with high expected growth rates over the next several decades from the development of vast expanses of sea areas with high winds and capacity factors of 40–45% using innovative low-cost floating wind turbine technologies that may be deployed in water depths ranging from 30 to several hundred meters. An offshore wind farm with a rated capacity of 1 GW and a lifespan of 25 years is on an energy-equivalent basis comparable to a 100 million barrel oil reservoir. Moreover, this energy resource is available just 100 meters above sea level as opposed to thousands of meters below it.

The valuation of a utility scale onshore or offshore wind farm as an energy asset may be carried out using the standard weighted average cost of capital (WACC) discounted cash flow method for a constant capital structure. Alternatively the adjusted present value (APV) method may be used for a varying leverage ratio and when tax shields and other incentives available to wind farm investments must be valued separately. Wind turbines are high-value capital assets that generate steady cash flows with an annualized volatility of about 10%. Utility-scale wind farm investments may therefore be structured using nonrecourse project finance with a leverage that may reach 70–80%. The risk embedded in debt and equity securities used to finance utility-scale wind farm investments depends on technical, environmental, and market factors. Their rational modeling permits the pricing of debt and equity financial

claims at various levels of leverage. Moreover, the availability of robust long-term statistical models of the mean wind speed and shorter term jump-diffusion models of wind speed fluctuations and of the market prices of electricity discussed above allows the pricing of *structured securities* like convertible debt and other derivatives that may be used to design an optimal capital structure, hedge financial exposures of wind farms as energy assets, and determine the optimal mix between fixed PPA versus fluctuating market price contracts for the delivery of electricity.

Investments in large-scale storage facilities for electricity generated by wind farms may be economically attractive since they would permit the storage of kilowatt-hours when electricity prices are low and wind speeds are high and the sale of electricity when prices are attractive. Such large-scale storage facilities include pumped water storage in large reservoirs above ground or below sea level, compressed air storage in large underground caverns, high-capacity batteries, and compressed hydrogen in tanks following the electrolysis of seawater by onshore or offshore wind farms. The valuation and optimal operation of such storage depends on the short-term volatility and longer term fluctuations of wind speeds and electricity prices. Therefore its value is analogous to that derived from *natural gas storage*. The availability of a stochastic price model for the spot and forward electricity prices allows the explicit valuation of such storage facilities using the methods presented above. This analysis would suggest the merits, size, and optimal management of utility-scale electricity storage facilities and would guide investments in these assets.

Canonical Correlation of Commodity and Shipping Forward Curves

The principal components analysis of the term structure of interest rates and of the forward curves in the commodities markets is a pow-

erful method for the parsimonious modeling of the evolution of a large number of highly correlated spot and forward securities in the respective market. Examples of the application of this statistical modeling method were discussed above.

The forward curves of distinct energy commodities, for example, crude oil, gasoline, and gasoil, are often correlated. The same applies to the FFA forward curves of distinct routes in the dry bulk and tanker shipping markets. Therefore the development of parsimonious and robust statistical models of the correlation structure of two or more forward curves is often necessary for the valuation of assets exposed to multiple commodities. This may be accomplished by carrying out a canonical correlation analysis of the block covariance matrix of the commodity forward curves of interest. The diagonal blocks are the intracommodity covariance matrices, which may be treated by the principal component analysis discussed above. The off-diagonal blocks are the intercommodity covariance matrices, which may be reduced by the canonical correlation analysis described in Basilevsky (1994) and Anderson (2003).

In a principal components analysis a small number of dominant factors is derived for each commodity forward curve, linear combinations of the traded futures contracts of varying tenors. In a canonical correlation analysis, for example of two commodity forward curves, portfolios of futures trading on each forward curve may be derived that are maximally correlated. The maximum correlation coefficient between the two curves is a summary metric that is independent of the tenor of the futures contracts used to derive each portfolio. The extension of this method to multiple commodity forward curves is straightforward. A canonical correlation analysis allows an in-depth study of the cross-correlation structure of multiple commodity and shipping markets and may be used in the development of cross hedging strategies, in the valuation of assets, and for risk management.

The canonical correlation of the forward curves of distinct routes in the dry bulk and tanker shipping markets was carried out by Hatziyiannis (2010). This study revealed various degrees of maximal correlations between shipping routes and a surprisingly large maximal correlation between the dry bulk and tanker markets. This suggests that there exist portfolios of FFAs trading on major routes in the dry bulk market that are highly correlated with FFA portfolios in the tanker market. The composition of these portfolios follows from the canonical correlation analysis. The implication is that a few liquid forward curves in shipping may be used for the hedging of exposures in routes with less liquid derivatives. Moreover, maximally correlated portfolios of spot and forward contracts may be used for the design of broad shipping indexes across shipping sectors and routes that may spur the liquidity of shipping derivatives.

The shipping forward curve principal components and canonical correlation analysis described above may be extended to include cross-correlated energy and other bulk commodity forward curves. Combined with the powerful methods of conditional multivariate statistics coupled with robust Bayesian Stein estimators of drifts, risk management strategies of energy and shipping assets may be developed and trading strategies involving paper assets may be derived.

Pricing of Shipping Options

The arbitrage-free pricing of shipping options is carried out along lines similar to those in the energy markets. A technical complexity of shipping derivatives is that shipping options settle against the arithmetic average of the underlying spot index. Shipping options may be priced either by modeling the evolution of the underlying index, or by modeling the evolution of the underlying futures contract. The first method is prevalent to date and is discussed in Alizadeh and Nomikos (2009). Yet, the second method

has a number of advantages. By modeling the underlying futures or FFA contract as a lognormal diffusion, the pricing of calls and puts may be carried out readily by using the Black formula. Moreover, the underlying futures or FFA contracts may be used for delta, gamma, and vega hedging of options exposures.

This approach of pricing shipping options has been adopted in the multifactor principal components model of the forward curve developed by Sclavounos and Ellefsen (2009). It leads to explicit expressions of the option prices and their Greeks and also allows for a volatility term structure, which is the result of the mean reversion of the factors driving the shipping forward curves. The explicit form of the Ornstein-Uhlenbeck diffusions governing the evolution of the factors leads to explicit algebraic expressions for the options and their sensitivities discussed in Ellefsen (2010).

Pricing of Credit Risk and Structured Securities in Shipping

Shipping fleets are primarily financed by debt issued by banks and other lending institutions, followed by equity raised by shipping firms in private placements or on public exchanges. The underlying assets financed by this capital are cargo ships, which have observable prices quoted by shipping brokers. Credit derivatives and other structured securities analogous to those in widespread use in other asset markets are not yet as widely traded in the shipping sector.

The pricing of credit risk is based on the fundamental structural form firm value method of Merton (1974) and the reduced form hazard rate method of Duffie and Singleton (2003). These valuation methods have enabled the pricing of derivatives written on individual credits—for example, credit default swaps—as well as derivatives written on baskets of credits. The values of the underlying entities in a basket and their default probabilities are correlated, and this dependency structure may be

modeled in its generality by using multivariate Gaussian statistics and in non-Gaussian settings copulae functions. This financial technology has enabled the design and pricing of an array of structured financial securities discussed in Duffie and Singleton (2003), Lando (2004), London (2007), and Cherubini and Della Lunga (2007).

Shipping credit risk may be modeled and priced using a hybrid model, which blends the structural and reduced form valuation methods discussed in Ammann (2001). The price of the assets of a shipping firm—the cargo vessels—is stochastic but observable, therefore recovery at default is known. The price of equity of public shipping firms is also observable and may be used to model the hazard rate, the probability of default, and hence the pricing of shipping debt by calibrating a hybrid credit risk model as described by Overhaus et al. (2007). Cargo ship prices within and across shipping sectors are correlated, and this dependency may be modeled by identifying common underlying factors via a principal components and canonical correlation analysis. The above attributes of the shipping sector may be introduced to price loans, convertible bonds, equity and credit linked notes, and other structured securities, which may be used to better manage shipping risk, reduce bank regulatory risk capital, and make available new and innovative sources of financing to shipowners.

KEY POINTS

- Producers of energy commodities and owners of shipping tonnage may take short positions in futures and freight forward agreements (FFAs) in order to hedge their forward delivery commitments against a decrease of prices.
- Consumers of energy commodities and shippers who charter cargo vessels may take long positions in futures and FFAs in order to hedge their forward commodity and freight rate exposures against rising prices.
- Power plants and refineries that transform an input commodity into an output commodity, for example, natural gas to electricity, crude oil to gasoline, may go long the futures of the input commodity and short the futures of the output commodity in order to protect their profit margins against adverse moves of the input/output commodity price spread.
- Liquid energy commodity forward curves convey information about the stochastic evolution of the spot price of the commodity.
- The stochastic dynamics of individual energy commodity and shipping forward curves may be modeled by a small number of independent statistical factors using a principal components analysis (PCA). The factors are portfolios of traded futures contracts, and their stochastic dynamics are governed by diffusions that may be derived in explicit form.
- The joint stochastic dynamics of cross-correlated commodity and shipping forward curves may be modeled by a small number of statistical factors using an intracommodity PCA curve and an intercommodity canonical correlation analysis (CCA).
- The parsimonious statistical factor modeling of the commodity and shipping forward curves may be used for the valuation and risk management of energy assets, structured securities, and portfolios of commodity and shipping derivatives.

REFERENCES

- Albanese, C., and Campolieti, G. (2006). *Advanced Derivatives Pricing and Risk Management*. Burlington, MA: Elsevier Academic Press.
- Alizadeh, A. H., and Nomikos, N. K. (2009). *Shipping Derivatives and Risk Management*. London: Palgrave Macmillan.
- Ammann, M. (2001). *Credit Risk Valuation*. Berlin: Springer Verlag.
- Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons.
- Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods: Theory and Applications*. New York: John Wiley & Sons.

- Bength, F. E., Bength, J. S., and Koekebakker, S. (2008). *Stochastic Modeling of Electricity and Related Markets*. Advanced Series of Statistical Science and Applied Probability. Singapore: World Scientific.
- Brennan, M. J., and Trigeorgis, L. (2000). *Project Flexibility, Agency and Competition: New Developments in the Theory and Application of Real Options*. Oxford: Oxford University Press.
- Campbell, J. Y., Lo, W. A., and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton: Princeton University Press.
- Carr, P., and Madan, D. (1998). Option valuation using the fast Fourier transform. *Journal of Computational Finance* 2: 61–73.
- Cherubini, U., and Della Lunga, G. (2007). *Structured Finance*. New York: John Wiley & Sons.
- Clelow, L., and Strickland, C. (2000). *Energy Derivatives: Pricing and Risk Management*. London: Lacima Publications.
- Cont, R., and Tankov, P. (2004). *Financial Modeling with Jump Processes*. Boca Raton, FL: Chapman & Hall/CRC.
- Copeland, T., and Antikarov, V. (2001). *Real Options: A Practitioner's Guide*. London: Monitor Group, TEXERE.
- Cortazar, G. and Naranjo, L. (2006). An N-factor Gaussian model of oil futures prices. *Journal of Futures Markets*, 26: 243–268.
- Culp, C. L., and Miller, M. H. (1999). *Corporate Hedging in Theory and Practice: Lessons from Metallgesellschaft*. London: RISK Books.
- Dixit, A. K., and Pindyck, R. S. (1994). *Investment under Uncertainty*. Princeton: Princeton University Press.
- Duffie, D. (2001). *Dynamic Asset Pricing Theory*, 3rd Edition. Princeton: Princeton University Press.
- Duffie, D., Pan, J., and Singleton, K. J. (2000). Transform analysis and asset pricing for affine jump diffusions. *Econometrica* 68: 1343–1376.
- Duffie, D., and Singleton, K. J. (2003). *Credit Risk: Pricing, Measurement and Management*. Princeton: Princeton University Press.
- Durbin, J., and Koopman, S. J. (2001). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.
- Ellefsen, P. E. (2010). Commodity market modeling and physical trading strategies. Master's thesis. Department of Mechanical Engineering, Massachusetts Institute of Technology.
- Eydeland, A., and Wolyniec, K. (2003). *Energy and Power Risk Management: New Developments in Modeling Pricing and Hedging*. New York: John Wiley & Sons.
- Gatheral, J. (2006). *The Volatility Surface: A Practitioner's Guide*. New York: John Wiley & Sons.
- Geman, H. (2005). *Commodities and Commodity Derivatives: Modeling and Pricing for Agriculturals, Metals and Energy*. New York: John Wiley & Sons.
- Gibson, R., and Schwartz, E. S. (1990). Stochastic convenience yield and the pricing of oil contingent claims. *The Journal of Finance* 45: 959–976.
- Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. Berlin: Springer Verlag.
- Greene, W. H. (2000). *Econometric Analysis*, 4th Edition. Upper Saddle River, NJ: Prentice Hall.
- Hatziyiannis, N. (2010). Canonical correlation of shipping forward curves. Master's thesis. Department of Mechanical Engineering, Massachusetts Institute of Technology.
- Haykin, S. (2001). *Kalman Filtering and Neural Networks*. New York: John Wiley & Sons.
- Heston, S. (1993). A closed form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies* 6: 327–343.
- Hull, J. C. (2003). *Options, Futures and Other Derivatives*, 5th Edition. Upper Saddle River, NJ: Prentice Hall.
- Joskow, P. L. (2006). Competitive electricity markets and investment in new generating capacity. MIT Working Paper. Center for Energy and Environmental Policy Research.
- Jouini, E., Cvitanic, J., and Musiela, M. (2001). *Option Pricing, Interest Rates and Risk Management*. Cambridge: Cambridge University Press.
- Kou, S. (2002). A jump-diffusion model for option pricing. *Management Science* 48: 1086–1101.
- Lando (2004). *Credit Risk Modeling: Theory and Applications*. Upper Saddle River, NJ: Princeton University Press.
- London, J. (2007). *Modeling Derivatives Applications*. London: Financial Times Press.
- Longstaff, F. A., and Schwartz, E. S. (2001). Valuing American options by simulation: A simple least-squares approach. *Review of Financial Studies* 14: 113–147.
- Lewis, A. L. (2005). *Option Valuation under Stochastic Volatility*. Newport Beach, CA: Finance Press.
- Li, M., Deng, S-J., and Zhou, J. (2008). Closed-form approximations for spread option prices and Greeks. *Journal of Derivatives* 15, 3: 58–80.
- Li, M., Zhou, J., and Deng, S-J. (2010). Multi-asset spread option pricing and hedging. *Quantitative Finance* 10, 3: 305–324.

- Merton, R. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29: 449–470.
- Miltersen, K. R., and Schwartz, E. S. (1998). Pricing of options on commodity futures with stochastic term structures of convenience yields and interest rates. *Journal of Financial and Quantitative Analysis* 33: 61–86.
- Musiela, M., and Rutkowski, M. (2005). *Martingale Methods in Financial Modelling*. Berlin: Springer Verlag.
- Oksendal, B., and Sulem, A. (2005). *Applied Stochastic Control of Jump Diffusions*. Berlin: Springer Verlag.
- Overhaus, M., Bermudez, A., Buehler, H., Ferraris, A., Jordinson, C., and Lamnouar, A. (2007). *Equity Hybrid Derivatives*. New York: John Wiley & Sons.
- Ronn, E. I. (2002). *Real Options and Energy Management*. London: RISK Publications.
- Ross, S. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13: 343–362.
- Samuelson, P. (1965). Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review* 6: 41–49.
- Schwartz, E. S. (1997). The stochastic behavior of commodity prices: Implication for valuation and hedging. *Journal of Finance* 3: 923–973.
- Schwartz, E. S., and Smith, J. E. (2000). The stochastic behavior of commodity prices: Implications for valuation and hedging. *Management Science* 46: 893–911.
- Sclavounos, P. D., and Ellefsen, P. E. (2009). Multi-factor model of correlated commodity forward curves for crude oil and shipping markets. Working Paper. Center for Environmental and Energy Policy Research (CEEPR). Massachusetts Institute of Technology, Sloan School of Management <http://web.mit.edu/ceepr/www/publications/workingpapers.html>
- Shreve, S. E. (2004). *Stochastic Calculus for Finance II. Continuous-Time Models*. Berlin: Springer Verlag.
- Singleton, K. J. (2006). *Empirical Dynamic Asset Pricing: Model Specification and Econometric Assessment*. Princeton: Princeton University Press.
- Yong, J., and Zhou, X. Y. (1999). *Stochastic Controls*. Berlin: Springer Verlag.

Index

- Absence of arbitrage principle, *I:99, I:127*. *See also* arbitrage, absence of
- ABS/MBS (asset-backed securities/mortgage-backed securities), *I:258–259, I:267*
- cash flow of, *III:4*
- comparisons to Treasury securities, *III:5*
- modeling for, *III:536*
- Accounting, *II:532, II:542–543*
- Accounting firms, watchdog function of, *II:542*
- Accounts receivable turnover ratio, *II:557–558*
- Active-passive decomposition model, *III:17, III:19–22, III:26*
- Activity ratios, *II:557–558, II:563*
- Adapted mesh, one year to maturity, *II:680f*
- Adjustable rate mortgages (ARMs). *See* ARMs (adjustable rate mortgages)
- Adjustments for changes in net working capital (ANWC), *II:25*
- Adverse selection, *III:76*
- Affine models, *III:554–557*
- Affine process, basic, *I:318–319, I:334n*
- Agency ratings, and model risk, *II:728–729*
- Airline stocks, *II:249–250, II:250f, II:250t, II:252t*
- Akaike Information Criterion (AIC), *II:703, II:717*
- Algorithmic trading, *II:117*
- Algorithms, *II:676–677, II:701–702, III:124*
- Allied Products Corp., cash flow of, *II:576*
- α -stable densities, *III:243f, III:244f*
- α -stable distributions
- defined, *II:738*
- discussion of, *III:233–238*
- fitting techniques for, *II:743–744*
- properties of, *II:739*
- simulations for, *II:750*
- subordinated representation of, *II:742–743*
- usefulness of, *III:242*
- and VaR, *II:748*
- variables with, *II:740*
- α -stable process, *III:499*
- Alternative risk measures proposed, *III:356–357*
- Amazon.com
- cash flows of, *II:568, II:568t*
- American International Group (AIG), stock prices of, *III:238*
- Amortization, *II:611, III:72–73*
- Analysis
- and Barra model, *II:244–248*
- bias in, *II:109*
- common-size, *II:561–563*
- crisis-scenario, *III:379–380*
- to determine integration, *II:514*
- formulas for quality, *II:239*
- fundamental, *II:243, II:248, II:253–254*
- interpretation of results, *III:42–44*
- mathematical, *I:18*
- model-generated, *III:41–42*
- multivariate, *II:48*
- statistical, *I:140, II:353–354*
- sum-of-the-parts, *II:43–44*
- vertical *vs.* horizontal common-size, *II:562*
- Analytics, aggregate, *II:269t*
- Anderson, Philip W., *III:275*
- Annual percentage rate (APR), *II:598, II:615–616*
- Annual standard deviation, *vs.* volatility, *III:534*
- Annuities
- balances in deferred, *II:610f*
- from bonds, *I:211–212*
- cash flows in, *II:604–607*
- future value factor, *II:605–606*
- ordinary, *II:605*
- present value factor, *II:605, II:606–607*
- valuation of due, *II:608–609*
- valuing deterred, *II:609–611*
- Anticipation, in stochastic integrals, *III:475*
- Approximation, quality of, *II:330–331*
- APT (arbitrage pricing theory), *I:116*
- Arbitrage
- absence of, *I:56, I:135, II:473*
- in continuous time, *I:121–123*
- convertible bond, *I:230*
- costless profits, *I:442*
- costless trades, *I:428t*
- defined, *I:99, I:119, I:123*
- in discrete-time, continuous state, *I:116–119*
- and equivalent martingale measures, *I:111–112*
- in multiperiod finite-state setting, *I:104–114*
- in one-period setting, *I:100–104*
- pricing of, *I:124, I:134–135, II:476*
- profit from, *I:221–222*
- and relative valuation models, *I:260*
- and state pricing, *I:55–56, I:102, I:130*
- test for costless profit, *I:441*
- trading strategy with, *I:105*
- types of, *I:55–56*
- using, *I:70–71*
- Arbitrage-free, *III:577, III:593–594*
- Arbitrage opportunities, *I:55, I:56, I:100, I:117, I:260–261, I:437*
- Arbitrage pricing theory (APT), *I:116*
- application of, *I:60–61*
- development of, *II:468, II:475–476*
- factors in, *II:138*
- key points on, *II:149–150*
- and portfolio optimization, *I:40*

- ARCH (autoregressive conditional heteroskedastic) models and behavior of errors, *II:362*
 defined, *I:176*
 in forecasting, *II:363*
 reasons for, *III:351*
 type of, *II:131*
 use of, *II:733–734*
- ARCH/GARCH models
 application to VaR, *II:365–366*
 behavior of, *II:361–362*
 discussion of, *II:362–366*
 generalizations of, *II:367–373*
 usefulness of, *II:366–367*
- ARCH/GARCH processes, *III:277*
- Area, approximation of, *II:589–590*, *II:589f*
- ARIMA (autoregressive integrated moving average) process, *II:509–510*
- ARMA (autoregressive moving average) models
 defined, *II:519*
 and Hankel matrices, *II:512*
 linearity of, *II:402*
 and Markov coefficients, *II:512*
 multivariate, *II:510–511*, *II:513–514*
 nonuniqueness of, *II:511*
 representations of, *II:508–512*
 and time properties, *II:733*
 univariate, *II:508–510*
- ARMA (autoregressive moving average) processes, *III:276–277*
- ARMs (adjustable rate mortgages), *III:25*, *III:71–72*, *III:72f*, *III:74*
- Arrays, in MATLAB and VBA, *III:420–421*, *III:457–458*, *III:466*
- Arrow, Kenneth, *II:467*, *II:699*
- Arrow-Debreu price, *I:53–55*. *See also* state price
- Arrow-Debreu securities, *I:458*, *I:463*
- Arthur, Bryan, *II:699*
- Artificial intelligence, *II:715*
- Asian fixed calls, with finite difference methods, *II:670t*
- Asian options, pricing, *III:642–643*
- Asset allocation
 advanced, *I:36*
 building blocks for, *I:38*
 modeling of, *I:42*
 standard approach to, *I:37–38*
- Asset-backed securities (ABS), *I:258*
- Asset-liability management (ALM), *II:303–304*, *III:125–126*
- Asset management, focus of, *I:35*
- Asset prices
 codependence of, *I:92*
 multiplicative model for, *I:86–87*, *I:88*
 negative, *I:84*, *I:88*
 statistical inference of models, *I:560*
- Asset pricing, *I:3*, *I:56–59*, *I:59–60*, *I:65–66*, *II:197*
- Asset return distributions, skewness of, *III:242*
- Asset returns
 characteristics of, *III:392*
 errors in estimation of, *III:140–141*
 generation of correlated, *I:380–381*
 log-normal distribution applied to, *III:223–225*
 models of, *III:381*
 normal distribution of, *I:40*
 real-world, *III:257*
 simulated vector, *I:380–381*
- Assets
 allocation of, *I:10*
 on the balance sheet, *II:533–534*
 carry costs, *I:424–425*
 correlation of company, *I:411*
 current *vs.* noncurrent, *II:533*
 deliverable, *I:483*
 discrete flows of, *I:425–426*
 expressing volatilities of, *III:396–397*
 financing of, *II:548*
 funding cost of, *I:531*
 future value of, *I:426t*, *I:427t*
 highly correlated, *I:192*
 intangible, *II:534*
 liquid, *II:551*
 management of, *II:558*
 market prices of, *I:486*
 new fixed, *II:25*
 prices of, *I:60*
 redundant, *I:51*
 representation of, *II:515*
 risk-free, *I:112–113*
 risky *vs.* risk-free, *I:5–6*
 shipping, *I:555*
 storage of physical, *I:439*, *I:442–443*, *I:560–561*
 values of after default events, *I:350*
- Asset swaps, *I:227–230*
- Assumptions
 about noise, *II:126*
 under CAPM, *I:68–69*
 errors in, *III:399*
 evaluation of, *II:696*
 homoskedasticity *vs.* heteroskedasticity, *II:360*
 importance of, *III:62*
 for linear models, *II:310–311*
 for linear regression models, *II:313*
 in scenario analysis, *II:289*
 simplification of, *III:397*
 using inefficient portfolio analysis, *I:288t*
 violations of, *I:475*
 zero mean return, *III:397*
- Attribution analysis, *II:188–189*
- AT&T stock, binomial experiment, *I:146–148*
- Audits, of financial statements, *II:532*
- Augmented Dickey-Fuller test (ADF), *II:387*, *II:389*, *II:390t*, *II:514*
- Autocorrelation, *II:328–329*, *II:503*, *II:733*
- Autoregressive conditional duration (ACD) model, *II:370*
- Autoregressive conditional heteroskedastic (ARCH) models. *See* ARCH (autoregressive conditional heteroskedastic) models
- Autoregressive integrated moving average (ARIMA) process, *II:509–510*
- Autoregressive models, *II:360–362*
- Autoregressive moving average (ARMA) models. *See* ARMA (autoregressive moving average) models
- AVaR. *See* average value at risk (AVaR)
- Average credit sales per day, calculation of, *II:553*
- Average daily volume (ADV), *II:63*
- Averages, equally weighted, *III:397–409*
- Average value at risk (AVaR) measure
 advantages of, *III:347*
 back-testing of, *III:338–340*
 boxplot of fluctuation of, *III:338f*
 and coherent risk measures, *III:333–334*
 computation of in practice, *III:336–338*
 computing for return distributions, *III:334–335*
 defined, *III:331–335*
 estimation from sample, *III:335–336*
 and ETL, *III:345–347*
 geometrically, *III:333f*
 graph of, *III:347f*
 higher-order, *III:342–343*
 historical method for, *III:336–337*
 hybrid method for, *III:337*
 minimization formula for, *III:343–344*
 Monte Carlo method for, *III:337–338*
 with the multivariate normal assumption, *III:336*
 of order one, *III:342–343*
 for stable distributions, *III:344–345*
 tail probability of, *III:332–333*
- Axiomatic systems, *III:152–153*

- Bachelier, Louis, *II:121–122, II:467, II:469–470, III:241–242, III:495*
- Back propagation (BP), *II:420*
- Back-testing
- binomial (Kupiec) approach, *III:363*
 - conditional testing (Christoffersen), *III:364–365*
 - diagnostic, *III:367–368*
 - example of, *II:748–751*
 - exceedance-based statistical approaches, *III:362–365*
 - in-sample *vs.* out-sample, *II:235–236*
 - need for, *III:361–362*
 - statistical, *III:362*
 - strengths/weaknesses of
 - exceedance-based, *III:365*
 - tests of independence, *III:363–364*
 - trading strategies, *II:236–237*
 - use of, *III:370*
 - using normal approximations, *III:363*
 - of VaRs, *III:365–367*
- Backward induction pricing technique, *III:26*
- Bailouts, *I:417*
- Balance sheets
- common-size, *II:562, II:562f*
 - information in, *II:533–536*
 - sample, *II:534t, II:546t*
 - structure of, *II:536*
 - XYZ, Inc. (example), *II:29t*
- Balls, drawing from urn, *III:174–177, III:175f, III:179–180*
- Bandwidth, *II:413–414, II:746*
- Bank accounts, and volatility, *III:472*
- Bank for International Settlements (BIS), definition of operational risk, *III:82*
- Bankruptcy, *I:350, I:366–369, II:577*
- Banks, use of VaR measures, *III:295*
- Barclays Global Risk Model, *II:173, II:193n, II:268*
- Barra models
- E3, *II:256, II:257t, II:261*
 - equity, *II:245–246*
 - fundamental data in, *II:246t*
 - fundamental factor, *II:244–248, II:248–250*
 - risk, *II:256*
 - use of, *II:254n*
- Barrier options, *II:683*
- Basel II Capital Accord, on operational risk, *III:86–87*
- Basic earning power ratio, *II:547, II:549*
- Bayes, Thomas, *I:140, I:196*
- Bayesian analysis
- empirical, *I:154–155*
 - estimation, *I:189*
 - hypothesis comparison, *I:156–157*
 - in parameter estimation, *II:78*
 - and probability, *I:140, I:148*
 - steps of decision making in, *I:141*
 - testing, *I:156–157*
 - use of, *I:18*
- Bayesian inference, *I:151, I:157–158, II:719*
- Bayesian Information Criterion (BIC), *II:703, II:717*
- Bayesian intervals, *I:156, I:170*
- Bayesian methods, and economic theory, *III:142*
- Bayes' theorem, *I:143–148, I:152*
- Behaviors, patterns of, *II:707–710, III:34–35*
- BEKK(1,1,K) model, *II:372*
- Beliefs
- about long-term volatility, *III:408–409*
 - posterior, *I:151–152*
 - prior, *I:152, I:159*
- Bellman's principle, *II:664–665*
- Benchmarks
- choice of, *II:114–115*
 - effect of taxes on, *II:74*
 - fair market, *III:626*
 - modeling of, *II:696*
 - portfolio, *II:272t*
 - for risk, *II:265, III:350, III:354–355*
 - risk in, *II:259*
 - tracking of, *II:67*
 - for trades, *II:117, III:624*
 - use of, *I:41–42, II:66–69*
- Benchmark spot rate curves, *I:222–223*
- Berkowitz transformation, application of, *III:366–367, III:368*
- Bernoulli model, parameter inference in, *II:726–727*
- Bernoulli trials, *I:81, III:170, III:174*
- Bessel function of the third kind, *III:232*
- Best bids/best asks, *II:449–450*
- Best practices, *I:416*
- Beta function, *III:222*
- Betas
- β_{1963} , *I:74–75*
 - β_{1964} , *I:75*
 - β_{1963} *vs.* β_{1964} , *I:76–77*
 - distribution of, *III:222*
 - meanings of, *I:74*
 - in portfolios, *II:273*
 - pricing model, *I:60–61, I:71–72*
 - propositions about, *I:75–77*
 - robust estimates of, *II:442–443*
 - in SL-CAPM models, *I:66–67*
 - two beta trap, *I:74–77*
- Bets, unintended, *II:261, II:263–264, II:264, II:265*
- Better building blocks, *I:36*
- Bias
- from data, *II:204*
 - discretization error, *III:641*
 - estimator, *III:641*
 - survivorship (look-ahead), *II:202, II:204, II:712–713, II:718*
- Bid-ask bounce, *II:455–457*
- Bid-ask spread
- aspects of, *III:597*
 - average hourly, *II:454f*
 - defined, *II:454*
 - under market conditions, *II:455f*
 - risk in, *III:372*
- Binomial experiment, *I:146–148*
- Black, Fischer, *II:468, II:476*
- Black and Scholes
- assumptions of, *I:510*
- Black-Derman-Toy (BDT) model
- defined, *I:492*
 - discussion of, *III:608–609*
 - features of, *III:549*
 - interest rate model, *III:616f*
 - as no arbitrage model, *III:604*
 - use of, *III:300*
- Black-Karasinski (BK) model, *III:548, III:607–608*
- binomial lattice, *III:611*
 - defined, *I:493*
 - features of, *III:604*
 - forms of, *III:600t*
 - interest rate trinomial lattice, *III:615f*
 - trinomial lattice, *III:616f*
- Black-Litterman model
- assumptions with, *I:196–197*
 - derivation of, *I:196–197*
 - discussion of, *I:195–201*
 - with investor's views and market equilibrium, *I:198–199*
 - mixed estimation procedure, *I:200*
 - use of for forecasting returns, *I:193–194, II:112*
 - use of in parameter estimation, *II:78*
 - variance of, *I:200*
- Black-Scholes formula
- for American options, *II:674*
 - with change of time, *III:522, III:524–525*
 - and diffusion equations, *II:654*
 - and Gaussian distribution, *II:732*
 - and Girsanov's theorem, *I:132–133*
 - statistical concepts for, *III:225*
 - use of, *I:126–127, I:136*
 - use of in MATLAB, *III:423–427, III:447*
 - use of with VBA, *III:462–463*
 - and valuation models, *I:271*
- Black-Scholes-Merton stock option pricing formula, *I:557*

- Black-Scholes model
 assumptions of, *I:512, III:655*
 and calibration, *II:681–682*
 for European options, *II:660–662, III:639–640*
 and hedging, *I:410*
 and Merton's model, *I:343*
 for pricing options, *I:487, I:509–510, I:522*
 usefulness of, *I:475*
 use of, *I:272*
 volatility in, *III:653*
- Black volatility, *III:548, III:550*
- Bohr, Niels, *I:123*
- Bond-price valuation model, *III:581–583*
- Bonds
 analytical models for, *I:271–273*
 annuities from, *I:211–212*
 calculating yields on, *II:618*
 callable, *I:24f, I:244–245, III:302–303, III:302f*
 capped floating rate, valuation of, *I:249f*
 changes in prices, *I:373–374*
 computing accrued interest and clean price of, *I:214–215*
 convertible, *I:230, I:271*
 corporate, *I:279, III:598–599*
 coupon-paying, *III:584–586*
 default-free, *I:223*
 determination of value of, *I:211–213*
 discount, *I:212*
 effective duration/convexity of, *I:255, I:256f*
 European convertible, *I:272*
 in European-style calls, *I:440*
 floating-coupon, *I:246–248, I:247f*
 floating-rate callable capped, *I:248*
 floating valuation, *I:253f*
 full (dirty) price, *I:214, I:370*
 futures contracts on, *I:498*
 general principles of valuation, *I:209–216*
 inflation-indexed, *I:278, I:279, I:283–290, I:290–294*
 input information for example, *III:613t*
 interest rate tree for, *I:244f*
 loading of specific, *II:279*
 modeling prices of, *I:490–494*
 and modified or effective duration, *III:299*
 nonpar, *I:232n*
 option-free, *I:241f, I:243*
 options on, *I:252–253, I:498–501, I:501–502*
 planned amortization class (PAC), *III:6*
 plot of convertible functions, *I:273f*
 prediction of yield spreads, *II:336–344*
 price/discount rate relationship, *I:215–216, I:215f*
 prices of, *I:213–214, I:278, I:382, II:727–728*
 prices with effective duration/convexity, *III:300t, III:301t*
 pricing for, *I:498–503, III:588*
 putable, effective duration of, *III:303–304, III:304f*
 regression data for spread application, *II:338–343t*
 relation to CDSs, *I:525–526*
 risk-free, *I:316*
 risk-neutral, *III:586*
 risk-neutral/equilibrium models for, *III:597–598*
 security levels of, *I:375t*
 spreads over time, *I:402f*
 straight, duration of, *III:301–302, III:301f*
 time path of, *I:216*
 valuation of, *I:213–215, I:216–223, I:223, II:730, III:576*
 valuing of, *I:213–214, I:244–246, I:246f*
 volatility of, *I:279*
- Book value, of companies, *II:535*
- Bootstrapping
 parametric, *II:428*
 of spot rate curve, *I:217–220*
 technique for, *II:711–712, II:746*
 usefulness of, *III:325*
 use of, *I:223, III:408*
- Borel functions, *III:508–509*
- Borel measures, *III:199, III:498*
- Borrowers, *III:5, III:70–71, III:74–75, III:598–600*
- Borrowing, *I:72–73, I:479–480*
- Boundary conditions, *II:660*
 need for, *II:661*
- Box-and-whiskers diagrams, use of, *III:329–330n*
- Boxplots, use of, *III:329–330n*
- Brennen-Schwartz model, *III:549*
- Brown, Robert, *III:476*
- Brownian motion, geometric (GBM), *I:95, III:656*
- Brownian motion (BM)
 arithmetic, *I:125, III:492, III:503*
 in binomial models, *I:114*
 bounds of, *III:473–474*
 canonical, *III:478*
 conditions defining, *III:483n*
 defined, *I:95, III:476–479*
 with drift, *III:491*
 early work on, *II:470*
 excursion of, *III:480*
 fractal properties of, *III:479–480, III:480f*
 generated by random walk, *III:479f*
 generating paths for in VBA, *III:463–465*
 geometric, *III:492–493, III:503, III:524*
 and Girsanov's theorem, *I:131, I:263*
 in Ito processes, *III:487–488*
 in Ito's formula, *III:488–489*
 and the Merton model, *I:306*
 one-dimensional standard, *III:477–478*
 paths of, *III:486, III:501–502, III:502f*
 path with deviation zones, *III:537f*
 process of, *I:269–270n*
 properties of, *III:479–481, III:501, III:536*
 in randomness calculations, *III:534–535*
 and stochastic integrals, *III:473*
 time-changed, *III:503–505*
 usefulness of, *III:495–496*
 use of, *I:262*
 variants of, *III:506*
- Bubbles, discovering, *II:396–399*
- Burmeister-Ibbotson-Roll-Ross (BIRR) model, *II:140*
- Burnout effect, *III:17–18, III:24, III:74*
- Burnout factor
 initializing of, *III:22*
- Business cycles, *I:351–352, I:408, II:430–431, II:432–433*
- Businesses, correlation within sectors, *I:411*
- Butterfly's wings, effect of, *II:645*
- Calculus, stochastic, *I:94–97*
- Calendarization, *II:43, II:487–488*
- Calibration
 of derivatives, *I:494*
 effect of, *III:619*
 under GIG model, *II:524*
 of local volatility, *II:681–685*
 need for, *III:604*
 to short forward curve, *III:545–546*
- Callable bonds, *I:462*
- Call options
 defined, *I:439*
 discrepancy measures across maturities of, *II:525t*
 early exercise of American-style, *I:442–443, I:449–450*
 European, *I:125, I:127–129, I:501, I:511, II:522–525, II:679*
 1998 prices of, *II:524t*
 value of, *I:259*

- Calls
 American-style, *I:441–442, I:449*
 error on value of, *II:668f*
 European-style, *I:440–441, I:448–449, I:448t*
- Canonical correlation analysis, *I:556*
- Capital asset pricing model (CAPM).
See CAPM (capital asset pricing model)
- Capital expenditures coverage ratio,
II:575–576
- Capital gains, taxes on, *II:73*
- Caplets, *I:249, III:589–590*
- CAPM
 multifactor, *II:475*
- CAPM (capital asset pricing model).
See also Roy CAPM; SL-CAPM
 application of, *I:60–61*
 areas of confusion, *I:67–68*
 for assessing operational risk,
III:92–93
 in asset pricing, *II:474*
 defined, *I:394*
 and discount factor model, *I:65–66*
 and investor risk, *I:73–74*
 using assumptions under, *I:68–69*
- Caps
 defined, *I:248–251*
 value of, *I:248, III:552–553*
 valuing of, with floors, *I:249–250, I:256*
- Carry, *I:423–426, I:481*
- Carry costs, *I:424–425, I:426, I:435, I:437–438, I:455n, I:481. See also* net cost of carry
- CART (classification and regression trees)
 defined, *II:375*
 example, input variables for, *II:379t*
 example, out-of-sample performance, *II:381t*
 fundamentals of, *II:376–377*
 in stock selection, *II:378–381*
 strengths and weaknesses of,
II:377–378
 uses of, *II:381*
- Cash-and-carry trade, *I:480, I:481, I:487*
- Cash concept, *II:567*
- Cash flows
 accounting for, *III:306*
 analysis of, *II:574–577, III:4–5*
 for bond class, *III:9t*
 of bonds, *I:211*
 cash flow at risk (CFaR), *III:376–378*
 classification of, *II:567*
 defined, *I:209–210, II:539, III:4*
 direct *vs.* indirect reporting method,
II:567
 discounted, *I:225*
 discrete, *I:429*
 distribution analysis *vs.* benchmark,
III:310
 estimation of, *I:209–210, II:21–23*
 expected, *I:211*
 factors in, *III:31–32, III:377*
 form residential mortgage loans,
III:62
 futures *vs.* forwards, *I:431t*
 future value of, *II:603f*
 influences on, *III:44*
 interest coverage ratio of, *II:561, II:575*
 interim, *I:482*
 for loan pool, *III:9t*
 measurement of, *II:565–566, III:14*
 monthly, *III:52–54, III:53t*
 net free (NFCF), *II:572–574, II:578*
 in OAS analysis, *I:259*
 perpetual stream of, *II:607–608*
 sources of, *II:540–541, II:569t*
 in state dependent models,
I:351–352
 statement of, *II:539–541, II:566–567*
 time patterns of, *II:607–611*
 and time value of money, *II:595–596*
 time value of series of, *II:602–607*
 for total return receivers, *I:542*
 for Treasuries, *I:219, III:564–565*
 types of in assessing liquidity risk,
III:378
 use of information on, *II:576–577*
 valuation of, *II:618–619*
vs. free cash flow, *II:22–23*
- Cash flow statements
 example of, *II:541*
 form of, *II:26t*
 information from, *II:577–578*
 reformatting of, *II:569t*
 restructuring of, *II:568*
 sample, *II:547t*
 use of, *II:24–26*
- Cash flow-to-debt ratio, *II:576*
- Cash-out refinancing, *III:66, III:69*
- Cash payments, *I:486–487, III:377*
- Categorizations, determining usefulness of, *II:335*
- Cauchy, Augustin, *II:655*
- Cauchy initial value problem, *II:655, II:656, II:656f, II:657*
- CAViAR (conditional autoregressive value at risk), *II:366*
- CDOs (collateralized debt obligations), *I:299, I:525, III:553, III:645*
- CDRs (conditional default rates)
 in cash flow calculators, *III:34*
 defaults measured by, *III:58–59*
 defined, *III:30–31*
 monthly, *III:62t*
 projections for, *III:35f*
 in transition matrices, *III:35f*
- CDSs (credit default swaps)
 basis, *I:232*
 bids on, *I:527*
 cash basis, *I:402*
 discussion of, *I:230–232*
 fixed premiums of, *I:530–531*
 hedging with, *I:418*
 illustration of, *I:527*
 initial value of, *I:538*
 maturity dates, *I:526*
 payoff and payment structure of,
I:534f
 premium payments, *I:231f, I:533–535*
 pricing models for, *I:538–539*
 pricing of by static replication,
I:530–532
 pricing of single-name, *I:532–538*
 quotations for, *I:413*
 risk and sensitivities of, *I:536–537*
 spread of, *I:526*
 unwinding of, *I:538*
 use of, *I:403, I:413, II:284*
 valuation of, *I:535–536*
 volume of market, *I:414*
- Central limit theorem
 defined, *I:149n, III:209–210, III:640*
 and the law of large numbers,
III:263–264
 and random number generation,
III:646
 and random variables, *II:732–733*
- Central tendencies, *II:353, II:354, II:355*
- Certainty equivalents, *II:723–724, II:724–725*
- CEV (constant elasticity of variance),
III:550, III:551f, III:654–655
- Chambers-Mallows-Stuck generator,
II:743–744
- Change of measures, *III:509–517, III:516t*
- Change of time methods (CTM)
 applications of, *III:522–527*
 discussion of, *III:519–522*
 general theory of, *III:520–521*
 main idea of, *III:519–520, III:527*
 in martingale settings, *III:522–523*
 in stochastic differential equation setting, *III:523*
- Chaos, defined, *II:653*
- Chaos: *Making a New Science* (Gleick),
II:714
- Characteristic function
vs. probability density function,
II:743

- Characteristic lines, *II:316, II:318t, II:344–348, II:345–347t*
- Chebyshev inequalities, *III:210, III:225*
- Chen model, *I:493*
- Chi-square distributions, *I:388–389, III:212–213*
- Cholesky factor, *I:380*
- Chow test, *II:336, II:343, II:344, II:350*
- CID (conditionally independent defaults) models, *I:320, I:321–322, I:333*
- CIR model, *I:498, I:500–501, I:502*
- Citigroup, *I:302, I:408f, I:409f*
- CLA (critical line algorithm), *I:73*
- Classes
criteria for, *II:494*
- Classical tempered stable (CTS) distribution, *II:741–742, II:741f, II:742f, II:743–744, III:512*
- Classification, and Bayes' Theorem, *I:145*
- Classification and regression trees (CART). *See* CART (classification and regression trees)
- Classing, procedure for, *II:494–498*
- Clearinghouses, *I:478*
- CME Group, *I:489–490*
- CMOs (collateralized mortgage obligations), *III:598, III:645*
- Coconut markets, *I:70*
- Coefficients
binomial, *III:171, III:187–191*
of determination, *II:315*
estimated, *II:336–337*
- Coherent risk measures, *III:327–329* and VaR, *III:329*
- Coins, fair/unfair, *III:169, III:326–327*
- Cointegrated models, *II:503*
- Cointegration
analysis of, *II:381t*
defined, *II:383*
empirical illustration of, *II:388–393*
technique of, *II:384–385*
testing for, *II:386–387*
test of, *II:394t, II:396t*
use of, *II:397*
- Collateralized debt obligations (CDOs), *I:299, I:525, III:553, III:645*
- Collateralized mortgage obligations (CMOs), *III:598, III:645*
- Collinearity, *II:329–330*
- Commodities, *I:279, I:556, I:566*
- Companies. *See* firms
- Comparison principals, *II:676*
- Comparisons *vs.* testing, *I:156*
- Complete markets, *I:103–104, I:119, I:133, I:461*
- Complexity, profiting from, *II:57–58*
- Complexity (Waldrop), *II:699*
- Complex numbers, *II:591–592, II:592f*
- Compounding. *See also* interest and annual percentage rates, *II:616*
continuous, *II:599, II:617*
determining number of periods, *II:602*
discrete *vs.* continuous, *III:570–571*
formula for growth rate, *II:8*
more than once per year, *II:598–599*
and present value, *II:618*
- Comprehensive Capital Analysis and Review, *I:300*
- Comprehensive Capital Assessment Review, *I:412*
- Computational burden, *III:643–644*
- Computers. *See also* various software applications
increased use of, *III:137–138*
introduction of into finance, *II:480*
modeling with, *I:511, II:695*
random walk generation of, *II:708*
in stochastic programming, *III:124, III:125–126*
- Concordance, defined, *I:327*
- Conditional autoregressive value at risk (CAViaR), *II:366*
- Conditional default rate (CDR). *See* CDRs (conditional default rates)
- Conditionally independent defaults (CID) models, *I:320, I:321–322, I:323*
- Conditioning/conditions, *I:24, II:307–308, II:361, II:645*
- Confidence, *I:200, I:201, II:723, III:319*
- Confidence intervals, *II:440, III:338t, III:399–400, III:400f*
- Conglomerate discounts, *II:43*
- Conseco, debt restructure of, *I:529*
- Consistency, notion of, *II:666–667*
- Constant elasticity of variance (CEV), *III:550, III:551f, III:654–655*
- Constant growth dividend discount model, *II:7–9*
- Constraints, portfolio
cardinality, *II:64–65*
common, *III:146*
commonly used, *II:62–66, II:84*
holding, *II:62–63*
minimum holding/transaction size, *II:65*
nonnegativity, *I:73*
real world, *II:224–225*
round lot, *II:65–66*
setting, *I:192*
turnover, *II:63*
on weights of, *I:191–192*
- Constraint sets, *I:21, I:28, I:29*
- Consumer Price Index (CPI), *I:277–278, I:291f, I:292, I:292f*
- Consumption, *I:59–60, II:360, III:570*
- Contagion, *I:320, I:324, I:333*
- Contingent claims
financial instruments as, *I:462*
incomplete markets for, *I:461–462*
unit, *I:458*
use of analysis, *I:463*
utility maximization in markets, *I:459–461*
value of, *I:458–459*
- Continuity, formal treatment of, *II:583–584*
- Continuous distribution function (c.d.f.), *III:167, III:196, III:205, III:345–346, III:345f*
- Continuous distribution function F(a), *III:196*
- Continuous time/continuous state, *III:578*
- Continuous-time processes, change of measure for, *III:511–512*
- Control flow statements in VBA, *III:458–460*
- Control methods, stochastic, *I:560*
- Convenience yields, *I:424, I:439*
- Convergence analysis, *II:667–668*
- Conversion, *I:274, I:445*
- Convexity
in callable bonds, *III:302–303*
defined, *I:258–259, III:309*
effective, *III:13, III:300–304, III:617t*
measurement of, *III:13–14, III:304–305*
negative, *III:14, III:49, III:303*
positive, *III:13*
use of, *III:299–300*
- Convex programming, *I:29, I:31–32*
- Cootner, Paul, *III:242*
- Copulas
advantages of, *III:284*
defined, *III:283*
mathematics of, *III:284–286*
usefulness of, *III:287*
visualization of bivariate independence, *III:285f*
visualization of Gaussian, *III:287f*
- Corner solutions, *I:200*
- Correlation coefficients
relation to R^2 , *II:316*
and Theil-Sen regression, *II:444*
use of, *III:286–287*
- Correlation matrices, *II:160t, II:163t, III:396–397*
- Correlations
in binomial distribution, *I:118*
computation of, *I:92–93*

- concept of, *III:283*
- drawbacks of, *III:283–284*
- between periodic increments, *III:540t*
- and portfolio risk, *I:11*
- robust estimates of, *II:443–446*
- serial, *II:220*
- undesirable, *I:293*
- use of, *II:271*
- Costs, net financing, *I:481*
- Cotton prices, model of, *III:383*
- Countable additivity, *III:158*
- Counterparts, robust, *II:81*
- Countries, low- vs. high inflation, *I:290*
- Coupon payments, *I:212, III:4*
- Coupon rates, computing of, *III:548–549*
- Courant-Friedrichs-Lewy (CFL) conditions, *II:657*
- Covariance
 - calculation of between assets, *I:8–9*
 - estimators for, *I:38–40, I:194–195*
 - matrix, *I:38–39, I:155, I:190*
 - relationship with correlation, *I:9*
 - reliability of sample estimates, *II:77*
 - use of, *II:370–371*
- Covariance matrices
 - decisions for interest rates, *III:406*
 - eigenvectors/eigenvalues, *II:160t*
 - equally weighted moving average, *III:402–403*
 - frequency of observations for, *III:404*
 - graphic of, *II:161t*
 - residuals of return process of, *II:162t*
 - of RiskMetrics™ Group, *III:412–413*
 - statistical methodology for, *III:398–399*
 - of ten stock returns, *II:159t*
 - use of, *II:158–159, II:169*
 - using EWMA in, *III:411*
- Coverage ratios, *II:560–561*
- Cox-Ingersoll-Ross (CIR) model, *I:260, I:491–492, I:547, I:548, III:546–547, III:656*
- Cox processes, *I:315–316, II:470–471*
- Cox-Ross-Rubenstein model, *I:510, I:522, II:678*
- CPI (Consumer Price Index), *I:277–278, I:291f, I:292, I:292f*
- CPRs (conditional prepayment rates). *See* prepayment, conditional
- CPR vector, *III:74. See also* prepayment, conditional
- Cramer, Harald, *II:470–471*
- Crank-Nicolson schemes, *II:666, II:669, II:674, II:680*
- Crank Nicolson-splitting (CN-S) schemes, *II:675*
- Crashmetrics, use of, *III:379, III:380*
- Credible intervals, *I:156*
- Credit-adjusted spread trees, *I:274*
- Credit crises
 - of 2007, *III:74*
 - of 2008, *III:381*
 - data from and DTS model, *I:396*
 - in Japan, *I:417*
- Credit curing, *III:73*
- Credit default swaps (CDSs). *See* CDSs (credit default swaps)
- Credit events
 - and credit loss, *I:379*
 - in default swaps, *I:526, I:528–530*
 - definitions of, *I:528*
 - descriptions of most used, *I:528t*
 - exchanges/payments in, *I:231f*
 - in MBS turnover, *III:66*
 - prepayments from, *III:49–50*
 - protection against, *I:230*
 - and simultaneous defaults, *I:323*
- Credit hedging, *I:405*
- Credit inputs, interaction of, *III:36–38*
- Credit loss
 - computation of, *I:382–383*
 - distribution of, *I:369f*
 - example of distribution of, *I:386f*
 - simulated, *I:389*
 - steps for simulation of, *I:379–380*
- Credit models, *I:300, I:302, I:303*
- Credit performance, evolution of, *III:32–36*
- Credit ratings
 - categories of, *I:362*
 - consumer, *I:302*
 - disadvantages of, *I:300–301*
 - implied, *I:381–382*
 - maturity of, *I:301*
 - reasons for, *I:300*
 - risks for, *II:280–281, II:280t*
 - use of, *I:309*
- Credit risk
 - common, *I:322*
 - counterparty, *I:413*
 - in credit default swaps, *I:535*
 - defined, *I:361*
 - distribution of, *I:377*
 - importance of, *III:81*
 - measures for, *I:386f*
 - modeling, *I:299–300, I:322, III:183*
 - quantification of, *I:369–372*
 - reports on, *II:278–281*
 - shipping, *I:566*
 - and spread duration, *I:391–392*
 - vs. cash flow risk, *III:377–378*
- Credit scores, *I:300–302, I:301–302, I:309, I:310n*
- Credit spreads
 - alternative models of, *I:405–406*
 - analysis with stock prices, *I:305t*
 - applications of, *I:404–405*
 - decomposition, *I:401–402*
 - drivers of, *I:402*
 - interpretation of, *I:403–404*
 - model specification, *I:403*
 - relationship with stock prices, *I:304*
 - risk in, *II:279t*
 - use of, *I:222–223*
- Credit support, evaluation of, *III:39–40*
- Credit value at risk (CVaR). *See* CVaR
- Crisis situations, estimating liquidity in, *III:378–380*
- Critical line algorithm (CLA), *I:73*
- Cross-trading, *II:85n*
- Cross-validation, leave-one-out, *II:413–414*
- Crude oil, *I:561, I:562*
- Cumulation, defined, *III:471*
- Cumulative default rate (CDX), *III:58*
- Cumulative frequency distributions, *II:493f, II:493t, II:498–499*
- formal presentation of, *II:492–493*
- Currency put options, *I:515*
- Current ratio, *II:554*
- Curve imbalances, *II:270–271*
- Curve options, *III:553*
- Curve risk, *II:275–278*
- CUSIPs/ticker symbols, changes in, *II:202–203*
- CVaR (credit value at risk), *I:384–385, I:385–386, II:68, II:85n, III:392t. See also* value at risk (VaR)
- Daily increments of volatility, *III:534*
- Daily log returns, *II:407–408*
- Dark pools, *II:450, II:454*
- Data. *See also* operational loss data
 - absolute, *II:487–488*
 - acquisition and processing of, *II:198*
 - alignment of, *II:202–203*
 - amount of, *I:196*
 - augmentation of, *I:186n*
 - availability of, *II:202, II:486*
 - backfilling of, *II:202*
 - bias of, *II:204, II:713*
 - bid-ask aggregation techniques for, *II:457f*
 - classification of, *II:499–500*
 - collection of, *II:102, II:103f*
 - cross-sectional, *II:201, II:488, II:488f*
 - in forecasting models, *II:230*
 - frequency of, *II:113, II:368, II:462–463, II:500*
 - fundamental, *II:246–247*
 - generation of, *II:295–296*

- Data (*Continued*)
- high-frequency (HFD) (*See* high-frequency data (HFD))
 - historical, *II:77–78, II:122, II:172*
 - housing bubble, *II:397–399*
 - importing into MATLAB, *III:433–434*
 - industry-specific, *II:105*
 - integrity of, *II:201–203*
 - levels and scale of, *II:486–487*
 - long-term, *III:389–390*
 - in mean-variance, *I:193–194*
 - misuse of, *II:108*
 - on operational loss, *III:99*
 - from OTC business, *II:486*
 - patterns in, *II:707–708*
 - pooling of, *III:96*
 - of precision, *I:158*
 - preliminary analysis of, *III:362*
 - problems in for operational risk, *III:97–98*
 - qualitative *vs.* quantitative, *II:486*
 - quality of, *II:204, II:211, II:452–453, II:486, II:695*
 - reasons for classification of, *II:493–494*
 - for relative valuation, *II:34–35*
 - restatements of, *II:202*
 - sampling of, *II:459f, II:711*
 - scarcity of, *II:699–700, II:703–704, II:718*
 - sorting and counting of, *II:488–491*
 - standardization of, *II:204, III:228*
 - structure/sample size of, *II:703*
 - types of, *II:486–488*
 - underlying signals, *II:111*
 - univariate, defined, *II:485*
 - working with, *II:201–206*
- Databases
- Compustat Point-In-Time, *II:238*
 - Factiva, *II:482*
 - Institutional Brokers Estimate System (IBES), *II:238*
 - structured, *II:482*
 - third-party, *II:198, II:211n*
- Data classes, criteria for, *II:500*
- Data generating processes (DGPs), *II:295–296, II:298f, II:502, II:702, III:278*
- Data periods, length of, *III:404*
- Data series, effect of large number of, *II:708–709*
- Data sets, training/test, *II:710–711*
- Data snooping, *II:700, II:710–712, II:714, II:717, II:718*
- Datini, Francesco, *II:479–480*
- Davis-Lo infectious defaults model, *I:324*
- Days payables outstanding (DPO), calculation of, *II:553–554*
- Days sales outstanding (DSO), calculation of, *II:553*
- DCF (discounted cash flow) models, *II:16, II:44–45*
- DDM (dividend discount models). *See* dividend discount models (DDM)
- Debt
- long-term, in financial statements, *II:542*
 - models of risky, *I:304–307*
 - restructuring of, *I:230*
 - risky, *I:307–308*
- Debt-to-assets ratio, *II:559*
- Debt-to-equity ratio, *II:559*
- Decomposition models
- active/passive, *III:19*
- Default correlation, *I:317–318*
- contagion, *I:353–354*
 - cyclical, *I:352, I:353*
 - linear, *I:320–321*
 - measures of, *I:320–321*
 - tools for modeling, *I:319–333*
- Default intensity, *III:225*
- Default models, *I:321–322, I:370f*
- Default probabilities
- adjustments in real time, *I:300–301*
 - between companies, *I:412–413*
 - cyclical rise and fall, *I:408f, I:409f*
 - defined, *I:299–300*
 - effect of business cycle on, *I:408*
 - effect of rating outlooks on, *I:365–366*
 - empirical approach to, *I:362–363*
 - five-year (Bank of America and Citigroup), *I:301f, I:302f*
 - merits of approaches to, *I:365*
 - Merton's approach to, *I:363–365*
 - probability of, *II:727, II:727f, II:728f*
 - and survival, *I:533–535*
 - and survival probability, *I:323–324*
 - term structure of, *I:303*
 - time span of, *I:302–303*
 - vs.* ratings and credit scores, *I:300–302*
 - for Washington Mutual, *I:415f, I:416f*
 - of Washington Mutual, *I:415f, I:416f*
- Defaults
- annual rates of, *I:363*
 - and Bernoulli distributions, *III:169–170*
 - calculation of monthly, *III:61t*
 - clustering of, *I:324–325*
 - contagion, *I:320*
 - copulas for times, *I:329–331*
 - correlation of between companies, *I:411*
 - cost of, *I:401, I:404f*
 - dollar amounts of, *III:59f*
 - effect of, *I:228, III:645*
 - event *vs.* liquidation, *I:349*
 - factors influencing, *III:74–75*
 - first passage model of, *I:349*
 - historical database of, *I:414*
 - intensity of, *I:330, I:414*
 - looping, *I:324–325*
 - measures of, *III:58–59*
 - in Merton approach, *I:306*
 - Moody's definition of, *I:363*
 - predictability of, *I:346–347*
 - and prepayments, *III:49–50, III:76–77*
 - process, relationship to recovery rate, *I:372*
 - pseudo intensities, *I:330*
 - rates of cumulative/conditional, *III:63*
 - recovery after, *I:316–317*
 - risk of, *I:210*
 - simulation of times, *I:322–324, I:325*
 - threshold of, *I:345–346*
 - times simulation of, *I:319*
 - triggers for, *I:347–348*
 - variables in, *I:307–308*
- Default swaps
- assumptions about, *I:531–532*
 - and credit events, *I:530*
 - digital, *I:537*
 - discussion of, *I:526–528*
 - market relationship with cash market, *I:530*
 - and restructuring, *I:528–529*
 - value of spread, *I:534*
- Default times, *I:332*
- Definite covariance matrix, *II:445*
- Deflators, *I:129, I:136*
- Degrees, in ordinary differential equations, *II:644–645*
- Degrees of freedom (DOF)
- across assets and time, *II:735–736*
 - in chi-square distribution, *III:212*
 - defined, *II:734*
 - for Dow Jones Industrial Average (DJIA), *II:735–737, II:737f*
 - prior distribution for, *I:177*
 - range of, *I:187n*
 - for S&P 500 index stock returns, *II:735–736, II:736f*
- Delinquency measures, *III:57–58*
- Delivery date, *I:478*
- Delta, *I:509, I:516–518, I:521*
- Delta-gamma approximation, *I:519, III:644–645*
- Delta hedging, *I:413, I:416, I:418, I:517*

- Delta profile, *I:518f*
- Densities
- beta, *III:108f*
 - Burr, *III:110f*
 - closed-form solutions for, *III:243*
 - exponential, *III:105–106, III:105f*
 - gamma, *III:108f*
 - Pareto, *III:109f*
 - posterior, *I:170f*
 - two-point lognormal, *III:111f*
- Density curves, *I:147f*
- Density functions
- asymmetric, *III:205f*
 - of beta distribution, *III:222f*
 - chi-square distributions, *III:213f*
 - common means, different variances, *III:203f*
 - computing probabilities from, *III:201*
 - discussion of, *III:197–200*
 - of *F*-distribution, *III:217f*
 - histogram of, *III:198f*
 - of log-normal distribution, *III:223f*
 - and normal distribution, *II:733*
 - and probability, *III:206*
 - rectangular distributions, *III:220*
 - requirements of, *III:198–200*
 - symmetric, *III:204f*
 - of *t*-distribution, *III:214f*
- Dependence, *I:326–327, II:305–308*
- Depreciation, *II:22*
- accumulated, *II:533–534*
 - expense *vs.* book value, *II:539f*
 - expense *vs.* carrying value, *II:540f*
 - in financial statements, *II:537–539*
 - on income statements, *II:536*
 - methods of allocation, *II:537–538*
- Derivatives
- construction of, *II:586–587*
 - described, *II:585–586*
 - embedded, *I:462*
 - energy, *I:558*
 - exotic, *I:558, I:559–560*
 - of functions, defined, *II:593*
 - and incomplete markets, *I:462*
 - interest rate, *III:589–590*
 - nonlinearity of, *III:644–645*
 - OTC, *I:538*
 - pricing of, *I:58, III:594–596*
 - pricing of financial, *III:642–643*
 - relationship with integrals, *II:590*
 - for shipping assets, *I:555, I:558, I:565–566*
 - use of instruments, *I:477*
 - valuation and hedging of, *I:558–560*
 - vanilla, *I:559*
- Derman, Emanuel, *II:694*
- Descriptors, *II:140, II:246–247, II:256*
- Determinants, *II:623*
- Deterministic methods
- usefulness of, *II:685*
- Diagonal VEC model (DVEC), *II:372*
- Dice, and probability, *III:152, III:153, III:155–156, III:156t*
- Dickey-Fuller statistic, *II:386–387*
- Dickey-Fuller tests, *II:514*
- Difference, notation of, *I:80*
- Differential equations
- classification of, *II:657–658*
 - defined, *I:95, II:644, II:657*
 - first-order system of, *II:646*
 - general solutions to, *II:645*
 - linear, *II:647–648*
 - linear ordinary, *II:644–645*
 - partial (PDE), *II:643, II:654–657*
 - stochastic, *II:643–644*
 - systems of ordinary, *II:645–646*
 - usefulness of, *II:658*
- Diffusion, *III:539, III:554–555*
- Diffusion invariance principle, *I:132*
- Dimensionality, curse of, *II:673, III:127*
- Dirac measures, *III:271*
- Directional measures, *II:428, II:429*
- Dirichlet boundary conditions, *II:666*
- Dirichlet distribution, *I:181–183, I:186–187n*
- Discounted cash flow (DCF) models, *II:16, II:44–45*
- Discount factors, *I:57–58, I:59–62, I:60, II:600–601*
- Discount function
- calculation of, *III:571*
 - defined, *III:563*
 - discussion of, *III:563–565*
 - forward rates from, *III:566–567*
 - graph of, *III:563f*
 - for on-the-run Treasuries, *III:564–565*
- Discounting, defined, *II:596*
- Discount rates, *I:211, I:212, I:215–216, II:6*
- Discovery heuristics, *II:711*
- Discrepancies, importance of small, *II:696*
- Discrete law, *III:165–169*
- Discrete maximum principle, *II:668*
- Discretization, *I:265, II:669f, II:672*
- Disentangling, *II:51–56*
- complexities of, *II:55–56*
 - predictive power of, *II:54–55*
 - return revelation of, *II:52–54*
 - usefulness of, *II:52, II:58*
- Dispersion measures, *III:352, III:353–354, III:357*
- Dispersion parameters, *III:202–205*
- Distress events, *I:351*
- Distributional measures, *II:428*
- Distribution analysis, cash flow, *III:310*
- Distribution function, *III:218f, III:224f*
- Distributions
- application of hypergeometric, *III:177–178*
 - beliefs about, *I:152–153*
 - Bernoulli, *III:169–170, III:185t*
 - beta, *I:148, III:108*
 - binomial, *I:81f, III:170–174, III:185t, III:363*
 - Burr, *III:109–110*
 - categories for extreme values, *II:752*
 - common loss, *III:112t*
 - commonly used, *III:225*
 - conditional, *III:219*
 - conditional posterior, *I:178–179, I:182–183, I:184–185*
 - conjugate prior, *I:154*
 - continuous probability, *III:195–196*
 - discrete, *III:185t*
 - discrete cumulative, *III:166*
 - discrete uniform, *III:183–184, III:185t, III:638f*
 - empirical, *II:498, III:104–105, III:105f*
 - exponential, *III:105–106*
 - finite-dimensional, *II:502*
 - of Fréchet, Gumbel and Weibull, *III:267f*
 - gamma, *III:107–108, III:221–222*
 - Gaussian, *III:210–212*
 - Gumbel, *III:228, III:230*
 - heavy-tailed, *I:186n, II:733, III:109, III:260*
 - hypergeometric, *III:174–178, III:185t*
 - indicating location of, *III:235*
 - infinitely divisible, *III:253–256, III:253t*
 - informative prior, *I:152–153*
 - inverted Wishart, *I:172*
 - light- *vs.* heavy-tailed, *III:111–112*
 - lognormal, *III:106, III:106f, III:538–539*
 - mixture loss, *III:110–111*
 - for modeling applications, *III:257*
 - multinomial, *III:179–182, III:185t*
 - non-Gaussian, *III:254*
 - noninformative prior, *I:153–154*
 - normal (*See* normal distributions)
 - parametric, *III:201*
 - Poisson, *I:142, III:182–183, III:185t, III:217–218*
 - Poisson probability, *III:187t*
 - posterior, *I:147–148, I:165, I:166–167, I:169–170, I:177, I:183–184*
 - power-law, *III:262–263*
 - predictive, *I:167*
 - prior, *I:177, I:181–182, I:196*
 - proposal, *I:183–184*
 - representation of stable and CTS, *II:742–743*

- Distributions (*Continued*)
 spherical, II:310
 stable, III:238, III:242, III:264–265, III:384 (*See also* α -stable distributions)
 subexponential, III:261–262
 tails of, III:112*f*, III:648
 tempered stable, III:257, III:382
 testing applied to truncated, III:367
- Diversification, II:57–58
 achieving, I:10
 and cap weighting, I:38
 and credit default swaps, I:413–414
 example of, I:15
 international, II:393–396
 Markowitz's work on, II:471
- Diversification effect, III:321
- Diversification indicators, I:192
- Dividend discount models (DDM)
 applied to electric utilities, II:12*t*
 applied to stocks, II:16–17
 basic, II:5
 constant growth, II:7–9, II:17–18
 defined, II:14
 finite life general, II:5–7
 free cash flow model, II:21–23
 intuition behind, II:18–19
 multiphase, II:9–10
 non-constant growth, II:18
 predictive power of, II:54
 in the real world, II:19–20
 stochastic, II:10–12, II:12*t*
- Dividend payout ratio, II:4, II:20
- Dividends
 expected growth in, II:19
 forecasting of, II:6
 measurement of, II:3–4, II:14
 per share, II:3–4
 reasons for not paying, II:27
 required rate of return, II:19
 and stock prices, II:4–5
- Dividend yield, II:4, II:19
- Documentation
 of model risk, II:696, II:697
- Dothan model, I:491, I:493
- Dow Jones Global Titans 500 (DJGTI), II:490*t*, II:491*t*
- Dow Jones Industrial Average (DJIA)
 in comparison of risk models, II:747–751
 components of, II:489*t*
 fitted stable tail index for, II:740*f*
 frequency distribution in, II:489*t*
 performance (January 2004 to June 2011), II:749*f*
 relative frequencies, II:491*t*
 stocks by share price, II:492*t*
- Drawing without replacement, III:174–177
- Drawing with replacement, III:170, III:174, III:179–180
- Drift
 effects of, III:537
 of interest rates, I:263
 in randomness calculations, III:535
 in random walks, I:84, I:86
 time increments of, I:83
 of time series, I:80
 as variable, III:536
- DTS (duration times spread), I:392, I:393–394, I:396–398
- Duffie-Singleton model, I:542–543
- Dupire's formula, II:682–683, II:685
- DuPont system, II:548–551, II:551*f*
- Duration
 calculations of real yield and inflation, I:286
 computing of, I:285
 defined, I:284, III:309
 effective, III:300–304, III:617*t*
 effective/option adjusted, III:13
 empirical, of common stock, II:318–322, II:319–322*t*
 estimation of, II:323*t*
 measurement of, III:12–13, III:304–305
 models of, II:461
 modified *vs.* effective, III:299
- Duration/convexity, effective, I:255, I:256*f*
- Duration times spread (DTS). *See* DTS (duration times spread)
- Durbin-Watson test, III:647
- Dynamical systems
 equilibrium solution of, II:653
 study of, II:651
- Dynamic conditional correlation (DCC) model, II:373
- Dynamic term structures, III:576–577, III:578–579, III:591
- Early exercise, I:447, I:455. *See* calls, American-style; options
- Earnings before interest, taxes, depreciation and amortization (EBITDA), II:566
- Earnings before interest and taxes (EBIT), II:23, II:547, II:556
- Earnings growth factor, II:223
- Earnings per share (EPS), II:20–21, II:38–39, II:537
- Earnings revisions factor, II:207, II:209*f*
- EBITDA/EV factor
 correlations with, II:226
 examples of, II:203, II:203*f*, II:207, II:208*f*
 in models, II:232, II:238–239
 use of, II:222–223
- Econometrics
 financial, II:295, II:298–300, II:301–303
 modeling of, II:373, II:654
- Economic cycles, I:537, II:42–43
- Economic intuition, II:715–716
- Economic laws, changes in, II:700
- Economy
 states of, I:49–50, II:518–519, III:476
 term structures in certain, III:567–568
 time periods of, II:515–516
- Economy as an Evolving Complex System, The* (Anderson, Arrow, & Pines), II:699
- Educated guesses, use of, I:511
- EE (explicit Euler) scheme, II:674, II:677–678
- Effective annual rate (EAR), interest, II:616–617
- Efficiency
 in estimation, III:641–642
- Efficient frontier, I:13–14, I:17*f*, I:289*f*
- Efficient market theory, II:396, III:92
- Eggs, rotten, I:457–458
- Eigenvalues, II:627–628, II:705, II:706–707*f*, II:707*t*
- Einstein, Albert, II:470
- Elements, defined, III:153–154
- Embedding problem, and change of time method, III:520
- Emerging markets, transaction costs in, III:628
- EM (expectation maximization) algorithm, II:146, II:165
- Empirical rule, III:210, III:225
- Endogenous parameterization, III:580–581
- Energy
 cargoes of, I:561–562
 commodity price models, I:556–558
 forward curves of, I:564–565
 power plants and refineries, I:563
 storage of, I:560–561, I:563–564
- Engle-Granger cointegration test, II:386–388, II:391–392, II:395
- Entropy, III:354
- EPS (earnings per share), II:20–21, II:38–39, II:537
- Equally weighted moving average, III:400–402, III:406–407, III:408–409
- Equal to earnings before interest and taxes (EBIT), II:23, II:547, II:556
- Equal-variance assumption, I:164, I:167
- Equations
 difference, homogenous *vs.* nonhomogenous, II:638

- difference *vs.* differential, *II:629*
 diffusion, *II:654–656, II:658n*
 error-correction, *II:391, II:395f*
 homogeneous linear difference,
 II:639–642, II:641f
 homogenous difference, *II:630–634,*
 II:631–632f, II:633–634f, II:642
 linear, *II:623–624*
 linear difference, systems of,
 II:637–639
 matrix characteristics of, *II:628*
 no arbitrage, *III:612, III:617–619*
 nonhomogeneous difference,
 II:634–637, II:635f, II:637–638f
 stochastic, *III:478*
- Equilibrium**
 and absolute valuation models,
 I:260
 defined, *II:385–386*
 dimensions of, *III:601*
 in dynamic term structure models,
 III:576
 expectations for, *II:112*
 expected returns from, *II:112*
 modeling of, *III:577, III:594*
 in supply and demand, *III:568*
- Equilibrium models**
 use of, *III:603–604*
- Equilibrium term structure models,**
 III:601
- Equities, I:279**
 investing in, *II:89–90*
- Equity**
 on the balance sheet, *II:535*
 changes in homeowner, *III:73*
 in homes, *III:69*
 as option on assets, *I:304–305*
 shareholders', *II:535*
- Equity markets, II:48**
- Equity multipliers, II:550**
- Equity risk factor models, II:173–178**
- Equivalent probability measures,**
 I:111, III:510–511
- Ergodicity, defined, II:405**
- Erlang distribution, III:221–222**
- Errors. See also estimation error;**
 standard errors
 absolute percentages of, *II:525f,*
 II:526f
 estimates of, *II:676*
 in financial models, *II:719*
 a posteriori estimates, *II:672–673*
 sources of, *II:720*
 terms for, *II:126*
 in variables problem, *II:220*
- Esscher transform, III:511, III:514**
- Estimates/estimation**
 confidence in, *I:199*
 consensus, *II:34–35*
- equations for, *I:348–349*
 in EVT, *III:272–274*
 factor models in, *II:154*
 with GARCH models, *II:364–365*
 in-house from firms, *II:35*
 maximum likelihood, *II:311–313*
 methodology for, *II:174–176*
 and PCA, *II:167f*
 posterior, *I:176*
 posterior point, *I:155–156*
 processes for, *I:193, II:176*
 properties of for EWMA, *III:410–411*
 robust, *I:189*
 techniques of, *II:330*
 use of, *II:304*
- Estimation errors**
 accumulation of, *II:78*
 in the Black-Litterman model, *I:201*
 covariance matrix of, *III:139–140*
 effect of, *I:18*
 pessimism in, *III:143*
 in portfolio optimization, *II:82,*
 III:138–139
 sensitivity to, *I:191*
 and uncertainty sets, *III:141*
- Estimation risk, I:193**
 minimizing, *III:145*
- Estimators**
 bias in, *III:641*
 efficiency in, *III:641–642*
 equally weighted average,
 III:400–402
 factor-based, *I:39*
 terms used to describe, *II:314*
 unbiased, *III:399*
 variance, *II:313*
- ETL (expected tail loss), III:355–356**
- Euler approximation, II:649–650,**
 II:649f, II:650f
- Euler constant, III:182**
- Euler schemes, explicit/implicit, II:666**
- Europe**
 common currency for, *II:393*
 risk factors of, *II:174*
- European call options**
 Black-Scholes formula for,
 III:639–640
 computed by different methods,
 III:650–651, III:651f
 explicit option pricing formula,
 III:526–527
 pricing by simulation in VBA,
 III:465–466
 pricing in Black-Scholes setting,
 III:649
 simulation of pricing, *III:444–445,*
 III:462–463
 and term structure models,
 III:544–545
- European Central Bank, I:300**
- Events**
 defined, *III:85, III:162, III:508*
 effects of macroeconomic, *II:243–244*
 extreme, *III:245–246, III:260–261,*
 III:407
 identification of, *II:516*
 mutually exclusive, *III:158*
 in probability, *III:156*
 rare, *III:645*
 rare *vs.* normal, *I:262*
 tail, *III:88n, III:111, III:118*
 three- δ , *III:381–382*
- EVT (extreme value theory). See**
 extreme value theory (EVT)
- EWMA (exponentially weighted**
 moving averages), III:409–413
- Exceedance observations, III:362–363**
- Exceedances, of VaR, III:325–326,**
 III:339
- Excel**
 accessing VBA in, *III:477*
 add-ins for, *I:93, III:651*
 data series correlation in, *I:92–93*
 determining corresponding
 probabilities in, *III:646*
 Excel Link, *III:434*
 Excel Solver, *II:70*
 interactions with MATLAB, *III:448*
 macros in, *III:449, III:454–455*
 notations in, *III:477n*
 random number generation in,
 III:645–646
 random walks with, *I:83, I:85, I:87,*
 I:90
 @RISK in, *II:12f*
 syntax for functions in, *III:456*
- Exchange-rate intervention, study on,**
 III:177–178
- Exercise prices, I:452, I:484, I:508**
- Expectation maximization (EM)**
 algorithm, *II:146, II:165*
- Expectations, conditional, I:122,**
 II:517–518, III:508–509
- Expectations hypothesis, III:568–569,**
 III:601n
- Expected shortfall (ES), I:385–386,**
 III:332. See also average value at
 risk (AVaR)
- Expected tail loss (ETL), III:291,**
 III:293f, III:345–347, III:347f,
 III:355–356
- Expected value (EV), I:511**
- Expenses, noncash, II:25**
- Experiments, possibility of, II:307**
- Explicit costs, defined, III:623**
- Explicit Euler (EE) scheme, II:674,**
 II:677–678
- Exponential density function, III:218f**

- Exponential distribution, *III*:217–219
 applications in finance, *III*:219
- Exponentially weighted moving averages (EWMA)
 discussion of, *III*:409–413
 forecasting model of, *III*:411
 properties of the estimates, *III*:410–411
 standard errors for, *III*:411–412
 statistical methodology in, *III*:409
 usefulness of, *III*:413–414
 volatility estimates for, *III*:410*f*
- Exposures
 calculation of, *II*:247*t*
 correlation between, *II*:186
 distribution of, *II*:250*f*, *II*:251*f*, *II*:254
 management of, *II*:182–183
 monitoring of portfolio, *II*:249–250
 name-specific, *II*:188
- Extrema, characterization of local, *I*:23
- Extremal random variables, *III*:267
- Extreme value distributions, generalized, *III*:269
- Extreme value theory (EVT), *II*:744–746, *III*:95, *III*:228
 defined, *III*:238
 for IID processes, *III*:265–274
 in IID sequences, *III*:275
 role of in modeling, *II*:753*n*
- Factor analysis
 application of, *II*:165
 based on information coefficients, *II*:222
 defined, *II*:141, *II*:169
 discussion of, *II*:164–166
 importance of, *II*:238
vs. principal component analysis, *II*:166–168
- Factor-based strategies
vs. risk models, *II*:236
- Factor-based trading, *II*:196–197
 model construction for, *II*:228–235
 performance evaluation of, *II*:225–228
- Factor exposures, *II*:247–248, *II*:275–283
- Factorials, computing of, *III*:456
- Factorization, defined, *II*:307
- Factor mimicking portfolio (FMP), *II*:214
- Factor model estimation, *II*:142–147, *II*:150
 alternative approaches and extensions, *II*:145–147
 applied to bond returns, *II*:144–145
 computational procedure for, *II*:142–144
 fixed N, *II*:143
 large N, *II*:143–144
- Factor models
 in the Black-Litterman framework, *I*:200
 commonly used, *II*:150
 considerations in, *II*:178
 cross-sectional, *II*:220–221
 defined, *II*:153
 fixed income, *II*:271–272
 in forecasting, *II*:230–231
 linear, *II*:154–156, *II*:168
 normal, *II*:156
 predictive, *II*:142
 static/dynamic, *II*:146–147, *II*:155
 in statistical methodology, *II*:141
 strict, *II*:155–156
 types of, *II*:138–142
 usefulness of, *II*:154, *II*:503
 use of, *I*:354, *II*:137, *II*:150, *II*:168, *II*:219–225
- Factor portfolios, *II*:224–225
- Factor premiums, cross-sectional methods for evaluation of, *II*:214–219
- Factor returns, *II*:191*t*, *II*:192*t*
 calculation of, *II*:248
- Factor risk models, *II*:113, *II*:119
- Factors
 adjustment of, *II*:205–206
 analysis of data of, *II*:206–211
 categories of, *II*:197
 choice of, *II*:232–235
 defined, *II*:196, *II*:211
 desirable properties of, *II*:200
 development of, *II*:198
 estimation of types of, *II*:156
 graph of, *II*:166*f*
 known, *II*:138–139
 K systematic, *II*:138–139
 latent, *II*:140–141, *II*:150
 loadings of, *II*:144, *II*:145*t*, *II*:155, *II*:166*t*, *II*:167*f*, *II*:168*t*
 market, *II*:176
 orthogonalization of, *II*:205–206
 relationship to time series, *II*:168*f*
 sorting of, *II*:215
 sources for, *II*:200–201
 statistical, *II*:197
 summary of well-known, *II*:196*t*
 transformations applied to, *II*:206
 use of multiple, *II*:141–142
- Failures, probability of, *II*:726–727
- Fair equilibrium, between multiple accounts, *II*:76
- Fair value
 determination of, *III*:584–585
- Fair value, assessment of, *II*:6–7
- Fama, Eugene, *II*:468, *II*:473–474
- Fama-French three-factor model, *II*:139–140, *II*:177
- Fama-MacBeth regression, *II*:220–221, *II*:224, *II*:227–228, *II*:228*f*, *II*:237, *II*:240*n*
- Fannie Mae/Freddie Mac, writedowns of, *III*:77*n*
- Fast Fourier transform algorithm, *II*:743
- Fat tails
 of asset return distributions, *III*:242
 in chaotic systems, *II*:653
 class \mathcal{L} , *III*:261–263
 comparison between risk models, *II*:749–750
 effects of, *II*:354
 importance of, *II*:524
 properties of, *III*:260–261
 in Student's *t* distribution, *II*:734
- Favorable selection, *III*:76–77
- F*-distribution, *III*:216–217
- Federal Reserve
 effects of on inflation risk premium, *I*:281
 study by Cleveland Bank, *III*:177–178
 timing of interventions of, *III*:178
- Feynman-Kac formulas, *II*:661
- FFAs (freight forward agreements), *I*:566
- Filtered probability spaces, *I*:314–315, *I*:334*n*
- Filtration, *II*:516–517, *III*:476–477, *III*:489–490, *III*:508
- Finance, three major revolutions in, *III*:350
- Finance companies, captive, *I*:366–369
- Finance theory
 development of, *II*:467–468
 effect of computers on, *II*:476
 in the nineteenth century, *II*:468–469, *II*:476
 in the 1960s, *II*:476
 in the 1970s, *II*:476
 stochastic laws in, *III*:472
 in the twentieth century, *II*:476
- Financial assets, price distribution of, *III*:349–350
- Financial crisis (2008), *III*:71
- Financial date, pro forma, *II*:542–543
- Financial distress, defined, *I*:351
- Financial institutions, model risk of, *II*:693
- Financial leverage ratios, *II*:559–561, *II*:563
- Financial modelers, mistakes of, *II*:707–710

- Financial planning, *III*:126–127, *III*:128, *III*:129
- Financial ratios, *II*:546, *II*:563–564
- Financial statements
 assumptions used in creating, *II*:532
 data in, *II*:563
 information in, *II*:533–542, *II*:543
 pro forma, *II*:22–23
 time statements for, *II*:532
 usefulness of, *II*:531
 use of, *II*:204–205, *II*:246
- Financial time series, *I*:79–80, *I*:386–387, *II*:415–416, *II*:503–504
- Financial variables, modeling of, *III*:280
- Find, in MATLAB, *III*:422
- Finite difference methods, *II*:648–652, *II*:656–657, *II*:665–666, *II*:674–675, *II*:676–677, *III*:19
- Finite element methods, *II*:669–670, *II*:672, *II*:679–681
- Finite element space, *II*:670–672
- Finite life general DDM, *II*:5–7
- Finite states, assumption of, *I*:100–101
- Firms
 assessment of, *II*:546–547
 and capital structure, *II*:473
 characteristics of, *II*:94, *II*:176–177, *II*:201
 clientele of, *II*:36
 comparable, *II*:34, *II*:35–36
 geographic location of, *II*:36
 history *vs.* future prospects, *II*:92
 phases of, *II*:9–10
 retained earnings of, *II*:20
 valuation of, *II*:26–27, *II*:473
 value of, *II*:27–31, *II*:39
vs. characteristics of group, *II*:90–91
- First boundary problem, *II*:655–656, *II*:657f
- First Interstate Bancorp, *I*:304
 analysis of credit spreads, *I*:305t
 debt ratings of, *I*:410
- First passage models (FPMs), *I*:342, *I*:344–348
- Fischer-Tippett theorem, *III*:266–267
- Fisher, Ronald, *I*:140
- Fisherian, defined, *I*:140
- Fisher's information matrix, *I*:160n
- Fisher's law, *II*:322–323
- Fixed-asset turnover ratio, *II*:558
- Fixed-charge coverage ratio, *II*:560–561
- Flesaker-Hughston (FH) model, *III*:548–549
- Flows, discrete, *I*:448–453
- FMP (factor mimicking portfolio), *II*:214
- Footnotes, in financial statements, *II*:541–542
- Ford Motor Company, *I*:408f, *I*:409f
- Forecastability, *II*:132
- Forecastability, concept of, *II*:123
- Forecast encompassing
 defined, *II*:230–231
- Forecasts
 of bid-ask spreads, *II*:456–457
 comparisons of, *II*:420–421
 contingency tables, *II*:429t
 development of, *II*:110–114
 directional, *II*:428
 effect on future of, *II*:122–123
 errors in, *II*:422f
 evaluation of, *II*:428–430, *III*:368–370
 machine-learning approach to, *II*:128
 measures of, *II*:429–430, *II*:430
 need for, *II*:110–111
 in neural networks, *II*:419–420
 one-step ahead, *II*:421f
 parametric bootstraps for, *II*:428–430
 response to macroeconomic shocks, *II*:55f
 usefulness of, *II*:131–132
 use of models for, *II*:302
 of volatility, *III*:412
- Foreclosures, *III*:31, *III*:75
- Forward contracts
 advantages of, *I*:430
 buying assets of, *I*:439
 defined, *I*:426, *I*:478
 equivalence to futures prices, *I*:432–433
 hedging with, *I*:429, *I*:429t
 as OTC instruments, *I*:479
 prepaid, *I*:428
 price paths of, *I*:428t
 short *vs.* long, *I*:437–438, *I*:438f
 valuing of, *I*:426–430
vs. futures, *I*:430–431, *I*:433
vs. options, *I*:437–439
- Forward curves
 graph of, *I*:434f
 modeling of, *I*:533, *I*:557–558, *I*:564–565
 normal *vs.* inverted, *I*:434
 of physical commodities, *I*:555
- Forward freight agreements (FFAs), *I*:555, *I*:558, *I*:566
- Forward measure, use of, *I*:543–544
- Forward rates
 calculation of, *I*:491, *III*:572
 defined, *I*:509–510
 from discount function, *III*:566–567
 implied, *III*:565–567
 models of, *III*:543–544
 from spot yields, *III*:566
 of term structure, *III*:586
- Fourier integrals, *II*:656
- Fourier methods, *I*:559–560
- Fourier transform, *III*:265
- FPMs (first passage models), *I*:342, *I*:344–348
- Fractals, *II*:653–654, *III*:278–280, *III*:479–480
- Franklin Tempelton Investment Funds, *II*:496t, *II*:497t, *II*:498t
- Fréchet distribution, *II*:754n, *III*:228, *III*:230, *III*:265, *III*:267, *III*:268
- Fréchet-Hoeffding copulas, *I*:327, *I*:329
- Freddie Mac, *II*:77n, *II*:754n, *III*:49
- Free cash flow (FCF), *II*:21–23
 analysis of, *II*:570–571
 calculation of, *II*:23–24, *II*:571–572
 defined, *II*:569–571, *II*:578
 expected for XYZ, Inc., *II*:30t
 financial adjustments to, *II*:25–26
 statement of, direct method, *II*:24–25, *II*:24t
 statement of, indirect method, *II*:24–25, *II*:24t
vs. cash flow, *II*:22–23
- Freedman-Diaconis rule, *II*:494, *II*:495, *II*:497
- Frequencies
 accumulating, *II*:491–492
 distributions of, *II*:488–491, *II*:499f
 empirical cumulative, *II*:492
 formal presentation of, *II*:491
- Frequentist, *I*:140, *I*:148
- Frictions, costs of, *II*:472–473
- Friedman, Milton, *I*:123
- Frontiers, true, estimated and actual efficient, *I*:190–191
- F_SCORE, use of, *II*:230–231
- F-test, *II*:336, *II*:337, *II*:344, *II*:425, *II*:426
- FTSE 100, volatility in, *III*:412–413
- Fuel costs, *I*:561, *I*:562–563. *See also* energy
- Full disclosure, defined, *II*:532
- Functional, defined, *I*:24
- Functional-coefficient autoregressive (FAR) model, *II*:417
- Functions
 affine, *I*:31
 Archimedean, *I*:329, *I*:330–331, *I*:331
 Bessel, of the third kind, *II*:591
 beta, *II*:591
 characteristic, *II*:591–592, *II*:593
 choosing and calibrating of, *I*:331–333
 Clayton, Frank, Gumbel, and Product, *I*:329

- Functions (*Continued*)
 continuous, *II:581–584, II:582f, II:583, II:592–593*
 continuous / discontinuous, *II:582f*
 convex, *I:24–27, I:25, I:25f, I:26f*
 convex quadratic, *I:26, I:31f*
 copula, *I:320, I:325–333, I:407–408*
 for default times, *I:329–331*
 defined, *I:24, I:333*
 density, *I:141*
 with derivatives, *II:585f*
 elementary, *III:474*
 elliptical, *I:328–329*
 empirical distribution, *III:270*
 factorial, *II:590–591*
 gamma, *II:591, II:591f, III:212*
 gradients of, *I:23*
 Heaviside, *II:418–419*
 hypergeometric, *III:256, III:257*
 indicator, *II:584–585, II:584f, II:593*
 likelihood function, *I:141–143, I:143f, I:144f, I:148, I:176, I:177*
 measurable, *III:159–160, III:160f, III:201*
 minimization and maximization of values, *I:22, I:22f*
 monotonically increasing, *II:587–588, II:588f*
 nonconvex quadratic, *I:26–27*
 nondecreasing, *III:154–155, III:155f*
 normal density, *III:226f*
 optimization of, *I:24*
 parameters of copulas, *I:331–332*
 properties of quasi-convex, *I:28*
 quasi-concave, *I:27–28, I:27f*
 right-continuous, *III:154–155, III:155f*
 surface of linear, *I:33f*
 with two local maxima, *I:23f*
 usefulness of, *I:411–412*
 utility, *I:4–5, I:14–15, I:461*
 Fund management, art of, *I:273*
 Fund separation theorems, *I:36*
 Futures
 Eurodollar, *I:503*
 hedging with, *I:433*
 market for housing, *II:396–397*
 prices of, and interest rates, *I:435n*
 telescoping positions of, *I:431–432*
 theoretical, *I:487*
 valuing of, *I:430–433*
 vs. forward contracts, *I:430–431*
 Futures contracts
 defined, *I:478*
 determining price of, *I:481*
 pricing model for, *I:479–481*
 theoretical price of, *I:481–484*
 vs. forward contracts, *I:433, I:478–479*
 Futures options, defined, *I:453*
 Future value, *II:618*
 determining of money, *II:596–600*
 Galerkin methods, principle of, *II:671*
 Gamma, *I:509, I:518–520*
 Gamma process, *III:498*
 Gamma profile, *I:519f*
 Gapping effect, *I:509*
 GARCH (generalized autoregressive conditional heteroskedastic) models
 asymmetric, *II:367–368*
 exponential (EGARCH), *II:367–368*
 extensions of, *III:657*
 factor models, *II:372*
 GARCH-M (GARCH in mean), *II:368*
 Markov-switching, *I:180–184*
 time aggregation in, *II:369–370*
 type of, *II:131*
 usefulness of, *III:414*
 use of, *I:175–176, I:185–186, II:371, II:733–734, III:388*
 and volatility, *I:179*
 weights in, *II:363–364*
 GARCH (1,1) model
 Bayesian estimation of, *I:176–180*
 defined, *II:364*
 results from, *II:366, II:366f*
 skewness of, *III:390–391*
 strengths of, *III:388–389*
 Student's *t*, *I:182*
 use of, *I:550–551, III:656–657*
 GARCH (1,1) process, *I:551t*
 Garman-Kohlhagen system, *I:510–511, I:522*
 Gaussian density, *III:98f*
 Gaussian model, *III:547–548*
 Gaussian processes, *III:280, III:504*
 Gaussian variables, and Brownian motion, *III:480–481*
 Gauss-Markov theorem, *II:314*
 GBM (geometric Brownian motion), *I:95, I:97*
 GDP (gross domestic product), *I:278, I:282, II:138, II:140*
 General inverse Gaussian (GIG) distribution, *II:523–524*
 Generalized autoregressive conditional heteroskedastic (GARCH) models. *See* GARCH (generalized autoregressive conditional heteroskedastic) models
 Generalized central limit theorem, *III:237, III:239*
 Generalized extreme value (GEV) distribution, *II:745, III:228–230, III:272–273*
 Generalized inverse Gaussian distribution, use of, *II:521–522*
 Generalized least squares (GLS), *I:198–199, II:328*
 Generalized tempered stable (GTS) processes, *III:512*
 Generally accepted accounting principles (GAAP), *II:21–22, II:531–532, II:542–543*
 Geometric mean reversion (GMR) model, *I:91–92*
 computation of, *I:91*
 Gibbs sampler, *I:172n, I:179, I:184–185*
 GIG models, calibration of, *II:526–527*
 Gini index of dissimilarity (Gini measure), *III:353–354*
 Ginnie Mae/Fannie Mae/Freddie Mac, actions of, *III:49*
 Girsanov's theorem
 and Black-Scholes option pricing formula, *I:132–133*
 with Brownian motion, *III:511*
 and equivalent martingale measures, *I:130–133*
 use of, *I:263, III:517*
 Glivenko-Cantelli theorem, *III:270, III:272, III:348n, III:646*
 Global Economy Workshop, Santa Fe Institute, *II:699*
 Global Industry Classification Standard (GICS®), *II:36–37, II:248*
 Global minimum variance (GMV) portfolios, *I:39*
 GMR (geometric mean reversion) model, *I:91–92*
 GMV (global minimum variance) portfolios, *I:15, I:194–195*
 GNP, growth rate of (1947–1991), *II:410–411, II:410f*
 Gradient methods, use of, *II:684*
 Granger causality, *II:395–396*
 Graphs, in MATLAB, *III:428–433*
 Greeks, the, *I:516–522*
 beta and omega, *I:522*
 delta, *I:516–518*
 gamma, *I:518–520*
 rho, *I:521–522*
 theta, *I:509, I:520–521*
 use of, *I:559, II:660, III:643–644*
 vega, *I:521*
 Greenspan, Alan, *I:140–141*
 Growth, *I:283f, II:239, II:597–598, II:601–602*
 Gumbel distribution, *III:265, III:267, III:268–269*

- Hamilton-Jacobi equations, *II:675*
- Hankel matrices, *II:512*
- Hansen-Jagannathan bound, *I:59, I:61–62*
- Harrison, Michael, *II:476*
- Hazard, defined, *III:85*
- Hazard (failure) rate, calculation of, *III:94–95*
- Heat diffusion equation, *II:470*
- Heath-Jarrow-Morton framework, *I:503, I:557*
- Heavy tails, *III:227, III:382*
- Hedge funds, and probit regression model, *II:349–350*
- Hedge ratios, *I:416–417, I:509*
- Hedges
 - importance of, *I:300*
 - improvement using DTS, *I:398*
 - in the Merton context, *I:409*
 - rebalancing of, *I:519*
 - risk-free, *I:532f*
- Hedge test, *I:409, I:411*
- Hedging
 - costs of, *I:514, II:725*
 - and credit default swaps, *I:413–414*
 - determining, *I:303–304*
 - with forward contracts, *I:429, I:429t*
 - of fuel costs, *I:561*
 - with futures, *I:433*
 - gamma, *I:519*
 - portfolio-level, *I:412–413*
 - of positions, *II:724–726*
 - ratio for, *II:725*
 - with swaps, *I:434–435*
 - transaction-level, *I:412*
 - usefulness of, *I:418*
 - use of, *I:125–126*
 - using macroeconomic indices, *I:414–417*
- Hessian matrix, *I:23–24, I:25, I:186n, III:645*
- Heston model, *I:547, I:548, I:552, II:682*
 - with change of time, *III:522*
- Heteroskedasticity, *II:220, II:359, II:360, II:403*
- HFD (high-frequency data). *See* high-frequency data (HFD)
- Higham's projection algorithm, *II:446*
- High-dimensional problems, *II:673*
- High-frequency data (HFD)
 - and bid-ask bounce, *II:454–457*
 - defined, *II:449–450*
 - generalizations to, *II:368–370*
 - Level I, *II:451–452, II:452f, II:453t*
 - Level II, *II:451*
 - properties of, *II:451, II:453t*
 - recording of, *II:450–451*
 - time intervals of, *II:457–462*
 - use of, *II:300, II:481*
 - volume of, *II:451–454*
- Hilbert spaces, *II:683*
- Hill estimator, *II:747, III:273–274*
- Historical method
 - drawbacks of, *III:413*
 - weighting of data in, *III:397–398*
- Hit rate, calculation of, *II:240n*
- HJM framework, *I:498*
- HJM methodology, *I:496–497*
- Holding period return, *I:6*
- Ho-Lee model
 - continuous variant for, *I:497*
 - defined, *I:492*
 - in history, *I:493*
 - interest rate lattice, *III:614f*
 - as short rate model, *III:23*
 - for short rates, *III:605*
 - as single factor model, *III:549*
- Home equity prepayment (HEP) curve, *III:55–56, III:56f*
- Homeowners, refinancing behavior of, *III:25*
- Home prices, *I:412, II:397f, II:399t, III:74–75*
- Homoskedasticity, *II:360, II:373*
- Horizon prices, *III:598*
- Housing, *II:396–399, III:48*
- Howard algorithm (policy iteration algorithm), *II:676–677, II:680*
- Hull-White (HW) models
 - binomial lattice, *III:610–611*
 - for calibration, *II:681*
 - defined, *I:492*
 - interest rate lattice, *III:614f*
 - and short rates, *III:545–546*
 - for short rates, *III:605*
 - trinomial lattice, *III:613, III:616f*
 - usefulness of, *I:503*
 - use of, *III:557, III:604*
 - valuing zero-coupon bond calls with, *I:500*
- Hume, David, *I:140*
- Hurst, Harold, *II:714*
- Hypercubes, use of, *III:648*
- IBM stock, log returns of, *II:407f*
- Ignorance, prior, *I:153–154*
- Implementation risk, *II:694*
- Implementation shortfall approach, *III:627*
- Implicit costs, *III:631*
- Implicit Euler (IE) scheme, *II:674, II:677–678*
- Implied forward rates, *III:565–567*
- Impurity, measures of, *II:377*
- Income, defined for public corporation, *II:21–22*
- Income statements
 - common-size, *II:562–563, II:562t*
 - defined, *II:536*
 - in financial statements, *II:536–537*
 - sample, *II:537t, II:547t*
 - structure of, *II:536*
 - XYZ Inc. (example), *II:28t*
- Income taxes. *See* taxes
- Independence, *I:372–373, II:624–625, III:363–364, III:368*
- Independence function, in VaR models, *III:365–366*
- Independently and identically distributed (IDD) concept, *I:164, I:171, II:127, III:274–280, III:367, III:414*
- Indexes
 - characteristics of efficient, *I:42t*
 - defined, *II:67*
 - of dissimilarity, *III:353–354*
 - equity, *I:15t, II:190t, II:262–263*
 - tail, *II:740–741, II:740f, III:234*
 - tracking of, *II:64, II:180*
 - use of weighted market cap, *I:38*
 - value weighted, *I:76–77*
 - volatility, *III:550–552, III:552f*
- Index returns, scenarios of, *II:190t, II:191t*
- Indifference curves, *I:4–5, I:5f, I:14*
- Industries, characteristics of, *II:36–37, II:39–40*
- Inference, *I:155–158, I:169t*
- Inflation
 - effect on after-tax real returns, *I:286–287*
 - and GDP growth, *I:282*
 - indexing for, *I:278–279*
 - in regression analysis, *II:323*
 - risk of, *II:282*
 - risk premiums for, *I:280–283*
 - seasonal factors in, *I:292*
 - shifts in, *I:285f*
 - volatility of, *I:281*
- Information
 - anticipation of, *III:476*
 - from arrays in MATLAB, *III:421*
 - completeness of, *I:353–354*
 - contained in high volatility stocks, *III:629*
 - and filtration, *III:517*
 - found in data, *II:486*
 - and information propagation, *II:515*
 - insufficient, *III:44*
 - integration of, *II:481–482*
 - overload of, *II:481*
 - prior in Bayesian analysis, *I:151–155, I:152*
 - propagation of, *I:104*

- Information (*Continued*)
 structures of, I:106f, II:515–517
 unstructured *vs.* semistructured,
 II:481–482
- Information coefficients (ICs), II:98–99,
 II:221–223, II:223f, II:227f, II:234
- Information ratios
 defined, II:86n, II:115, II:119, II:237
 determining, II:100f
 for portfolio sorts, II:219
 use of, II:99–100
- Information sets, II:123
- Information structures
 defined, II:518
- Information technology, role of,
 II:480–481
- Ingersoll models, I:271–273, I:275f
- Initial conditions, fixing of, II:502
- Initial margins, I:478
- Initial value problems, II:639
- Inner quartile range (IQR), II:494
- Innovations, II:126
- Insurance, credit, I:413–414
- Integrals, II:588–590, II:593. *See also*
 stochastic integrals
- Integrated series, and trends,
 II:512–514
- Integration, stochastic, III:472, III:473,
 III:483
- Intelligence, general, II:154
- Intensity-based frameworks, and the
 Poisson process, I:315
- Interarrival time, III:219, III:225
- Intercepts, treatment of, II:334–335
- Interest
 accumulated, II:604–605, II:604f
 annual *vs.* quarterly compounding,
 II:599f
 compound, II:597, II:597f
 computing accrued, and clean price,
 I:214–215
 coverage ratio, II:560
 defined, II:596
 determining unknown rates,
 II:601–602
 effective annual rate (EAR),
 II:616–617
 mortgage, II:398
 simple *vs.* compound, II:596
 terms of, II:619
 from TIPS, I:277
- Interest rate models
 binomial, III:173–174, III:174f
 classes of, III:600
 confusions about, III:600
 importance of, III:600
 properties of lattices, III:610
 realistic, arbitrage-free, III:599
 risk-neutral/arbitrage-free, III:597
- Interest rate paths, III:6–9, III:7, III:8t
- Interest rate risk, III:12–14
- Interest rates
 absolute *vs.* relative changes in,
 III:533–534
 approaches in determining future,
 III:591
 binomial model of, III:173–174
 binomial trees, I:236, I:236f, I:237f,
 I:240f, I:244, I:244f, III:174f
 borrowing *vs.* lending, I:482–483
 calculation of, II:613–618
 calibration of, I:495
 caps/caplets of, III:589–590
 caps on, I:248–249
 categories of term structure, III:561
 computing sensitivities, III:22–23
 continuous, I:428, I:439–488
 derivatives of, III:589–590
 determination of appropriate,
 I:210–211
 distribution of, III:538–539
 dynamic of process, I:262
 effect of, I:514–515
 effect of shocks, III:23
 effect on putable bonds, III:303–304
 future course of, III:567, III:573
 and futures prices, I:435n
 importance of models, III:600
 jumps of, III:539–541
 jumpy and continuous, III:539f
 long *vs.* short, III:538
 market spot/forward, I:495t
 mean reversion of, III:7
 modeling of, I:261–265, I:267, I:318,
 I:491, I:503, III:212–213
 multiple, II:599–600
 negative, III:538
 nominal, II:615–616
 and option prices, I:486–487
 and prepayment risk, III:48
 risk-free, I:442
 shocks/shifts to, III:585–596
 short-rate, I:491–494, III:595
 simulation of, III:541
 stochastic, I:344, I:346
 structures of, III:573, III:576
 use of for control, I:489
 volatility of, III:405, III:533
- Intermarket relations, no-arbitrage,
 I:453–455
- Internal consistency rule, in OAS
 analysis, I:265
- Internal rate of return (IRR), II:617–618
 in MBSs, III:36
- International Monetary Fund
 Global Stability Report, I:299
- International Swap and Derivatives
 Association (ISDA). *See* ISDA
- Interpolated spread (I-spread), I:227
- Interrate relationship, arbitrage-free,
 III:544
- Intertemporal dependence, and risk,
 III:351
- Intertrade duration, II:460–461,
 II:462t
- Intertrade intervals, II:460–461
- Intervals, credible, I:170
- Interval scales, data on, II:487
- Intrinsic value, I:441, I:511, I:513,
 II:16–17
- Invariance property, III:328–329
- Inventory, II:542, II:557
- Inverse Gaussian process, III:499
- Investment, goals of, II:114–115
- Investment management, III:146
- Investment processes
 activities of integrated, II:61
 evaluation of results of, II:117–118
 model creation, II:96
 monitoring of performance, II:104
 quantitative, II:95, II:95f
 quantitative equity, II:95f, II:96f,
 II:105
 research, II:95–102
 sell-structured, II:108
 steps for equity investment, II:119
 testing of, II:109
- Investment risk measures, III:350–351
- Investments, I:77–78n, II:50–51,
 II:617–618
- Investment strategies, II:66–67,
 II:198
- Investment styles, quantamental,
 II:93–94, II:93f
- Investors
 behavior of, II:207, II:504
 comfort with risk, I:193
 completeness of information of,
 I:353–354
 focus of, I:299, II:90–91
 fundamental *vs.* quantitative,
 II:90–94, II:91f, II:92f, II:105
 goals/objectives of, II:114–115,
 II:179, III:631
 individual accounts of, II:74
 monotonic preferences of, I:57
 number of stocks considered, II:91
 preferences of, I:5, I:260, II:48, II:56,
 II:92–93
 prior beliefs of, II:727
 real-world, II:132
 risk aversion of, II:82–83, II:729
 SL-CAPM assumptions about, I:66
 sophistication of, II:108
 in uncertain markets, II:54
 views of, I:197–199
- Invisible hand, notion of, II:468–469

- ISDA (International Swap and Derivatives Association)
 Credit Derivative Definitions (1999), *I:230, I:528*
 Master Agreement, *I:538*
 organized auctions, *I:526–527*
 supplement definition, *I:230*
- I-spread (interpolated spread), *I:227*
- Ito, Kiyosi, *II:470*
- Ito definition, *III:486–487*
- Ito integrals, *I:122, III:475, III:481, III:490–491*
- Ito isometry, *III:475*
- Ito processes
 defined, *I:95*
 generic univariate, *I:125*
 and Girsanov's theorem, *I:131*
 under HJM methodology, *I:497*
 properties of, *III:487–488*
 and smooth maps, *III:493*
- Ito's formula, *I:126, III:488–489*
- Ito's lemma
 defined, *I:98*
 discussion of, *I:95–97*
 in estimation, *I:348*
 and the Heston model, *I:548*
- James-Stein shrinkage estimator, *I:194*
- Japan, credit crisis in, *I:417*
- Jarrow-Turnbull model, *I:307*
- Jarrow-Yu propensity model, *I:324–325*
- Jeffreys' prior, *I:153, I:160n, I:171–172*
- Jensen's inequality, *I:86, III:569*
- Jevons, Stanley, *II:468*
- Johansen-Juselius cointegration tests, *II:391–393, III:395*
- Joint jumps/defaults, *I:322–324*
- Joint survival probability, *I:323–324*
- Jordan diagonal blocks, *II:641–642*
- Jorion shrinkage estimator, *I:194, I:202*
- Jump-diffusion, *III:554–557, III:657*
- Jumps
 default, *I:322–324*
 diffusions, *I:559–560*
 downward, *I:347*
 idiosyncratic, *I:323*
 incorporation of, *I:93–94*
 in interest rates, *III:539–541*
 joint, *I:322–324*
 processes of, *III:496*
 pure processes, *III:497–501, III:506*
 size of, *III:540*
- Kalotay-Williams-Fabozzi (KWF) model, *III:604, III:606–607, III:615f*
- Kamakura Corporation, *I:301, I:307, I:308–309, I:310n*
- Kappa, *I:521*
- Karush-Kuhn-Tucker conditions (KKT conditions), *I:28–29*
- Kendall's tau, *I:327, I:332*
- Kernel regression, *II:403, II:412–413, II:415*
- Kernels, *II:412, II:413f, II:746*
- Kernel smoothers, *II:413*
- Keynes, John Maynard, *II:471*
- Key rate durations (KRD), *II:276, III:311–315, III:317*
- Key rates, *II:276, III:311*
- Kim-Rachev (KR) process, *III:512–513*
- KKT conditions (Karush-Kuhn-Tucker conditions), *I:28–29, I:31, I:32*
- KoBoL distribution, *III:257n*
- Kolmogorov extension theorem, *III:477–478*
- Kolmogorov-Smirnov (KS) test, *II:430, III:366, III:647*
- Kolmogorov equation, use of, *III:581*
- Kreps, David, *II:476*
- Krispy Kreme Doughnuts, *II:574–575, II:574f*
- Kronecker product, *I:172, I:173n*
- Kuiper test, *III:366*
- Kurtosis, *I:41, III:234*
- Lag operator L , *II:504–506, II:507, II:629–630*
- Lagrange multipliers, *I:28, I:29–31, I:30, I:32*
- Lag times, *II:387, III:31*
- Laplace transforms, *II:647–648*
- Last trades, price and size of, *II:450*
- Lattice frameworks
 bushy trees in, *I:265, I:266f*
 calibration of, *I:238–240*
 fair, *I:235*
 interest rate, *I:235–236, I:236–238*
 one-factor model, *I:236f*
 for pricing options, *I:487*
 usefulness of, *I:235*
 use of, *I:240, I:265–266, III:14*
 value at nodes, *I:237–238*
 1-year rates, *I:238f, I:239f*
- Law of iterated expectations, *I:110, I:122, II:308*
- Law of large numbers, *I:267, I:270n, III:263–264, III:275*
- Law of one α , *II:50*
- Law of one price (LOP), *I:52–55, I:99–100, I:102, I:119, I:260*
- LCS (liquidity cost score), *I:402*
 use of, *I:403*
- LDIs (liability-driven investments), *I:36*
- LD (loss on default), *I:370–371*
- Leases, in financial statements, *II:542*
- Least-square methods, *II:683–685*
- Leavens, D. H., *I:10*
- Legal loss data
 Cruz study, *III:113, III:115t*
 Lewis study, *III:117, III:117t*
- Lehman Brothers, bankruptcy of, *I:413*
- Level (parallel) effect, *II:145*
- Lévy-Khinchine formula, *III:253–254, III:257*
- Lévy measures, *III:254, III:254t*
- Lévy processes
 and Brownian motion, *III:504*
 in calibration, *II:682*
 change of measure for, *III:511–512*
 conditions for, *III:505*
 construction of, *III:506*
 from Girsanov's theorem, *III:511*
 and Poisson process, *III:496*
 as stochastic process, *III:505–506*
 as subordinators, *III:521*
 for tempered stable processes, *III:512–514, III:514t*
 and time change, *III:527*
- Lévy stable distribution, *III:242, III:339, III:382–386, III:392*
- LGD (loss given default), *I:366, I:370, I:371*
- Liabilities, *II:533, II:534–535, III:132*
- Liability-driven investments (LDIs), *I:36*
- Liability-hedging portfolios (LHPs), *I:36*
- LIBOR (London Interbank Offered Rate)
 and asset swaps, *I:227*
 changes in, by type, *III:539–540*
 curve of, *I:226*
 interest rate models, *I:494*
 market model of, *III:589*
 spread of, *I:530*
 in total return swaps, *I:541*
 use of in calibration, *III:7*
- Likelihood maximization, *I:176*
- Likelihood ratio statistic, *II:425*
- Limited liability rule, *I:363*
- Limit order books, use of, *III:625, III:632n*
- Lintner, John, *II:474*
- Lipschitz condition, *II:658n, III:489, III:490*
- Liquidation
 effect of, *II:186*
 procedures for, *I:350–351*
 process models for, *I:349–351*
 time of, *I:350*
 vs. default event, *I:349*
- Liquidity
 assumption of, *III:371*
 in backtesting, *II:235*
 changes in, *I:405*

- Liquidity (*Continued*)
 cost of, I:401
 creation of, III:624–625, III:631
 defined, III:372, III:380
 effect of, II:284
 estimating in crises, III:378–380
 in financial analysis, II:551–555
 and LCS, I:404
 and market costs, III:624
 measures of, II:554–555
 premiums on, I:294, I:307
 ratios for, II:555
 in risk modeling, II:693
 shortages in, I:347–348
 and TIPS, I:293, I:294
 and transaction costs, III:624–625
- Liquidity-at-risk (LAR), III:376–378
- Liquidity cost, III:373–374, III:375–376
- Liquidity cost score (LCS), I:402, I:403
- Liquidity preference hypothesis, III:570
- Liquidity ratios, II:563
- Liquidity risk, II:282, III:380
- Ljung-Box statistics, II:407, II:421, II:422, II:427–428
- LnMix models, calibration of, II:526–527
- Loading, standardization of, II:177
- Loan pools, III:8–9
- Loans
 amortization of, II:606–607, II:611–613
 amortization table for, II:612t
 delinquent, III:63
 fixed rate, fully amortized schedule, II:614t
 floating rate, II:613
 fully amortizing, II:611
 modified, III:32
 nonperforming, III:75
 notation for delinquent, III:45n
 recoverability of, III:31–32
 refinancing of, III:68–69
 repayment of, II:612f, II:613f
 term schedule, II:615t
- Loan-to-value ratios (LTVs), III:31–32, III:69, III:73, III:74–75
- Location parameters, I:160n, III:201–202
- Location-scale invariance property (Gaussian distribution), II:732
- Logarithmic Ornstein-Uhlenbeck (log-OU) processes, I:557–558
- Logarithmic returns, III:211–212, III:225
- Logistic distribution, II:350
- Logistic regression, I:307, I:308, I:310
- Logit regression models, II:349–350, II:350
- Log-Laplace transform, III:255–256
- Lognormal distribution, III:222–225, III:392
- Lognormal mixture (LnMix) distribution, II:524–525
- Lognormal variables, I:86
- Log returns, I:85–86, I:88
- London Interbank Offered Rate (LIBOR). *See* LIBOR
- Lookback options, I:114, III:24
- Lookback periods, III:402, III:407
- LOP (law of one price). *See* law of one price (LOP)
- Lorenz, Edward, II:653
- Loss distributions, conditional, III:340–341
- Losses. *See also* operational losses
 allocation of, III:32
 analysis of in backtesting, III:338
 collateral *vs.* tranche, III:36
 computation of, I:383
 defined, III:85
 estimation of cumulative, III:39–40
 expected, I:369–370, I:373–374
 expected *vs.* unexpected, I:369, I:375–376
 internal *vs.* external, III:83–84
 median of conditional, III:348n
 projected, III:37f
 restricting severity of, I:385–386
 severity of, III:44
 unexpected, I:371–372, I:374–375
- Loss functions, I:160n, III:369
- Loss given default (LGD), I:366, I:370, I:371
- Loss matrix analysis, III:40–41
- Loss on default (LD), I:370–371
- Loss severity, III:30–31, III:60–62, III:97–99
- Lottery tickets, I:462
- Lower partial moment risk measure, III:356
- Lundbert, Filip, II:467, II:470–471
- Macroeconomic influences, defined, II:197
- Magnitude measures, II:429–430
- Maintenance margins, I:478
- Major indexes, modeling return distributions for, III:388–392
- Malliavin calculus, III:644
- Management, active, II:115
- Mandelbrot, Benoit, II:653, II:738, III:234, III:241–242
- Manufactured housing prepayment (MHP) curve, III:56
- Marginalization, II:335
- Marginal rate of growth, III:197–198
- Marginal rate of substitution, I:60
- Margin calls, exposure to, III:377
- Market cap *vs.* firm value, II:39
- Market completeness, I:52, I:105
- Market efficiency, I:68–73, II:121, II:473–474
- Market equilibrium
 and investor's views, I:198–199
- Market impact
 costs of, III:623–624, III:627
 defined, II:69
 forecasting/modeling of, III:628–631
 forecasting models for, III:632
 forecasting of, III:628–629, III:629–631
 measurement of, III:626–628
 between multiple accounts, II:75–76
 in portfolio construction, II:116
 and transaction costs, II:70
- Market model regression, II:139
- Market opportunity, two state, I:460f
- Market portfolios, I:66–67, I:72–73
- Market prices, I:57, III:372
- Market risk
 approaches to estimation of, III:380
 in bonds, III:595
 in CAPM, I:68–69, II:474
 importance of, III:81
 models for, III:361–362
 premium for, I:203n, I:404
- Markets
 approach to segmented, II:48–51
 arbitrage-free, I:118
 complete, I:51–52, III:578
 complex, II:49
 effect of uncertainty in on bid-ask spreads, II:455–456
 efficiency of, II:15–16
 frictionless, I:261
 incomplete, I:461–462
 liquidity of, III:372
 models of, III:589
 for options and futures, I:453–454
 perfect, II:472
 properties of modern, III:575–576
 sensitivities to value-related variables, II:54t
 simple, I:70
 systematic fluctuations in, II:172–173
 unified approach to, II:49
 up/down, defined, II:347
- Market sectors, defined, III:560
- Market standards, I:257
- Market structure, and exposure, II:269–270
- Market timing, II:260
- Market transactions, upstairs, III:630–631, III:632n

- Market weights, *II:269t*
- Markov chain approximations, *II:678*
- Markov chain Monte Carlo (MCMC)
methods, *II:410f, II:417–418*
- Markov coefficients, *II:506–507, II:512*
- Markov matrix, *I:368*
- Markov models, *I:114*
- Markov processes
in dynamic term structures, *III:579*
hidden, *I:182*
use of, *III:509, III:517*
- Markov property, *I:82, I:180–181, I:183, II:661, III:193n*
- Markov switching (MS) models
discussion of, *I:180–184*
and fat tails, *III:277–278*
stationarity with, *III:275*
usefulness of, *II:433*
use of, *II:409–411, II:411t*
- Markowitz, Harry M., *I:38, I:140, II:467, II:471–472, III:137, III:351–352*
- Markowitz constraint sets, *I:69, I:72*
- Markowitz diversification, *I:10–11, I:11*
- Markowitz efficient frontiers, *I:191f*
- Markowitz model
in financial planning, *III:126*
- Mark-to-market (MTM)
calculation of value, *I:535–536, I:536t*
defined, *I:535*
and telescoping futures, *I:431–432*
- Marshall and Siegel, *II:694*
- Marshall-Olkin copula, *I:323–324, I:329*
- Martingale measures, equivalent
and arbitrage, *I:111–112, I:124*
and complete markets, *I:133*
defined, *I:110–111*
and Girsanov's theorem, *I:130–133*
and state prices, *I:133–134*
use of, *I:130–131*
working with, *I:135*
- Martingales
with change of time methods
(CTM), *III:522–523*
defined, *II:124, II:126, II:519*
development of concept, *II:469–470*
equivalent, *II:476*
measures of, *I:110–111*
use of conditions, *I:116*
use of in forward rates, *III:586*
- Mathematical theory, importance of
advances in, *III:145*
- Mathworks, website of, *III:418*
- MATLAB
array operations in, *III:420–421*
basic mathematical operations in,
III:419–420
- construction of vectors/matrices,
III:420
- control flow statements in,
III:427–428
- desktop, *III:419f*
- European call option pricing with,
III:444–445
- functions built into, *III:421–422*
- graphs in, *III:428–433, III:429–430f, III:431f*
- interactions with other software,
III:433–434
- M-files in, *III:418–419, III:423, III:447*
- operations in, *III:447*
- optimization in, *III:434–444, III:435t*
- Optimization Tool, *III:435–436, III:436f, III:440f, III:441f*
- overview of desktop and editor,
III:418–419
- quadprog function, *II:70*
- quadratic optimization with,
III:441–444
- random number generation,
III:444
- for simulations, *III:651*
- Sobol sequences in, *III:445–446*
- for stable distributions, *III:344*
- surf function in, *III:432–433*
- syntax of, *III:426–427*
- toolboxes in, *III:417–418*
- user-defined functions in,
III:423–427
- Matrices
augmented, *II:624*
characteristic polynomial of, *II:628*
coefficient, *II:624*
companion, *II:639–640*
defined, *II:622*
diagonal, *II:622–623, II:640*
eigenvalues of random, *II:704–705*
eigenvectors of, *II:640–641*
in MATLAB, *III:422, III:432*
operations on, *II:626–627*
ranks of, *II:623, II:628*
square, *II:622–623, II:626–627*
symmetric, *II:623*
traces of, *II:623*
transition, *III:32–33, III:32t, III:33t, III:35f*
types of, *II:622, II:628*
- Matrix differential equations, *III:492*
- Maturity value (lump sum), from
bonds, *I:211*
- Maxima, *III:265–269, III:266f*
- Maximum Description Length
principle, *II:703*
- Maximum eigenvalue test, *II:392–393*
- Maximum likelihood (ML)
approach, *I:141, I:348*
methods, *II:348–349, II:737–738, III:273*
principal, *II:312*
- Maximum principle, *II:662, II:667*
- Max-stable distributions, *III:269, III:339–340*
- MBA (Mortgage Bankers Association)
refi index, *III:70, III:70f*
- MBS (mortgage-backed securities),
I:258
agency vs. nonagency, *III:48*
cash flow characteristics of, *III:48*
default assumptions about, *III:8*
negative convexity of, *III:49*
performance of, *III:74*
prices of, *III:26*
projected long-term performance of,
III:34f
time-related factors in, *III:73–74*
valuation of, *III:62*
valuing of, *III:645*
- MBS (mortgage-backed securities),
nonagency
analysis of, *III:44–45*
defined, *III:48*
estimation of returns, *III:36–44*
evaluation of, *III:29*
factors impacting returns of,
III:30–32
yield tables for, *III:41t*
- Mean absolute deviation (MAD),
III:353
- Mean absolute moment (MAM(q)),
III:353
- Mean colog (M-colog), *III:354*
- Mean entropy (M-entropy), *III:354*
- Mean excess function, *II:746–747*
- Mean/first moment, *III:201–202*
- Mean residual life function, *II:754n*
- Mean reversion
discussion of, *I:88–92*
geometric, *I:91–92*
in HW models, *III:605*
and market stability, *III:537–538*
models of, *I:97*
parameter estimation, *I:90–91*
risk-neutral asset model, *III:526*
simulation of, *I:90*
in spot rate models, *III:580*
stabilization by, *III:538*
within a trinomial setting, *III:604*
- Mean-reverting asset model (MRAM),
III:525–526
- Means, *I:148, I:155, I:380, III:166–167*
- Mean-variance
efficiency, *I:190–191*
efficient portfolios, *I:13, I:68, I:69–70*

- Mean-variance (*Continued*)
 nonrobust formulation, III:139–140
 optimization, I:192
 constraints on, I:191
 estimation errors and, I:17–18
 practical problems in, I:190–194
 risk aversion formulation, II:70
 Mean variance analysis, I:3, I:15f,
 I:201, II:471–472, III:352
 Measurement levels, in descriptive
 statistics, II:486–487
 Media effects, III:70
 Median, I:155, I:159n, II:40
 Median tail loss (MTL), III:341
 Mencken, H. L., II:57
 Menger, Carl, II:468
 Mercurio-Moraleda model, I:493–494
 Merton, Robert, I:299, I:310, II:468,
 II:475, II:476
 Merton model
 advantages and criticisms of,
 I:344
 applied to probability of default,
 I:363–365
 with Black-Scholes approach,
 I:305–306
 default probabilities with, I:307–308
 discussion of, I:343–344
 drawbacks of, I:410
 with early default, I:306
 evidence on performance, I:308–309
 as first modern structural model,
 I:313, I:341
 in history, I:491
 with jumps in asset values, I:306
 portfolio-level hedging with,
 I:411–413
 with stochastic interest rates, I:306
 and transaction-level hedging,
 I:408–410
 usefulness of, I:410, I:411–412,
 I:417–418
 use of, I:304, I:305, I:510
 variations on, I:306–307
 Methodology, equally weighted,
 III:399
 Methods
 quantile, II:354–356
 Methods pathwise, III:643
 Metropolis-Hastings (M-H) algorithm,
 I:178
 M-H algorithm, I:179
 MIB 30, III:402–403, III:402f, III:403f
 Microsoft, II:722f. *See also* Excel
 Midsquare technique, III:647
 Migration mode
 calculation of expected/unexpected
 losses under, I:376t
 expected loss under, I:373–374
 Miller, Merton, II:467, II:473
 MiniMax (MM) risk measure, III:356
 Minimization problems, solutions to,
 II:683–684
 Minimum-overall-variance portfolio,
 I:69
 Minority interest, on the balance
 sheet, II:536
 Mispricing, risk of, II:691–692
 Model creep, II:694
 Model diagnosis, III:367–368
 Model estimation, in non-IDD
 framework, III:278
 Modeling
 calibration of structure, III:549–550
 changes in mathematical, II:480–481
 discrete *vs.* continuous time, III:562
 dynamic, II:105
 issues in, II:299
 nonlinear time series, II:427–428,
 II:430–433
 quantitative, II:481
 Modeling techniques
 non-parametric/nonlinear, II:375
 Model risk
 of agency ratings, II:728–729
 awareness of, I:145, II:695–696
 with computer models, II:695
 consequences of, II:729–730
 contribution to bond pricing,
 II:727–728
 defined, I:331, II:691, II:697
 discussion of, II:714–715
 diversification of, II:378
 endogenous, II:694–695, II:697
 in financial institutions, II:693
 guidelines for institutions,
 II:696–697
 management of, II:695–697, II:697
 misspecification of, II:199
 and robustness, II:301
 of simple portfolio, II:721–726
 sources of, II:692–695
 Models. *See also* operational risk
 models
 accuracy in, III:321
 adjustment, II:502
 advantages of reduced-form, I:533
 analytical tractability of, III:549–550
 APD, III:18, III:20–22, III:21f, III:26
 application of, II:694
 appropriate use of classes of,
 III:597–598
 arbitrage-free, III:600
 autopredictive, II:502
 averages across, II:715
 bilinear, II:403–404
 binomial, I:114–116, I:119
 binomial stochastic, II:10–11
 block maxima, II:745
 choosing, III:550–552
 comparison of, III:617
 compatibility of, III:373
 complexity of, II:704, II:717
 computer, I:511, II:695
 conditional normal, II:733–734
 conditional parametric fat-tailed,
 II:744
 conditioning, II:105
 construction of, II:232–235
 for continuous processes, I:123
 creation of, II:100–102
 cross-sectional, II:174–175, II:175t
 cumulative return of, II:234
 defined, II:691, II:697
 to describe default processes, I:313
 description and estimation of,
 II:256–257
 designing the next, III:590–591
 determining, II:299–300
 disclosure of, I:410
 documentation of, II:696
 dynamic factor, II:128, II:131,
 III:126–127
 dynamic term structure, III:591
 econometric, II:295, II:304
 equilibrium forms of, III:599–600
 equity risk, II:174, II:178–191, II:192
 error correction, II:381t, II:387–388,
 II:394–395
 evidence of performance, I:308–309,
 II:233
 examples of multifactor, II:139–140
 financial, I:139, II:479–480
 forecasting, II:112, II:303–304
 for forecasting, III:411
 formulation of, III:128–131
 fundamental factor, II:244, II:248
 generally, II:360–362
 Gordon-Shapiro, II:17–18
 Heath-Jarrow-Morton, III:586–587,
 III:589
 hidden-variable, II:128, II:131
 linear, II:264, II:310–311, II:348,
 II:507–508
 linear autoregressive, II:128,
 II:130–131
 linear regression, I:91, I:163–170,
 II:360, II:414–415
 liquidation process, I:342
 martingale, II:127–128, III:520–521
 MGARCH, II:371–372
 model-vetting procedure, II:696–697
 moving average, III:414
 multifactor, II:231–232, III:92
 multivariate extensions of,
 II:370–373
 no arbitrage, III:604

- nonlinear, *II:402–421, II:417–418*
 penalty functions in, *II:703*
 performance measurement of, *II:301*
 predictive regressive, *II:130*
 predictive return, *II:128–131*
 for pricing, *II:127–128*
 pricing errors in, *I:322*
 principals for engineering,
II:482–483
 probabilistic, *II:299*
 properties of good, *I:320*
 ranking alternative, *III:368–370*
 recalibration of, *II:713–714*
 reduced form default, *I:310, I:313*
 regressive, *II:128, II:129–130*
 relative valuation, *I:260*
 return forecasting, *II:119*
 returns of, *II:233f*
 robustness of, *II:301*
 selection of, *I:145, II:298, II:692–693,*
II:699–701
 short-rate, *I:494*
 single-index market, *II:317–318*
 static, *II:297, III:573*
 static regressive, *II:129–130*
 static *vs.* dynamic, *II:295–296, II:304*
 statistical, *II:175, II:175t*
 stochastic, *I:557, III:124–125*
 structural, *I:305, I:313–314, I:341–342*
 structural *vs.* reduced, *I:532–533*
 subordinated, *II:742–743*
 temporal aggregation of, *II:369*
 testing of, *II:126–127, II:696–697*
 time horizon of, *II:300–301*
 time-series, *II:175, II:175t*
 tree, *II:381, III:22–23*
 tuning of, *III:580–581*
 two-factor, *I:494*
 univariate regression, *I:165*
 usefulness of, *II:122*
 use of in practice, *I:494–496, III:600t*
- Models, lattice**
 binomial, *III:610, III:610f*
 Black-Karasinski (BK) lattice, *III:611*
 Hull White binomial, *III:610–611*
 Hull White trinomial, *III:613*
 trinomial, *III:610, III:610f,*
III:611–612
- Models, selection of**
 components of, *II:717*
 generally, *II:715–717*
 importance of, *II:700*
 machine learning approach to,
II:701–703, II:717
 uncertainty/noise in, *II:716–717*
 use of statistical tools in, *II:230*
- Modified Accelerated Cost Recovery
 System (MACRS), *II:538***
- Modified Restructuring clause, *I:529***
- Modified tempered stable (MTS)
 processes, *III:513***
- Modigliani, Franco, *II:467, II:473*
- Modigliani-Miller theorem, *I:343,*
I:344, II:473, II:476
- Moment ratio estimators, *III:274*
- Moments**
 exponential, *III:255–256*
 first, *III:201–202*
 of higher order, *III:202–205*
 integration of, *II:367–368*
 raw, *II:739*
 second, *III:202*
 types of, *II:125*
- Momentum**
 formula for analysis of, *II:239*
 portfolios based on, *II:181*
- Momentum factor, *II:226–227*
- Money, future value of, *II:596–600*
- Money funds, European options on,
I:498–499
- Money markets, *I:279, I:282, I:314,*
II:244
- Monotonicity property, *III:327*
- Monte Carlo methods**
 advantages of, *II:672*
 approach to estimation, *I:193*
 defined, *I:273*
 examples of, *III:637–639*
 foundations of, *I:377–378*
 for interest rate structure, *I:494*
 main ideas of, *III:637–642*
 for nonlinear state-space modeling,
II:417–418
 stochastic content of, *I:378*
 usefulness of, *I:389*
 use of, *I:266–268, III:651*
 of VaR calculation, *III:324–325*
- Monte Carlo simulations**
 for credit loss, *I:379–380*
 effect of sampling process, *I:384*
 in fixed income valuation modeling,
III:6–12
 sequences in, *I:378–379*
 speed of, *III:644*
 use of, *III:10–11, III:642*
- Moody's diversity score, use of,
I:332
- Moody's Investors Service, *I:362*
- Moody's KMV, *I:364–365*
- Mortgage-backed securities (MBS). *See*
 MBS (mortgage-backed
 securities)
- Mortgage Bankers Association (MBA)
 method, *III:57–58*
- Mortgagee pools
 composition of, *III:52*
 defined, *III:23, III:65*
 nonperforming loans and, *III:75*
 population of, *III:19*
 seasoning of, *III:20, III:22*
- Mortgages, *III:48–49, III:65, III:69,*
III:71
- Mosaic Company, distribution of price
 changes of, *II:723f*
- Mossin, Jan, *II:468, II:474*
- Moving averages, infinite, *II:504–508*
- MSCI Barra model, *II:140*
- MSCI EM, historical distributions of,
III:391f
- MSCI-Germany Index, *I:143*
- MSCI World Index, *I:15–17*
 analysis of 18 countries, *I:16t*
- MS GARCH model, *I:185–186*
 estimation of, *I:182*
 sampling algorithm for, *I:184*
- MSR (maximum Sharpe Ratio), *I:36–37*
- MS-VAR models, *II:131*
- Multiaccount optimization, *II:75–77*
- Multicollinearity, *II:221*
- Multilayer perceptrons, *II:419*
- Multinomial/polynomial coefficients,
III:191–192
- Multivariate normal distribution, in
 MATLAB, *III:432–433, III:433f*
- Multivariate random walks, *II:124*
- Multivariate stationary series,
II:506–507
- Multivariate *t* distribution, loss
 simulation, *I:388–389*
- Nadaraya-Watson estimator, *II:412,*
II:415
- Natural conjugate priors, *I:160n*
- Navigation, fuel-efficient, *I:562–563*
- Near-misses, management of,
III:84–85
- Net cash flow, defined, *II:541*
- Net cost of carry, *I:424–425, I:428,*
I:437, I:439–440, I:455
- Net free cash flow (NFCF), *II:572–574,*
II:578
- Net profit margin, *II:556*
- Net working capital-to-sales ratio,
II:554–555
- Network investment models,
III:129–130, III:129f
- Neumann boundary condition, *II:666,*
II:671
- Neural networks, *II:403, II:418–421,*
II:418f, II:701–702
- Newey-West corrections, *II:220*
- NIG distribution, *III:257n*
- 9/11 attacks, effects of, *III:402–403*
- No-arbitrage condition, in certain
 economy, *III:567–568*
- No arbitrage models, use of, *III:604*
- No-arbitrage relations, *I:423*

- Noise
 continuous-time, III:486
 in financial models, II:721–722
 in model selection, II:716–717
 models for, II:726
 reduction of, II:51–52
- Noise, white
 defined, I:82, II:297
 qualities of, II:127
 sequences, II:312, II:313
 in stochastic differential equations, III:486
 strict, II:125
vs. colored noise, III:275
- Nonlinear additive AR (NAAR)
 model, II:417
- Nonlinear dynamics and chaos, II:645, II:652–654
- Nonlinearity, II:433
 in econometrics, II:401–403
 tests of, II:421–427
- Non-normal probability distributions, II:480
- Nonparametric methods, II:411–416
- Normal distributions, I:81, I:82*f*, I:177–178, III:638*f*
 and AVaR, III:334
 comparison with α -stable, III:234*f*
 fundamentals of, II:731–734
 inverse Gaussian, III:231–233, III:232*f*, III:233*f* (*See also* Gaussian distribution)
 likelihood function, I:142–143
 for logarithmic returns, III:211–212
 mixtures of for downside risk estimation, III:387–388
 for modeling operational risk, III:98–99
 multivariate, and tail dependence, I:387
 properties of, II:732–733, III:209–210
 relaxing assumption of, I:386–387
 standard, III:208
 standardized residuals from, II:751
 use of, II:752*n*
 using to approximate binomial distribution, III:211
 for various parameter values, III:209*f*
vs. normal inverse Gaussian distribution, III:232–233
- Normal mean, and posterior tradeoff, I:158–159
- Normal tempered stable (NTS) processes, III:513
- Normative theory, I:3
- Notes, step-up callable, I:251–252, I:251*f*, I:252*f*
- Novikov condition, I:131–132
- NTS distribution, III:257*n*
- Null hypothesis, I:157, I:170, III:362
- Numeraire, change of, III:588–589
- Numerical approximation, I:265
- Numerical models for bonds, I:273–275
- OAS (option-adjusted spread). *See* option-adjusted spread
- Obligations, deliverable, I:231, I:526
- Observations, frequency of, III:404
- Occam's razor, in model selection, II:696
- Odds ratio, posterior, I:157
- Office of Thrift Supervision (OTS) method, III:57–58
- Oil industry, free cash flows of, II:570
- OLS (ordinary least squares). *See* ordinary least squares (OLS)
- Open classes, II:493–494
- Operating cash flow (OCF), II:23
- Operating cycles, II:551–554
- Operating profit margin, II:556
- Operational loss data
 de Fontnouvelle, Rosengren, and Jordan study, III:116–117, III:116*t*
 empirical evidence with, III:112–118
 Moscadelli study, III:113, III:116, III:116*t*
 Müller study, III:113, III:114*f*, III:115*t*
 Reynolds-Syer study, III:117–118
 Rosenberg-Schuermann study, III:118
- Operational losses
 and bank size, III:83
 definitions of types, III:84*t*
 direct *vs.* indirect, III:84–85
 expected *vs.* unexpected, III:85
 histogram of, III:104*f*
 histogram of severity distribution, III:95*f*
 historical data on, III:96
 near-miss, III:84–85
 process of arriving at data, III:96–97
 process of occurrence, III:86*f*
 recording of, III:97
 severity of, III:104*f*
 time lags in, III:96–97
 types of, III:81, III:88
- Operational loss models
 approaches to, III:103–104
 assumptions in, III:104
 nonparametric approach, III:103–104, III:104–105, III:118
 parametric approach, III:104, III:105–110, III:118
 types of, III:118
- Operational risk
 classifications of, III:83–88, III:87–88, III:87*f*, III:88
 defined, III:81–83, III:88
 event types with descriptions, III:86*t*
 indicators of, III:83
 models of, III:91–96
 nature of, III:99
 and reputational risk, III:88
 sources of, III:82
- Operational risk/event/loss types, distinctions between, III:85–87
- Operational risk models
 actuarial (statistical) models, III:95
 bottom-up, III:92*f*, III:94–96, III:99
 causal, III:94
 expense-based, III:93
 income-based, III:93
 multifactor causal models, III:95
 operating leverage, III:93
 process-based, III:94–95
 proprietary, III:96
 reliability, III:94–95
 top down, III:92–94, III:99
 types of, III:91–92
- Operations
 addition, II:625, II:626
 defined, II:628
 inverse and adjoint, II:626–627
 multiplication, II:625–626, II:626
 transpose, II:625, II:626
 vector, II:625–626
- Operators in sets, defined, III:154
- Ophelimity, concept of, II:469
- Opportunity cost, I:435, I:438, I:439, II:596, III:623
- Optimal exercise, I:515–516
- Optimization
 algorithms for, III:124
 complexity of, II:82
 constrained, I:28–34
 defined, III:434–435
 local *vs.* global, II:378
 in MATLAB, III:434–444
 unconstrained, I:22–28
- Optimization theory, I:21
- Optimization Toolbox, in MATLAB, III:435–436, III:436*f*
- Optimizers, using, II:115–116, II:483
- Option-adjusted spread (OAS)
 calculation of, I:253–255
 defined, I:254, III:11
 demonstrated, I:254*f*
 determination of, I:259
 implementation of, I:257
 and market value, I:258
 results from example, III:617*t*
 and risk factors, III:599

- rules-of-thumb for analysis, *I:264–265*
- usefulness of, *III:3*
- values of, *I:267, I:268*
- variance between dealers, *I:257–258*
- Option premium, *I:508–509*
- time/intrinsic values of, *I:513*
- Option premium profiles, *I:512, I:512f*
- Option prices
 - components of, *I:484–485, I:511–512*
 - factors influencing, *I:486–487, I:486t, I:487–488, I:522–523*
 - models for, *I:490*
- Options
 - American, *II:664–665, II:669–670, II:674–679, II:679–681*
 - American-style, *I:444, I:454–455, I:490*
 - Asian, *II:663–664, II:668–669, III:642–643*
 - on the average, *II:663–664*
 - barrier, *II:662–663*
 - basic properties of, *I:507–508*
 - basket, *II:662, II:672*
 - Bermudean, *II:663–664, III:597*
 - buying assets of, *I:439*
 - costs of, *I:441–442, III:11–12*
 - difference from forwards, *I:437–439*
 - early exercise of, *I:442–443, I:447*
 - Eurodollar, *I:489*
 - European, *I:125, I:127–129, II:660–664, II:665–674*
 - European-style, *I:444–445, I:454*
 - European-style *vs.* American-style, *I:453t, I:455n, I:508, I:515–516*
 - and expected volatility, *I:486*
 - expiration/maturity dates of, *I:484*
 - factors affecting value of, *I:474*
 - formulas for pricing, *III:522, III:527*
 - in/out of/at-the-money, *I:485*
 - long *vs.* short call, *I:437–439, I:438f*
 - lookback, *II:663, II:672, II:673f*
 - on the maximum, *II:663*
 - models of, *I:510–511*
 - no-arbitrage futures, *I:453*
 - price relations for, *I:448t*
 - pricing of, *I:124–129, I:455t, I:484–488, I:507, III:408*
 - theoretical valuation of, *I:508–509*
 - time premiums of, *I:485*
 - time to expiration of, *I:486*
 - types of, *I:484*
 - valuing of, *I:252–253, III:639*
 - vanilla, *II:661, III:655*
 - volatility of, *I:488*
- Orders
 - in differential equations, *II:643, II:644–645*
 - fleeting limit, *III:625*
 - limit, *III:625, III:631*
 - market, *III:625, III:631*
- Order statistics, *III:269–270*
- bivariate, *III:293–295*
- joint probability distributions for, *III:291–292*
- use of, *III:289*
- for VaR and ETL, *III:292t*
- in VaR calculations, *III:291*
- Ordinary differential equations (ODE), *II:644–645, II:646–648, II:648–652, II:649f*
- Ordinary least squares (OLS)
 - alternate weighting of, *II:438–439*
 - estimation of factor loadings matrix with, *II:165*
 - in maximum likelihood estimates, *II:313–314*
 - pictorial representations of, *II:437–438, II:438f*
 - squared errors in, *II:439–440*
 - use of, *I:165, I:172n, II:353*
 - vs.* Theil-Sen estimates of beta, *II:442f*
 - vs.* Theil-Sen regression, *II:441t*
- Ornstein-Uhlenbeck process
 - with change of time, *III:523*
 - and mean reversion, *I:263, I:264f*
 - solutions to, *III:492*
 - use of, *I:89, I:95*
 - and volatility, *III:656*
- Outcomes, identification and
 - evaluation of worst-case, *III:379–380*
- Outliers
 - in data sets, *II:200*
 - detection and management of, *II:206*
 - effect of, *II:355f, II:442–443*
 - and market crashes, *II:503*
 - in OLS methods, *II:354*
 - in quantile methods, *II:355–356*
 - and the Thiel-Sen regression algorithm, *II:440*
- Out-of-sample methodology, *II:238*
- Pair trading, *II:710*
- P-almost surely (P-a.s.) occurring events, *III:158*
- Parallel yield curve shift assumption, *III:12–13*
- Parameters
 - calibration of, *II:693*
 - density functions for values, *III:229f, III:230f, III:231f*
 - distributions of, *II:721*
 - estimation of for random walk, *I:83*
 - robust estimation of, *II:77–78*
 - stable, *III:246f*
- Parametric methods, use of, *II:522*
- Parametric models, *II:522–523, II:526–527*
- Par asset swap spreads, *I:530, I:531*
- Par CDS spread, *I:531*
- Par-coupon curve, *III:561*
- Pareto, Vilfredo, *II:467, II:468–469, II:474*
- Pareto(2) distribution, *II:441*
- Pareto distributions
 - density function of, *II:738*
 - generalized (GPD), *II:745–746, II:747, III:230–231*
 - in loss distributions, *III:108–109*
 - parameters for determining, *II:738*
 - stable, *II:738–741*
 - stable/varying density, *II:739f*
 - tails of, *II:751*
- Pareto law, *II:469*
- Pareto-Lévy stable distribution, *III:242*
- Partial differential equations (PDEs)
 - for American options, *II:664–665*
 - equations for option pricing, *II:660–665*
 - framework for, *I:261, I:265, II:675, III:555*
 - pricing European options with, *II:665–674*
 - usefulness of, *II:659–660*
 - use of, *III:18–19*
- Partitioning, binary recursive, *II:376–377, II:376f*
- Paths
 - in Brownian motion, *III:501, III:502f*
 - dependence, *III:18–19*
 - stochastic, *II:297*
- Payments, *I:229, II:611–612*
- Payment shock, *III:72*
- Payoff-rate process, *I:121–122*
- Payoffs, *III:466, III:638–639*
- PCA (principal components analysis). *See* principal component analysis (PCA)
- Pearson skewness, *III:204–205*
- Pension funds, constraints of, *II:62*
- Pension plans, *II:541, III:132*
- P/E (price/earnings) ratio, *II:20–21, II:38*
- Percentage rates, annual *vs.* effective, *II:615–617*
- Percolation models, *III:276*
- Performance attribution, *II:57, II:58, II:104, II:188–189, II:252–253, II:253t*
- Performance-seeking portfolios (PSPs), *I:36, I:37*
- Perpetuities, *II:607–608*
- Pharmaceutical companies, *II:7–8, II:11, II:244*

- Phillips-Perron statistic, *II:386, II:398*
- Pickand-Balkema-de Haan theorem, *II:746*
- Pickand estimator, *III:273*
- Pliska, Stankey, *II:476*
- Plot function, in MATLAB, *III:428–432*
- P-null sets, *III:197*
- Pochhammer symbol, *III:256*
- Poincaré, Henri, *II:469*
- POINT[®]
- features of, *II:193n, II:291n*
 - modeling with, *II:182*
 - screen shot of, *II:287f, II:288f*
 - use of, *II:179, II:189, II:286–287*
- Point processes, *III:270–272*
- Poisson-Merton jump process, distribution tails for, *III:540–541*
- Poisson-Merton jump variable, *III:540*
- Poisson processes
- compounded, *III:497*
 - homogeneous, *III:270–271*
 - and jumps, *I:93, III:498, III:540*
 - for modeling durations, *II:461*
 - as stochastic process, *III:496, III:497, III:506*
 - use of, *I:262, I:315–316*
- Poisson variables, distribution of, *III:271f*
- Policy iteration algorithm (Howard algorithm), *II:676–677*
- Polyhedral sets, *I:33, I:33f*
- Polynomial fitting of trend stationary process, *II:702–703, II:702f*
- Population profiles, in transition matrices, *III:32–34*
- Portfolio allocation, example using MATLAB, *III:436–441*
- Portfolio management
- approaches to, *II:108–110*
 - checklist for robust, *III:144*
 - for credit risk, *I:416–417*
 - of large portfolios, *III:325*
 - and mean-variance framework, *I:196*
 - real world, *I:190*
 - software for, *II:75 (See also Excel)*
 - tax-aware, *II:74–75*
 - using Bayesian techniques, *I:196*
- Portfolio managers, *III:444–445*
- approaches used by, *II:108–109*
 - enhanced indexers, *II:268*
 - example of, *III:436–441, III:437t*
 - questions considered by, *II:277*
 - specialization of, *II:48–49*
 - traditional vs. quantitative, *II:109, II:110t*
 - types of, *II:179, II:286*
- Portfolio optimization
- for American options, *II:678*
 - classical mean-variance problem, *III:441–444*
 - constraints on, *II:62*
 - defined, *I:36*
 - formulation of theory, *II:476*
 - max-min problem, *III:139*
 - models of, *II:84–85n*
 - robust, *III:146*
 - techniques of, *II:115–116*
 - uncertainty in, *I:192–193, II:82–83*
- Portfolios. *See* constraints, portfolio allocation of, *I:192–193, II:72*
- assessment of risk factors of, *III:637–638*
- benchmark, *I:41–42, II:180*
- building efficient, *II:115*
- bullet vs. barbell, *III:308t, III:309t*
- bullet vs. barbell (hypothetical), *III:308*
- cap-weighted, *I:38f*
- centering optimal, *I:199*
- considerations for rebalancing of, *II:75*
- construction of, *I:37–38, II:56–57, II:102–104, II:102f, II:114–116, II:179–184, II:261–264, II:286–287, II:301–303*
- cor-plus, and DTS, *I:398*
- credit bond, hedging of, *I:405*
- data on, *II:365t*
- diversification of, *I:10–12*
- efficient, *I:12, I:77, I:288f, I:289f, I:290f*
- efficient set of, *I:13*
- efficient vs. feasible, *I:13*
- efficient vs. optimal, *I:5*
- examples of, *II:261t, II:262t*
- expected returns from, *I:6–7, I:7, I:12t, I:69t, I:195*
- factor exposures in, *II:183t, II:184t, II:263t, II:264t*
- factor model approach to, *II:224*
- feasible and efficient, *I:12–14*
- feasible set of, *I:12–13, I:13f*
- index-tracking, *II:186*
- information content of, *I:192*
- long-short, *II:181–182, II:226f*
- management of fixed-income, *I:391*
- and market completeness, *I:50–52*
- mean-variance efficient, *I:66, I:69f*
- mean-variance optimization of, *II:79*
- momentum, *II:182f*
- monitoring of, *II:106*
- MSR (maximum Sharpe Ratio), *I:36–37*
- normalized, *II:157*
- optimal, *I:14–15, I:14f, I:15–17, II:181t*
- optimization-based approach to, *II:224–225*
- optimization of, *I:17–18, I:40, II:56–57, II:301–303*
- optimized, *II:116*
- performance-seeking, *I:36*
- quadratic approximation for value, *III:644–645*
- rebalancing of, *II:287–288*
- replication of, *II:476*
- resampling of, *I:189, II:78–80, II:84*
- returns of, *I:6–7*
- risk control in, *II:181–182*
- riskless, *I:509*
- with risky assets, *I:12–17*
- robust optimization of, *II:80–84*
- rule-based, *II:116*
- selection of, *I:3–19, III:351–353, III:356*
- self-financing, *II:660–661*
- stress tests for, *I:412*
- tangency, *I:36–37*
- tilting of, *II:263–264*
- tracking, *II:187t*
- weighting in, *I:50–51, II:64–65*
- weights of, *I:191–192*
- yield simulations of, *I:284–285*
- Portfolio sorts
- based on EBITDA/EV factor, *II:216–217, II:216f*
 - based on revisions factor, *II:217–218, II:217f*
 - based on share repurchase factor, *II:218, II:218f*
 - information ratios for, *II:219*
 - results from, *II:225f*
 - use of, *II:214–219*
- Portfolio trades, arbitrage, *I:440t*
- Position distribution and likelihood function, *I:142–143*
- Positive homogeneity property, *III:327–328*
- Posterior distribution, *I:159, I:165*
- Posterior odds ratio, *I:157*
- Posterior tradeoff, and normal mean, *I:158–159*
- Power conditional value at risk measure, *III:356*
- Power law, *III:234–235*
- Power plants/refineries, valuation and hedging of, *I:563*
- Power sets, *III:156, III:156t*
- Precision, *I:158, II:702*
- Predictability, *II:122–127*
- Predictions, *I:167, II:124*
- Predictive return modes, adoption of, *II:128–129*

- Preferred habitat hypothesis, III:569–570
- Prepayments
burnout, III:19
calculating speeds of, III:50–56
in cash-flow yields, III:4
conditional rate of (CPR), III:30, III:50–51, III:58–59
defaults and involuntary, III:59, III:74–77
defined, III:50
disincentives for, III:7–8
drivers of, III:77
effect of time on rates of, III:73–74
evaluation of, III:62
factors influencing speeds of, III:69–74
fundamentals of, III:66–69
for home equity securities, III:55–56
interactions with defaults, III:76–77
interest rate path dependency of, III:6
lag in, III:24–25
levels of analysis, III:50
lock-ins, III:73
modeling of, I:258, I:267, I:268, III:63n, III:598–600
practical interpretations of, III:20
rates of, III:74
reasons for, III:48
risk of, II:281, II:281t
S-curves for, III:67–68, III:67f
sources of, III:23–24
voluntary, III:38
voluntary *vs.* involuntary, III:30, III:75–76
- Prepay modeling, III:19–20
rational exercise, III:25
- Present value, I:268n, II:19, II:603–604, II:609, III:9–10
- Price/earnings (P/E) ratio, II:20–21, II:38
- Price patterns, scaling in, III:279
- Price processes, bonds, I:128
- Prices
bid/ask, III:625
Black-Scholes, II:673–674
changes in, II:722f, II:723f, II:742, III:305–306, III:305t
compression of, III:303
computing clean, I:214–215
dirty, I:382
distribution of, I:510
estimating changes in bond, I:373–374
flexible and sticky in CPI basket, I:292
formula for discounted, I:110
marked-to-market, I:430
modeling realistic, I:93–94
natural logarithm of, I:85
path-dependent, III:193n
strike, I:484–485, I:486
truncation of, III:304
vs. value, I:455n
- Price time series, autocorrelation in, III:274
- Pricing
backward induction, III:18
formulas for relationships, I:105–110
grids for, III:18–19
linear, I:52–55
models for, II:127–128
rational, I:53
risk-neutral, I:533, I:544
rule representation, I:260–261
use of trees, III:22–23
- Principal component analysis (PCA)
compared to factor analysis, II:166–168
concept of, II:157
defined, II:147, II:276
discussed, II:157–164
illustration of, II:158–163
with stable distributions, II:163–164
usefulness of, II:158
use of, I:39–40, II:142, II:168–169
- Principal components, defined, II:148, II:159
- Principal components analysis (PCA), I:556
- Prior elicitation, informative, I:152–153, I:159
- Prior precision, I:158
- Priors, I:153, I:165–167, I:168, I:171–172
- Probabilistic decision theory, II:719–721, II:729
- Probabilities
in Bayesian framework, I:140, I:144, I:146–148
conditional, I:117, II:517–518, III:477
formulas for conditional, I:108t
interpretation of, II:123
in models, II:299
posterior, I:140, I:144
prior, I:140, I:144
prior beliefs about, I:147
realistic, III:596–597
as relative frequencies, III:152
risk-adjusted, I:264
risk neutral, I:58–59, I:59, I:102, I:104, I:111–114, I:115–116, I:117, III:594–596
- Probability density function (PDF), III:384–385
- Probability distributions
binomial, III:186t
continuous, III:578
for drawing black balls, III:176–177
inverting the cumulating, III:646
for prepayment models, III:598
for rate of return, I:7t, I:9t
use of, III:638, III:645–646
- Probability-integral transformation (PIT), III:365
- Probability law, III:161
- Probability measures, III:157–159, III:594–597
- Probability of default (PD). *See* default probabilities
- Probability theory, II:133, II:700–701
- Probit regression models, II:348–349, II:350
- Processes
absolute volatility of, III:474
exponential, III:498
martingale, I:119, I:262–263, III:509, III:517
non-decreasing, III:503–505
normal tempered stable, III:504–505
predictable, II:132–133
subordinated, III:387–388
weakly stationary, II:360–361
- Process maps, III:94
- Proctor & Gamble, cash flows of, II:567–568, II:568t, II:571–573, II:573t
- Product transitions, III:66, III:71–73
- Profit, riskless, I:480
- Profitability ratios, II:555–557, II:563
- Profit margin ratios, II:555–556
- Profit opportunities, I:261
- Programming, linear, I:29, I:32–33
- Programming, stochastic
defined, III:123–124
in finance, III:125–126
general multistage model for
financial planning, III:128–132
use of scenario trees in, III:131–132
vs. continuous-time models, III:127–128
vs. other methods in finance, III:126–128
- Projected successive over relaxation (PSOR) method, II:677
- Projections, as-was, usefulness of, II:38
- Propagation effect, III:351
- Prospectus prepayment curve (PPC), III:54–55, III:56
- Protection, buying/selling of, I:230–231
- 100 PSA (Public Securities Association prepayment benchmark), III:51–52, III:55
- Pseudo-random numbers, generation of, III:647

- PSPs (performance-seeking portfolios), *I:36, I:37*
- Public Securities Association (PSA) prepayment benchmark, *III:51–55, III:51f, III:62–63*
- Pull to par value, *I:216*
- Pure returns, *II:51*
- Put-call parity, *I:437*
for American-style options, *I:446–448, I:452–453, I:452t*
for European options, *I:499*
for European-style options, *I:444–446, I:445t, I:451, I:451t*
perfect substitutes in European-style, *I:445t*
relations of, *I:446*
- Put-call parity relationship, *I:445, I:446, I:485*
- Put options, *I:439*
- Puts, American-style
early exercise of, *I:444, I:450–451*
error on value of, *II:677t, II:678t*
lower price bound, *I:443–444, I:450*
numerical results for, *II:677–678*
- Puts, European-style
arbitrage trades, *I:443t*
lower price bound, *I:443, I:450*
- Pyrrho's lemma, *II:330, II:331*
- Q-statistic of squared residuals, *II:422*
- Quadratic objective, two-dimensional, *I:29f*
- Quadratic programming, *I:29, I:33–34*
- Quadratic variation, *III:474*
- Quantiles
development of regression, *II:356*
methods, *II:354–356*
plot (QQ-plot) of, *III:272*
use of regression, *II:353–354, II:356–357*
- Quantitative methods, *II:483*
- Quantitative portfolio allocation, use of, *I:17–18*
- Quantitative strategies, backtesting of, *I:201*
- Quintile returns, *II:97–98*
- Quotes
delayed, *II:454*
discrepancies in, *II:453–454*
histograms from simple returns, *II:458f*
methods for sampling, *II:457–460*
mid-quote closing, *II:460f*
mid-quote format, *II:456*
mid-quote time-interpolated, *II:460f*
quantile plots of, *II:459f, II:461f*
- R^2 , adjusted, *II:315–316*
- Radon-Nikodym derivative, *I:111, I:130, I:133–134, III:510–511, III:515*
- Ramp, loans on, *III:52*
- Randomized operational time, *III:521*
- Randomness, *I:164, III:534–537, III:580*
- Random numbers
clusters in, *III:649–650*
generation of, *III:645–647*
practicality of, *III:647*
reproducing series of, *III:646*
simulations of, *III:650f*
- Random walks
advanced models of, *I:92–94*
arithmetic, *I:82–84, I:97, II:125*
for Brownian motion, *III:478–479*
computation of, *I:83, I:85, I:87, I:90*
correlated, *I:92–93, II:502–503*
defined, *III:486*
in forecastability, *II:127*
generation of, *I:85*
geometric, *I:84–88, I:89, I:97*
and linear nonstationary models, *II:508*
multivariate, *I:93*
parameters of, *I:87–88*
polynomial fitting of, *II:704f*
simulation of, *I:87*
and standard deviation, *II:385*
500-step samples, *II:708f*
strict, *II:126*
use of, *II:132, III:474*
variables in, *I:83–84*
- Range notes, valuing, *I:252*
- RAS Asset Management, *III:624*
- Rate-and-term refinancing, *III:66*
- Rating agencies, *I:300, III:44*
effect of actions of, *I:367–369*
role of, *I:362*
- Rating migration, *I:362, I:367–369*
- Rating outlooks, *I:365–366*
- Ratings
maturity of, *I:301*
- Ratings-based step-ups, *I:352*
- Rating transitions, *I:368, I:368t, I:381*
- Ratios
analysis of, *II:575–576*
classification of, *II:545–546*
defined, *II:545*
quick (acid test), *II:554*
scales of, *II:487*
- Real estate prices, effect of, *III:44*
- Real yield duration, calculation of, *I:286*
- Receipts, depository, *II:36*
- Recoveries, in foreclosures, *III:75*
- Recovery percentages, *III:30–31*
- Recovery rates
calibration of assumption, *I:537–538*
for captive finance companies, *I:366–367*
and credit risk, *I:362*
dealing with, *I:334n*
on defaulted securities, *I:367t*
drivers of, *I:372*
modeling of, *I:316–317*
random, *I:383*
relationship to default process, *I:372, I:376*
time dimension to, *I:366–377*
- Rectangular distribution, *III:219–221*
- Recursive out-of-sample test, *II:236*
- Recursive valuation process, *I:244*
- Reduced form models, usefulness of, *I:412*
- Redundant assets/securities, *I:51*
- Reference entities, *I:526*
- Reference priors, *I:159–160n*
- Refinancing
and ARMs, *III:72*
categories of, *III:48*
discussion of, *III:68–69*
rate-and-term, *III:68*
speed of, *III:25–26*
threshold model, *III:18*
- Refinancing, paths of rates, *III:8t*
- RefiSMM(Price) function, *III:25–26*
- Regime switching, *I:173n*
- Regression
binary, *III:364*
properties of, *II:309–310*
spurious, *II:384, II:385*
stepwise, *II:331*
- Regression analysis
results for dummy variable regression, *II:348t*
usefulness of, *II:305*
use in finance, *II:316–328*
variables in, *II:330*
- Regression coefficients, testing of, *I:170*
- Regression disturbances, *I:164*
- Regression equations, *II:309–310*
- Regression function, *II:309*
- Regression models, *I:168–169, I:170–172, II:302*
- Regressions
estimation of linear, *II:311–314*
explanatory power of, *II:315–316*
linear, *II:310–311*
and linear models, *II:308–311*
pitfalls of, *II:329–330*
sampling distributions of, *II:314*
spurious, *II:329*
- Regression theory, classical, *II:237*
- Regressors, *II:308–310, II:311, II:330*

- Reg T (Treasury Regulation T), *I:67*
- Relative valuation analysis
 hypothetical example of, *II:40–45*
 hypothetical results, *II:40t*
 implications of hypothetical,
II:41–42
 low or negative numbers in, *II:42–43*
- Relative valuation methods
 choice of valuation multiples in,
II:38–39
 usefulness of, *II:45*
 use of, *II:33–34, II:45*
- Replication, *I:526*
- Reports, *II:200–201, II:283–286*
- Research, process of quantitative,
II:717f
- Residuals, *II:220, II:328–329*
- Restructuring, *I:528–530, I:529, I:529t, I:530, I:537*
- Return covariance matrix formula,
II:141
- Return distributions, *III:333f, III:388–392*
- Return effects, *II:47–48, II:51, II:51f*
- Return generating function, *II:256*
- Return on assets, *II:547–548, II:548–550*
- Return on equity (ROE), *II:37–38, II:41–42, II:548, II:550*
- Return on investment ratios,
II:547–551, II:548, II:563
- Returns
 active, *II:115*
 arithmetic *vs.* geometric average,
II:598
 defined, *II:598*
 estimated moments of, *II:204*
 estimates of expected, *I:190–191*
 ex ante, *I:7*
 excess, *I:66, I:67, I:74*
 expected, *I:71–72, II:13–14, II:112*
 ex post, *I:6*
 fat tails of conditional distribution,
II:753n
 finite variance of, *III:383–384*
 forecasting of, *II:111–112, II:362*
 historical, *II:285f, III:389t*
 monthly *vs.* size-related variables,
II:52t
 naïve, *II:51, II:53f*
 naïve *vs.* pure, *II:52f, II:53–54*
 Nasdaq, Dow Jones, bond, *II:365f*
 pure, *II:51, II:53f, II:54t*
 robust estimators for, *I:40–41*
 rolling 24-month, *II:229f*
 systematic *vs.* idiosyncratic, *II:173*
 time-series properties of, *II:733–734*
- Returns to factors, *II:248*
- Return to maturity expectations
 hypothesis, *III:569*
- Return volatility, excess and DTS,
I:396–397
- Reverse optimization, *I:203n*
- Riemann-Lebesgue integrals, *III:483*
- Riemann-Stieltjes integrals, *I:122, III:473–474, III:487*
- Riemann sum, *II:743–744*
- Risk. *See also* operational risk
 alternative definitions of, *III:350*
 analyzing with multifactor models,
II:184–188
 assessment of, *III:640–641*
 asymmetry of, *III:350–351*
 budgeting of, *II:115, II:286–287*
 of CAPM investors, *I:73–74*
 changes in, *II:368, III:351*
 coherent measures of, *III:327–329*
 collective, *II:470*
 common factor/specific, *II:258*
 controlling, *I:397*
 correlated, *II:271t*
 correlated *vs.* isolated, *II:271*
 counterparty, *I:478, I:479*
 decomposition of, *II:250–253, II:257–261, II:265*
 and descriptors, *II:140*
 downside, *III:382*
 effect of correlation of asset returns
 on portfolio, *I:11–12*
 effect of number of stocks on,
II:249f
 estimation of, *I:40*
 in financial assets, *I:369*
 forecasting of, *II:112–113*
 fundamental, *II:199*
 funding, *II:199*
 horizon, *II:199*
 idiosyncratic, *II:178, II:188, II:188t, II:283, II:285t, II:291*
 idiosyncratic *vs.* systematic, *I:40–41*
 implementation, *II:199*
 including spread in estimation of,
I:399
 indexes of, *II:140, II:256*
 interest rate, *I:521–522, III:4*
 issue specific, *II:283t*
 liquidity, *II:199*
 main sources of, *II:211*
 market price of, *III:579, III:588, III:591*
 model (*See* model risk)
 modeling, *III:11*
 momentum, *II:181t*
 as multidimensional phenomenon,
III:350
 noise trader, *II:199*
 perspective on, *II:91–92*
 portfolio, *I:7–10, I:9–10, I:11, II:180t*
 repayment, *III:48*
 price movement costs, *II:69*
 quantification of, *I:4, I:7–8*
 realized, *II:118*
 reinvestment, *III:4–5*
 relativity of, *III:350*
 residual, *II:258–259*
 by sector, *II:185t*
 in securities, *I:73*
 sources of, *II:173–174, II:251f, II:274, II:281–282*
 systematic, *II:186*
 tail, *I:384, I:385*
 true *vs.* uncertainty, *II:721*
 in a two-asset portfolio, *I:8*
 in wind farm investments, *I:563–564*
- Risk analysis, *II:268–286, II:273t, II:274t, II:275t*
- Risk aversion, *I:404*
 in analysis, *III:570*
 coefficient for, *I:59*
 functions, *III:339f*
 of investors, *I:191*
 and portfolio management, *I:37*
- Risk-based pricing, *III:70*
- Risk decomposition
 active, *II:259–260, II:259f*
 active systematic-active residual,
II:260, II:260f
 insights of, *II:252*
 overview of, *II:261f*
 summary of, *II:260–261*
 systematic-residual, *II:258–259, II:259f*
 total risk, *II:258, II:258f*
- Risk exposures, *I:394, I:521*
- Risk factors
 allocation of, *I:398*
 constraints on, *II:63–64*
 identification of, *II:256*
 macroeconomic, *I:415–416*
 missing, *II:693*
 systematic, *II:268, II:474*
 unsystematic, *II:474*
- Riskiness, determining, *I:145*
- Risk management
 internal models of, *III:289–290*
 in investment process, *II:104*
 portfolio, *III:643–644*
 in portfolio construction, *II:303*
 and quasi-convex functions, *I:28*
- Risk measures, safety-first, *III:352, III:354–356, III:357*
- RiskMetrics™ Group
 approach of, *III:322–323*
 comparison with FTSE100 volatility,
III:413f
 methodology of, *III:412–413*
 software of, *III:413*
 website of, *III:412*

- Risk models
 applications of, *II:286–290*
 comparisons among, *II:747–751*
 defined, *II:692*
 equity, *II:172–173, II:192–193, II:255, II:264*
 indicator, *III:93–94*
 and market volatility, *II:748*
 multifactor, *II:257–258*
 principal of, *II:292n*
 and uncertainty, *II:724*
 use of, *II:171–172, II:268, II:290*
- Risk neutral, use of term, *III:593–594*
- Risk neutral density (RND)
 concept of, *II:521*
 fitting data to models of, *II:526–527*
 generally, *II:527*
 parametric models for, *II:523–525*
- Risk oversight, *II:303*
- Risk premiums
 for default, *III:599*
 importance of, *III:587*
 quantifying, *III:580–581*
 of time value, *I:513*
 as a variable in discount bond prices, *III:581*
 variables, *I:403, I:405*
- Risk reports
 credit risk, *II:278–281*
 detailed, *II:272–286*
 factor exposure, *II:275–283*
 implied volatility, *II:282*
 inflation, *II:282*
 issue-level, *II:283–285*
 liquidity, *II:282*
 prepayment risk, *II:281*
 risk source interaction, *II:281–282*
 scenario analysis, *II:285–286*
 summary, *II:272–275*
 tax-policy, *II:282–283*
- Risk tolerance, *II:720–721, II:725, II:729f*
- Risky bonds, investment in, *II:726–729*
- Robot analogy, *III:594*
- Robust covariance matrix, *II:446*
- Robust optimization, *II:83, III:141–142*
- Robust portfolio optimization, *I:17–18, I:193, III:138–142*
 effect on performance, *III:144*
 need for research in, *III:145–146*
 practical considerations for, *III:144–145*
 in practice, *III:142–144*
- Rolling windows, use of, *II:371*
- Roots
 complex, *II:632–634, II:636–637*
 in homogenous difference equations, *II:642*
 real, *II:630–632, II:635–636*
- Ross, Stephen, *II:468, II:475*
- Rounding, impact of, *III:306n*
- Roy CAPM, *I:67, I:69, I:70*
- Ruin problem, development of, *II:470–471*
- Runge-Kutta method, *II:650–652, II:651f, II:652f*
- Russell 1000, *II:213, II:236–237*
- Saddle points, *I:23, I:23f, I:30*
- Sales, net credit, *II:557–558*
- Samples
 effect of size, *I:158–159, I:159f, III:407*
 importance of size, *III:152*
 and model complexity, *II:703–707*
 in probability, *III:153*
 selection of, *II:716*
- Sampling
 antithetic, *I:383*
 importance, *I:384, III:648–649*
 stratified, *II:115, III:648*
- Sampling error, *III:396*
- Samuelson, Paul, *I:556, II:468, II:473–474*
- Sandmann-Sondermann model, *I:493*
- Sarbanes-Oxley Act (2002), *II:542*
- Scalar products, *II:625–626*
- Scale parameters, *I:160n*
- Scaling laws, use of, *III:280*
- Scaling vs. self-similarity, *III:278–280*
- Scenario analysis
 constraints on, *III:130*
 factor-based, *II:189–192, II:193*
 for operational risk, *III:93*
 usefulness of, *II:179*
 use of, *II:288–290, III:378*
- Scenarios
 defined, *III:128*
 defining, *II:189*
 generation of, *III:128–132*
 network representation of, *III:129f*
 number needed of, *III:640–641*
- Scholes, Myron, *II:468, II:476*
- Schönbucher-Schubert (SS) approach, *I:329–331*
- Schwarz criterion, *II:387, II:389*
- Scorecard Approach, *III:100n*
- Scott model, *II:681–682*
- SDMs (state dependent models), *I:342, I:351–352*
- Secrecy, in economics, *II:716*
- Sector views, implementation of, *II:182–184*
- Securities
 alteration of cash flows of, *I:210*
 arbitrage-free value of, *I:261*
 baskets of, *I:483–484*
 convertible, *I:462*
 creating weights for, *II:102–104, II:103f*
 evaluation of, *I:50*
 fixed income, *I:209–210, II:268*
 formula for prices, *I:107*
 non-Treasury, *I:222–223, I:223t*
 of other countries, *I:226*
 payoffs of, *I:49–50, I:116–117, I:121–122*
 pricing European-style, *III:642*
 primary, *I:458*
 primitive, *I:51*
 private label (*See* MBS (mortgage-backed securities), nonagency)
 ranking of, *I:200–201*
 redundant, *I:124*
 risk-free, *I:115*
 selection of, *I:225–226*
 structured, *I:564, I:565–566*
 supply and demand schedule of, *III:626f*
 valuing credit-risky, *III:645*
 variables on losses, *I:370*
- Securities and Exchange Commission (SEC)
 filings with, *II:532*
- Security levels, two-bond portfolio, *I:382t*
- Selection, adverse vs. favorable, *III:76–77*
- Self-exciting TAR (SETAR) model, *II:405*
- Self-similarity, *III:278–280*
- Selling price, expected future, *II:19–20*
- Semimartingales, settings in change of time, *III:520–521*
- Semi-parametric models
 tail in, *II:744–747*
- Semiparametric/nonparametric methods, use of, *II:522*
- Semivariance, as alternative to variance, *III:352*
- Sensitivity, *III:643–644*
- Sensitivity analysis, *I:192, II:235*
- Sequences, *I:378, III:649–651, III:650*
- Series, *II:299, II:386, II:507–508, II:512*
- SETAR model, *II:425–426*
- Set of feasible points, *I:28, I:31*
- Set operations, defined, *III:153–154*
- Sets, *III:154*
- Settlement date, *I:478*
- Settlements, *I:526–528*
- Shareholders
 common, *II:4*
 equity of, *II:535*
 negative equity of, *II:42*
 preferred, *II:4–5*
 statement of equity, *II:541*

- Shares, repurchases of, *II:207, II:210f, II:211, II:215–216, II:227*
- Sharpe, William, *I:75, II:468, II:474*
- Sharpe-Lintner CAPM (SL-CAPM), *I:66–67, I:75, I:78n*
- Sharpe ratios, *I:40, I:62, I:193*
- Sharpe's single-index model, *I:74–75*
- Shipping options, pricing of, *I:565*
- Shortfall, expected, *I:385–386*
- Short positions, *I:67*
- Short rate models, *III:543–545, III:545–550, III:552–554, III:557, III:604–610*
- Short rates, *III:212–213, III:541, III:549, III:595–596*
- Short selling
 constraints on, *I:67*
 effect of constraints on, *I:17, I:191–192, II:461*
 effect of on efficient frontiers, *I:17f*
 example, *I:480–481*
 as hedging route, *I:409*
 in inefficient markets, *I:71f*
 and market efficiency, *I:70–71*
 net portfolio value, *I:433t*
 and OAS, *I:259*
 and real estate, *II:396–397*
 in reverse cash-and-carry trade, *I:483*
 for terminal wealth positions, *I:460–461*
 using futures, *I:432–433*
- Shrinkage
 estimation of, *I:192, I:194–195, I:201–202, III:142*
 optimal intensity of, *I:202n–203n*
 use of estimators, *II:78*
- δ -algebra, *III:15, III:157*
- δ -fields
 defined, *III:508*
- Signals (forecasting variables), use of
 in forecasting returns, *II:111–112*
 evaluation of, *II:111–112*
- Similarity, selecting criteria for, *II:35*
- Simulated average life, *III:12*
- Simulations
 credit loss, *I:378–380*
 defined, *III:637*
 efficiency of, *I:384*
 financial applications of, *III:642–645*
 process of, *III:638*
 technique of, *III:444–445*
- Single firm models, *I:343–352*
- Single monthly mortality rate (SMM), *III:50–51, III:58*
- Skewness
 defined, *III:238–239*
 and density function, *III:204–205*
 indicating, *III:235*
 and the Student's *t*-distribution, *III:387*
 treatment of stocks with, *I:41*
- Sklar's theorem, *I:326, III:288*
- Skorokhod embedding problem, *III:504*
- Slackness conditions, complementary, *I:32*
- SL-CAPM (Sharpe-Lintner CAPM), *I:66–67, I:75, I:78n*
- Slope elasticity measure, *III:315, III:317*
- Smith, Adam, *II:468, II:472*
- Smoothing, in nonparametric methods, *II:411–412*
- Smoothing constant, *III:409–410*
- Smoothly truncated stable distribution (STS distribution), *III:245–246*
- Smooth transition AR (STAR) model, *II:408–409*
- Sobol sequences, pricing European call options with, *III:445–446*
- Software
 case sensitivity of, *III:434*
 comments in MATLAB code, *III:427*
 developments in, *II:481–482*
 macros in, *III:450–452, III:450f, III:460, III:466*
 pseudo-random number generation, *III:646–647*
 random number generation commands, *III:645–647*
- RiskMetrics Group, *III:413, III:644*
- simulation, *III:651f*
 for stable distributions, *III:344, III:383*
 stochastic programming applications, *III:126*
 use of third party, *II:481*
- Solutions, stability of, *II:652–653*
- Solvers, in MATLAB, *III:435*
- Space in probability, *III:156, III:157*
- Sparse tensor product, *II:673*
- S&P 60 Canada index, *I:550–552, I:550t, I:553f*
- Spearman, Charles, *II:153–154*
- Spearman model, *II:153–154*
- Spearman's rho, *I:327, I:332, I:336n*
- Splits, in recursive partitioning, *II:376–377*
- Spot curves, with key rate shifts, *III:313f, III:314f*
- Spot price models, energy commodities, *I:556–557*
- Spot rates
 arbitrage-free evolution of, *I:557–558*
 bootstrapping of curve, *I:217–220*
 calculation of, *III:581*
 and cash flows in OAS analysis, *I:259*
 changes in, *III:311, III:312f, III:312t*
 computing, *I:219–220*
 under continuous compounding, *III:571*
 defined, *III:595*
 effect of changes in, *I:514, III:313–314, III:314t*
 and forward rates, *III:572*
 models of, *III:579–581*
 paths of monthly, *III:9–10, III:10t*
 theoretical, *I:217*
 Treasury, *I:217*
 uses for, *I:222*
- Spot yields, *III:565, III:566, III:571*
- Spread analysis, *II:290t*
 table of, *II:290t*
- Spread duration, beta-adjusted, *I:394*
- Spreads
 absolute and relative change volatility, *I:396f*
 change in, *I:392, I:393, I:394f, I:399*
 determining for asset swaps, *I:227–228*
 level vs. volatility of, *I:397*
 measurement of, *II:336–337*
 measure of exposure to change in, *I:397*
 nominal, use of, *III:5*
 option-adjusted, *I:253–255, I:254f*
 reasons for, *I:210–211*
 relative vs. absolute modeling, *I:393*
 volatility vs. level, *I:394–396, I:395f*
 zero-volatility, *III:5*
- Squared Gaussian (SqG) model, *III:547–548*
- Square-root rule, *III:534*
- SR-SARV model class, *II:370*
- St. Petersburg paradox, *III:480*
- Stability
 notion of, *II:667*
 in Paretian distribution, *II:739–741*
 property of, *II:740–741, III:236–237, III:244–245*
- Stable density functions, *III:236f*
- Stable Paretian model, α -stable distribution in, *II:748*
- Standard Default Assumption (SDA) convention, *III:59–60, III:60f*
- Standard deviations
 and covariance, *I:9*
 defined, *III:168*
 mean, *III:353*
 posterior, *I:155*
 related to variance, *III:203–204*
 rolling, *II:362–363*

- Standard deviations (*Continued*)
 and scale of possible outcomes,
III:168f
 for tail, *III:341*
- Standard errors. *See also* errors
 for average estimators, *III:400–402*
 defined, *III:399*
 estimation of, *III:640*
 of the estimator, *III:400*
 for exponentially weighted moving
 averages (EWMA), *III:411–412*
 reduction of, *III:648*
- Standard normality, testing for,
III:366–367
- Standard North American contract
 (SNAC), *I:529*
- Standard & Poors 500
 auto correlation functions of, *II:389t*
 cointegration regression, *II:390t*
 daily close, *III:402f*
 daily returns (2003), *III:326f*
 distributions of, *III:384f*
 error correction model, *II:391t*
 historical distributions of, *III:390f*
 index and dividends (1962–2006),
II:388f
 parameter estimates of, *III:385t*,
III:387t, *III:388t*
 return and excess return data
 (2005), *II:316–317t*
 stationarity test for, *II:389t*
 time scaling of, *III:383f*
 worst returns for, *III:382t*
- State dependent models (SDMs), *I:342*,
I:351–352
- Statement of stockholders' equity,
II:541
- State price deflators
 defined, *I:103*, *I:129–130*
 determining, *I:118–119*, *I:124*
 formulas for, *I:107–108*, *I:109–110*
 in multiperiod settings, *I:105*
 and trading strategy, *I:106*
- State prices
 and arbitrage, *I:55–56*
 condition, *I:54*
 defined, *I:101–102*
 and equivalent martingale
 measures, *I:133–134*
 vectors, *I:53–55*, *I:58*, *I:119*
- States, probabilities of, *I:115*
- States of the world, *I:457–458*, *I:459*,
II:306, *II:308*, *II:720*
- State space, *I:269n*
- Static factor models, *II:150*
- Stationary series, trend *vs.* difference,
II:512–513
- Stationary univariate moving average,
II:506
- Statistical concepts, importance of,
II:126–127
- Statistical factors, *II:177*
- Statistical learning, *II:298*
- Statistical methodology, EWMA,
III:409
- Statistical tests, inconsistencies in,
II:335–336
- Statistics, *II:387*, *II:499*
- Stein paradox, *I:194*
- Stein-Stein model, *II:682*
- Step-up callable notes, valuing of,
I:251–252
- Stochastic, defined, *III:162*
- Stochastic control (SC), *III:124*
- Stochastic differential equations
 (SDEs)
 binomial/trinomial solutions to,
III:610–613
 with change of time methods,
III:523
 defined, *II:658*
 examples of, *III:523–524*
 generalization to several
 dimensions with, *III:490–491*
 intuition behind, *III:486–487*
 modeling states of the world with,
III:127
 for MRAM equation, *III:525–526*
 setting of change of time, *III:521*
 solution of, *III:491–493*
 steps to definition, *III:487*
 usefulness of, *III:493*
 use of, *II:295*, *III:485–486*,
III:489–490, *III:536*, *III:603*,
III:619
- Stochastic discount factor, *I:57–58*
- Stochastic integrals
 defined, *III:481–482*
 intuition behind, *III:473–475*
 in Ito processes, *III:487*
 properties of, *III:482–483*
 steps in defining, *III:474–475*
- Stochastic processes
 behavior of, *I:262*
 characteristic function of, *III:496*
 characteristics of, *II:360*
 continuous-time, *III:496*, *III:506*
 defined, *I:263–264*, *I:269n*, *II:518*,
III:476, *III:496*
 discrete time, *II:501*
 properties of, *II:515*
 representation of, *II:514–515*
 and scaling, *III:279*
 specification of, *II:692–693*
- Stochastic programs
 features of, *III:124*, *III:132*
- Stochastic time series, linear,
II:401–402
- Stochastic volatility models (SVMs)
 with change of time, *III:520*
 continuous-time, *III:656*
 discrete, *III:656–657*
 importance of, *III:658*
 for modeling derivatives,
III:655–656
 multifactor models for, *III:657–658*
 and subordinators, *III:521–522*
 use of, *III:653*, *III:656*
- Stock indexes
 interim cash flows in, *I:482*
 risk control against, *II:262–263*
- Stock markets
 bubbles in, *II:386*
 as complex system, *II:47–48*
 1987 crash, *II:521*, *III:585–586*
 dynamic relationships among,
II:393–396
 effects of crises, *III:233–234*
 variables effects on different sectors
 of, *II:55*
- Stock options, valuation of long-term,
I:449
- Stock price models
 binomial, *III:161*, *III:171–173*, *III:173f*
 multinomial, *III:180–182*, *III:181f*,
III:184
 probability distribution of
 two-period, *III:181t*
- Stock prices
 anomalies in, *II:111t*
 behavior of, *II:58*
 correlation of, *I:92–93*
 and dividends, *II:4–5*
 lognormal, *III:655–656*
 processes of, *I:125*
- Stock research, main areas of, *II:244t*
- Stock returns, *II:56*, *II:159f*
- Stocks
 batting average of, *II:99*, *II:99f*
 characteristics of, *II:204*
 common, *II:4*, *II:316–322*
 cross-sectional, *II:197*
 defined, *II:106*
 defining parameters of, *II:49*
 determinants of, *II:245f*
 execution price of, *III:626*
 fair value *vs.* expected return, *II:13f*
 finding value for XYZ, Inc., *II:31t*
 information coefficient of, *II:98f*
 information sources for, *II:90f*
 measures of consistency, *II:99–100*
 mispriced, *II:6–7*
 quantitative research metrics tests,
II:97–99
 quintile spread of, *II:97f*
 relative ranking of, *I:196–197*
 review of correlations, *II:101f*

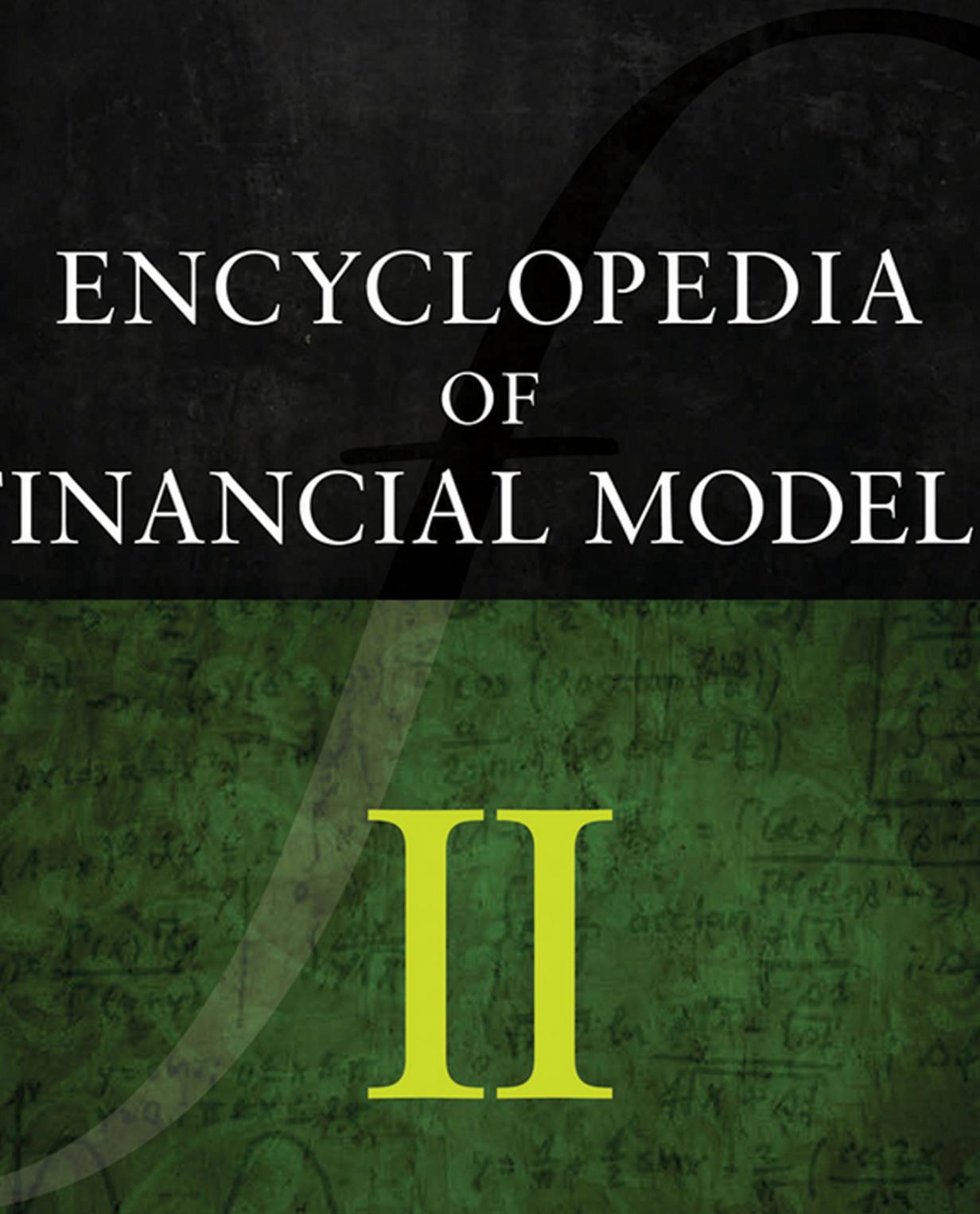
- sale/terminal price of, *II:5*
 short selling of, *I:432–433*
 similarities between, *II:245f*
 sorting of, *II:215*
 testing of, *II:95, II:96f*
 that pay no dividend, *II:17*
 use of, *II:90*
 valuation of, *II:6, II:8–9, II:14, II:18–19*
 weightings of, *II:101f*
- Stock selection**
 models for, *II:197*
 in quantitative equity investment process, *II:105*
 quantitative model, *II:94–95*
 for retail sector, *II:94f*
 strategies for, *II:195*
 tree for, *II:379–381, II:380f*
- Stopping times, *II:685***
Straontonovich, Ruslan, *II:470*
Strategies, backtesting of, *II:235–236*
Stress tests, *I:412, I:417, I:418, III:93, III:596–597*
Strike price, *I:509, I:514*
Strong Law of Large Numbers (SLLN), *I:270n, III:263–264*
Structural breaks, *I:167, III:274–275*
Student's *t* distribution
 applications to stock returns, *III:215–216*
 and AVaR, *III:334–335*
 classical, *II:734–738*
 density function of, *II:735*
 discussion of, *III:213–216*
 distribution function of, *III:215f*
 for downside risk estimation, *III:386–387*
 fitting and simulation of, *II:737–738*
 heavy tails of, *I:160n, I:176, II:747–748, II:751, III:227–228*
 limitations of, *II:736*
 in modeling credit risk, *I:387–388*
 normals representation in, *I:177–178*
 skewed, *II:736–737, II:753n*
 skewness of, *III:390*
 standard deviation of, *I:173n*
 symmetry of, *III:387*
 tails of, *III:392*
 use of, *I:153–154, I:172n, III:234*
- Student's *t*-test, *II:219***
Sturge's rule, *II:495*
Style analysis, *II:189*
Style factors, *II:247*
Style indexes, *II:48*
Stylized facts, *II:503–504*
Subadditivity property, *III:328*
Subordinated processes, *I:186n, III:277, III:521–522*
- Successive over relaxation (SOR) method, *II:677***
Summation stability property (Gaussian distribution), *II:732–733*
Supervisory Capital Assessment Program, *I:300, I:412*
Support, defined, *II:200*
Survey bias, *I:293*
Survival probability, *I:533–535*
Swap agreements, *I:434, I:435–436n*
Swap curves, *I:226, II:275–276*
Swap rates, *I:226, III:536f*
Swaps
 with change of time method, *III:522*
 covariance/correlation, *I:547–548, I:549–550, I:552*
 duration-matched, *I:285*
 freight rate, *I:558*
 modeling and pricing of, *I:548–550*
 summary of studies on, *I:546t*
 valuing of, *I:434–435*
- Swap spread (SS) risk, *II:278, II:278t***
Swaptions, *I:502–503, III:550*
Synergies, in conglomerates, *II:43–44*
Systematic risk, *II:290*
Systems
 homogenous, *II:624*
 linear, *II:624*
 types of, *II:47, II:58*
- Tailing the hedge, defined, *I:433***
Tail losses
 in loss functions, *III:369–370*
Tail probability, *III:320*
Tail risk, *I:377, I:385, II:752*
Tails
 across assets through time, *II:735–736*
 behavior of in operational losses, *III:111–112*
 in density functions, *III:203*
 dependence, *I:327–328, I:387*
 Gaussian, *III:98–99, III:260*
 heavy, *II:734–744, III:238*
 modeling heaviness of, *II:742–743*
 for normal and STS distributions, *III:246t*
 power tail decay property, *II:739, III:244*
 properties of, *III:261–262*
 tempering of, *II:741*
- Takeovers, probability of, *I:144–145***
Tangential contour lines, *I:29–30, I:30f, I:32f*
Tanker market, *I:565*
TAR-F test, *II:426*
TAR(1) series, simulated time plot of, *II:404f*
- Tatonnement, concept of, *II:468***
Taxes
 and bonds, *I:226*
 capital gains, *II:73*
 cash, *II:573*
 for cash/futures transactions, *I:484*
 complexity of, *II:73–74*
 deferred income, *II:535, II:538*
 effect on returns, *II:83–84, II:84, II:85n*
 in financial statements, *II:541*
 impact of, *I:286–287*
 incorporating expense of, *II:73–75*
 managing implications of, *III:146*
 and Treasury strips, *I:218*
- Tax policy risk, *II:282–283***
Technology, effect of on relative values, *II:37*
Telescoping futures strategy, *I:433*
Tempered stable distributions
 discussions of, *III:246–252, III:384–386*
 generalized (GTS), *III:249*
 Kim-Rachev (KRTS), *III:251–252*
 modified (MTS), *III:249–250*
 normal (NTS), *III:250–251*
 probability densities of, *III:247f, III:248f, III:250f, III:252f*
 rapidly decreasing (RDTS), *III:252*
 tempering function in, *III:254, III:258n*
- Tempered stable processes, *III:499–501, III:500t, III:512–517***
Tempering functions, *III:254, III:255t*
Templates, for data storage, *II:204*
Terminal profit, options and forwards, *I:438f, I:439f*
Terminal values, *II:45*
Terminology
 of delinquency, default and loss, *III:56*
 of prepayment, *III:49–50*
 standard, of tree models, *II:376*
- Term structure**
 in contiguous time, *III:572–573*
 continuous time models of, *III:570–571*
 defined, *III:560*
 eclectic theory of, *III:570*
 of forward rates, *III:586*
 mathematical relationships of, *III:562*
 modeling of, *I:490–494, III:560*
 of partial differential equations, *III:583–584*
 in real world, *III:568–570*
- Term structure modeling**
 applications of, *III:584–586*
 arbitrage-free, *III:594*

- Term structure modeling (*Continued*)
 calibration of, III:580–581
 discount function in, III:565
 discussion of, III:560–561
- Term structure models
 approaches to, III:603–604
 defined, I:262, I:263
 discrete time, III:562–563
 discussion of, III:561–562
 of interest rates, I:314
 internal consistency checks for, III:581
 with no mean reversion, III:613–616
 for OAS, I:265–267
 quantitative, III:563
 static *vs.* dynamic, III:561–562
- Term structures, III:567–568, III:570, III:579, III:587
- Tests
 Anderson-Darling (AD), III:112–113
 BDS statistic, II:423–424, II:427
 bispectral, II:422–423
 cointegration, II:708–710
 Kolmogorov-Smirnov (KS), III:112–113
 monotonic relation (MR), II:219
 nonlinearity, II:426–427, II:427^t
 nonparametric, II:422–424
 out-of-sample *vs.* in-sample, II:236
 parametric, II:424–426
 RESET, II:424–425
 run tests, III:364
 threshold, II:425–426
 for uniformity, III:366
- TEV (tracking error volatility), II:180, II:186, II:272–274, II:286–287
- Theil-Sen regression algorithm, II:440–442, II:443–446, II:444^t
- The Internal Measurement Approach (BIS), III:100n
- Theoretical value, determination of, III:10–11
- Théorie de la Spéculation (The Theory of Speculation)* (Bachelier), II:121–122, II:469
- Theory of point processes, II:470–471
- Three Mile Island power plant crisis, II:51–52
- Three-stage growth model, II:9–10
- Threshold autoregressive (TAR) models, II:404–408
- Thresholds, II:746–747
- Through the cycle, defined, I:302–303, I:309–310
- Thurstone, Louis Leon, II:154
- Tick data. *See* high-frequency data (HFD)
- Time
 in differential equations, II:643–644
 physical *vs.* intrinsic scales of, II:742
 use of for financial data, II:546–547
- Time aggregation, II:369
- Time decay, I:509, I:513, I:521^f
- Time dependency, capture of, II:362–363
- Time discretization, II:666, II:679
- Time increments
 models of, I:79
 in parameter estimation, I:83
- Time intervals, size of, II:300–301
- Time lags, II:299–300
- Time points, spacing of, II:501
- Time premiums, I:485
- Time series
 autocorrelation of, II:331
 causal, II:504
 concepts of, II:501–503
 continuity of, I:80
 defined, II:501–502, II:519
 fractal nature of, III:480
 importance of, II:360
 multivariate, II:502
 stationary, II:502
 stationary/nonstationary, II:299
 for stock prices, II:296
- Time to expiry, I:513
- Time value, I:513, I:513^f, II:595–596
- TIPS (Treasury inflation-protected securities)
 and after-tax inflation risk, I:287
 apparent real yield premium, I:293^f
 effect of inflation and flexible price CPI, I:292^f
 features of, I:277
 and flexible price CPI, I:291^f
 and inflation, I:290, I:294
 performance link with short-term inflation, I:291–292
 real yields on, I:278
 spread to nominal yield curve, I:281^f
 volatility of, I:288–290, I:294
vs. real yield, I:293–294
 10-year data, I:279–280
 yield of, I:284
 yields from, I:278
- TLF model, strengths of, III:388–389
- Total asset turnover ratio, II:558
- Total return reports, II:237^t
- Total return swaps, I:540–542, I:541–542
- Trace test statistic, II:392
- Tracking error
 actual *vs.* predicted, II:69
 alternate definitions of, II:67–68
 defined, II:115, II:119
 estimates of future, II:69
 as measure of consistency, II:99–100
 reduction of, II:262–263
 standard definition, II:67
 with TIPS, I:293
- Tracking error volatility (TEV). *See* TEV (tracking error volatility)
- Trade optimizers, role of, II:116–117
- Trades
 amount needed for market impact, III:624
 cash-and-carry, I:487
 crossing of, II:75
 importance of execution of, III:623, III:631
 measurement of size, III:628
 in portfolio construction, II:104, II:116–117
 round-trip time of, II:451
 size effects of, III:372, III:630
 speed of, II:105
 timing of, III:628–629
- Trading costs, II:118, III:627–628, III:631–632
- Trading gains, defined, I:122, I:123
- Trading horizons, extending, III:624
- Trading lists, II:289^t
- Trading strategies
 backtesting of, II:236–237
 categories of, II:195
 in continuous-state, continuous-time, I:122
 development of factor-based, II:197–198, II:211
 factor-based, II:195, II:232–235
 factor weights in, II:233^f
 in multiperiod settings, I:105
 risk to, II:198–200
 self-financing, I:126–127, I:136
- Trading venues, electronic, II:57
- Training windows, moving, II:713–714
- Tranches, III:38, III:39^t, III:45
- Transaction costs
 in backtesting, II:235
 in benchmarking, II:67
 components of, II:119
 consideration of, II:64, II:85–86n
 dimensions of, III:631
 effect of, I:483
 figuring, II:85n
 fixed, II:72–73
 forecasting of, II:113–114
 incorporation of, II:69–73, II:84
 international, III:629
 linear, II:70
 and liquidity, III:624–625
 managing, III:146
 measurement of, III:626
 piecewise-linear, II:70–72, II:71^f

- quadratic, *II:72*
 in risk modeling, *II:693*
 types of, *III:623*
- Transformations, nonlinear,
III:630–631
- Transition probabilities, *I:368, I:381t*
- Treasuries
 correlations of, *III:405t*
 covariance matrix of, *III:406t*
 curve risk, *II:277t*
 discount function for, *III:564–565*
 futures, *I:482*
 inflation-indexed, *I:286*
 movements of, *III:403f*
 on-the-run, *I:227, III:7, III:560*
 par yield curve, *I:218t*
 spot rates, *I:220*
 3-month, *II:415–416, II:416f*
 volatility of, *III:404–406, III:406t*
- Treasury bill rates, weekly data, *I:89f*
- Treasury inflation-protected securities (TIPS). *See* TIPS (Treasury inflation-protected securities)
- Treasury Regulation T (Reg T), *I:67*
- Treasury securities, *I:210–211*
 comparable, defined, *III:5*
 in futures contracts, *I:483*
 hypothetical, illustration of duration/convexity, *III:308–310, III:308t*
 maturities of, *I:226*
 options on, *I:490*
 par rates for, *I:217*
 prediction of 10-year yield, *II:322–328*
 valuation of, *I:216*
 yield of, *II:324–327t*
- Treasury strips, *I:218t, I:220–221, I:286, III:560*
- Treasury yield curves, *I:226, III:561*
- Trees/lattices
 adjusted to current market price, *I:496f*
 bushy trees, *I:265, I:266f*
 calibrated, *I:495*
 convertible bond value, *I:274–275*
 extended pricing tree, *III:23f*
 from historical data, *III:131f*
 pruning of, *II:377*
 stock price, *I:274*
 three-period scenario, *III:131f*
 trinomial, *I:81, I:273, I:495–496*
 use of in modeling, *I:494–496*
- Trees/lattices, binomial
 building of, *I:273*
 for convertible bonds, *I:275f*
 discussion of, *I:80–81*
 interest rate, *I:244*
 model of, *I:273–275*
- stock price model, *III:173*
- term structure evolution, *I:495f*
 use of, *I:114–115, I:114f*
- Trends
 deterministic, *II:383*
 in financial time series, *II:504*
 and integrated series, *II:512–514*
 stochastic, *II:383, II:384*
- Treynor-Black model, *I:203n*
- Trinomial stochastic models, *II:11–12*
- Truncated Lévy flight (TLF), *III:382, III:384–386*
 IDD in, *III:386*
 time scaling of, *III:385f*
- Truncation, *III:385–386*
- Truth in Savings Act, *II:615*
- T*-statistic, *II:240n, II:336, II:350, II:390*
- Tuple, defined, *III:157*
- Turnover
 assessment of, *III:68*
 defined, *III:66*
 in MBSs, *III:48*
 in portfolios, *II:234, II:235*
- Two beta trap, *I:74–77*
- Two-factor models, *III:553–554*
- Two-stage growth model, *II:9*
- U.K. index-linked gilts, tax treatment of, *I:287*
- Uncertainties
 and Bayesian statistics, *I:140*
 in measurement processes, *II:367*
 modeling of, *II:306, III:124, III:131–132*
 and model risk, *II:729*
 quantification of, *I:101*
 representation of, *III:128*
 time behavior of, *II:359*
- Uncertainty sets
 effect of size of, *III:143*
 in portfolio allocation, *II:80*
 selection of, *III:140–141*
 structured, *III:143–144*
 in three dimensions, *II:81f*
 use of, *III:138, III:140*
- Uncertain volatility model, *II:673–674*
- Underperformance, finding reasons for, *II:118*
- Underwater, on homeowner's equity, *III:73*
- Unemployment rate
 as an economic measure, *II:398*
 application of TAR models to, *II:405–406*
 characteristics of series, *II:430*
 forecasts from, *II:433*
 performance of forecasting, *II:432–433, II:432t*
 and risk, *II:292n*
- test of nonlinearity, *II:431, II:431t*
- time plot of, *II:406f, II:430f*
- Uniqueness, theorem of, *III:490*
- Unit root series, *II:385*
- Univariate linear regression model, *I:163–170*
- Univariate stationary series, *II:504*
- U.S. Bankruptcy Code. *See also* bankruptcy
 Chapter 7, *I:350*
 Chapter 11, *I:342, I:350*
 Utility, *I:56, II:469, II:471, II:719–720*
- Validation, out of sample, *II:711*
- Valuation
 arbitrage-free, *I:216–217, I:220–222, I:221t*
 and cash flows, *I:223*
 defined, *I:209*
 effect of business cycle on, *I:303–304*
 fundamental principle of, *I:209*
 with Monte Carlo simulation, *III:6–12*
 of natural gas/oil storage, *I:560–561*
 of non-Treasury securities, *I:222–223*
 relative, *I:225, II:34–40, II:44–45*
 risk-neutral, *I:557, III:595–596, III:601*
 total firm, *II:21–23*
 uncertainty in, *II:15*
 use of lattices for, *I:240*
- Value
 absolute *vs.* relative basis of, *I:259–260*
 analysis of relative, *I:225*
 arbitrage-free, *I:221*
 book *vs.* market of firms, *II:559–560*
 determining present, *II:600–601*
 formulas for analysis of, *II:238–239*
 identification of relative, *I:405*
 intrinsic, *I:484–485*
 present, discounted, *II:601f*
 relative, *I:405, II:37–38*
vs. price, *I:455n*
- Value at risk (VaR). *See also* CVaR (credit value at risk)
 in backtesting, *II:748*
 backtesting of, *II:749f, III:325–327, III:365–367*
 boxplot of, *III:325f*
 and coherent risk measures, *III:329*
 conditional, *III:332, III:355–356, III:382*
 deficiencies in, *I:407, III:321, III:331–332, III:347*
 defined, *II:754n, III:319–322*
 density and distribution functions, *III:320f*

- Value at risk (VaR) (*Continued*)
determining from simulation,
III:639f
distribution-free confidence
intervals for, *III:292–293*
estimation of, *II:366, III:289–290,*
III:373–376, III:644, III:644t
exceedances of, *III:325–326*
IDD in, *III:290*
interest rate covariance matrix in,
III:403
levels of confidence with,
III:290–291
liquidity-adjusted, *III:374, III:376*
in low market volatility, *II:748*
measurements by, *II:354*
methods of computation, *III:323*
modeling of, *II:130–131, III:375–376*
and model risk, *II:695*
normal against confidence level,
III:294f
portfolio problem, *I:193*
in practice, *III:321–325*
relative spreads between
predictions, *II:750f, II:751f,*
II:752f
as safety-first risk measure,
III:355
standard normal distribution of,
III:324t
use of, *II:365*
vs. deviation measures, *III:320–321*
- Value of operations, process for
finding, *II:30t*
- Values, lagged, *II:130*
- Van der Korput sequences, *III:650*
- Variables
antithetic, *III:647–648*
application of macro, *II:193n*
behavior of, *III:152–153*
categorical, *II:333–334, II:350*
classification, *II:176*
declaration of in VBL, *III:457–458*
dependence between, *II:306–307*
dependent categorical, *II:348–350*
dependent/independent in CAPM,
I:67
dichotomous, *II:350*
dummy, *II:334*
exogenous *vs.* endogenous, *II:692*
fat-tailed, *III:280*
independent and identically
distributed, *II:125*
independent categorical, *II:333–348*
interactions between, *II:378*
large numbers of, *II:147*
macroeconomic, *II:54–55, II:177*
in maximum likelihood
calculations, *II:312–313*
mixing of categorical and
quantitative, *II:334–335*
nonstationary, *II:388–393*
as observation or measurement,
II:306
random, *I:159n*
in regression analysis, *II:330*
separable, *II:647*
slope, *III:553*
split formation of, *III:130f*
spread, *II:336*
standardization of, *II:205*
stationary, *II:385, II:386*
stationary/nonstationary, *II:384–386*
stochastic, *III:159–164*
use of dummy, *II:335, II:343–344*
- Variables, random, *II:297*
 α -stable, *III:242–244, III:244–245*
Bernoulli, *III:169*
continuous, *III:200–201, III:205–206*
on countable spaces, *III:160–161,*
III:166
defined, *III:162*
discrete, *III:165*
infinitely divisible, *III:253*
in probability, *III:159–164*
sequences of, *I:389*
on uncountable spaces, *III:161–162*
use of, *I:82*
- Variance gamma process, *III:499,*
III:504
- Variance matrix, *II:370–371*
- Variances
addressing inequality of, *I:168*
based on covariance matrix, *II:161t,*
II:163t, II:164f
conditional, *I:180*
conditional/unconditional, *II:361*
in dispersion parameters,
III:202–203
equal, *I:164*
as measure of risk, *I:8*
in probability, *III:167–169*
reduction in, *III:647–651*
unequal, *I:167–168, I:172*
- Variances/covariances, *II:112–113,*
II:302–303, III:395–396
- Variance swaps, *I:545–547, I:549,*
I:552
- Variational formulation, and finite
element space, *II:670–672*
- Variation margins, *I:478*
- Vasicek model
with change of time, *III:523–524*
for coupon-bond call options,
I:501–502
distribution of, *I:493*
in history, *I:491*
for short rates, *III:545–546*
use of, *I:89, I:497*
valuing zero-coupon bond calls
with, *I:499–500*
- VBA (Visual Basic for Applications)
built-in numeric functions of, *III:456*
comments in, *III:453*
control flow statements, *III:458–460*
debugging in, *III:461*
debugging tools of, *III:461, III:477*
example programs, *III:449–452,*
III:461–466
in Excel, *III:449, III:450f*
FactorialFun1, *III:455–456*
functions, user-defined, *III:463f*
functions in, *III:477*
generating Brownian motion paths
in, *III:463–465*
If statements, *III:459*
For loops, *III:458–459*
methods (actions) in, *III:452–453*
modules, defined, *III:455*
as object-oriented language, *III:452,*
III:466
objects in, *III:452*
operators in, *III:459–460*
Option Explicit command, *III:458*
pricing European call options,
III:465–466
programing of input dialog boxes,
III:460–461
programming tips for, *III:454–461*
properties in, *III:453*
random numbers in, *III:464–465*
subroutines and user-defined
functions in, *III:466–477*
subroutines *vs.* user-defined
functions in, *III:455–457*
use of Option Explicit command,
III:458
user-defined functions, *III:463f*
user interaction with, *III:460–461*
variable declaration in, *III:457–458*
With/End structure in, *III:453–454*
writing code in, *III:453–454*
- Vech notation, *II:371–372*
- VEC model, *II:372*
- Vector autoregressive (VAR) model,
II:393
- Vectors, *II:621–622, II:625–626, II:628*
- Vega, *I:521*
- Vichara Technology, *III:41–42, III:43t*
- Visual Basic for Applications (VBA).
See VBA
- Volatilities
absolute *vs.* relative, *III:404–405*
actual, *I:514*
aim of models of, *I:176*
analysis of, *II:270–272*
and ARCH models, *II:409*

- assumptions about, *III:7*
 calculation of, *II:272, III:534t*
 calculation of daily, *III:533–534*
 calibration of local, *II:681–685*
 clustering of, *II:359, II:716, III:402*
 confidence intervals for, *III:399–400*
 constant, *III:653*
 decisions for measuring, *III:403–404*
 defined, *III:533, III:653*
 with different mean reversions,
III:538f
 of the diffusion, *I:125*
 effect of local, *III:609*
 effect on hedging, *I:517–518*
 of energy commodities, *I:556–557*
 estimation of, *II:368–369*
 in EWMA estimates, *III:410–411*
 exposure to, *II:252f, II:252t*
 forecasts of, *I:179–180, II:172,*
II:367–368
 in FTSE 100, *III:412–413*
 historical, *I:513, III:534, III:654*
 hypothetical modelers of, *III:408*
 implied, *I:513–514, II:282, II:662,*
III:654
 in interest rate structure models,
I:492
 jump-diffusion, *III:657*
 level-dependent, *III:654–655,*
III:656
 local, *II:681, II:682–683, III:655*
 as a measure, *I:545, II:373*
 measurement of, *I:393, III:403–406*
 minimization of, *II:179*
 in models, *II:302*
 models of, *II:428*
 in option pricing, *I:513–514*
 patterns in, *I:395*
 in random walks, *I:84*
 and risk, *II:270*
 in risk-neutral measures, *III:587*
 smile of, *III:557*
 and the smoothing constant,
III:409–410
 states of, *I:180–181*
 stochastic, *I:94, I:547, I:548,*
III:655–658, III:656, III:658
 stochastic models, *II:681*
 time increments of, *I:83*
 of time series, *I:80*
 time-varying, *II:733–734*
 types of, *III:658*
 vs. annual standard deviation,
III:534
 Volatility clustering, *III:242, III:388*
 Volatility curves, *III:534–535,*
III:535t
- Volatility measures, nonstochastic,
III:654–655
 Volatility multiples, use of,
III:536
 Volatility risk, *I:509*
 Volatility skew, *III:550, III:551f,*
III:555–556, III:654
 measuring, *III:550*
 Volatility smile, *II:681, III:555–557,*
III:556f, III:654, III:656
 Volatility swaps, *I:545–547, I:552*
 for S&P Canada index (example),
I:550–552
 valuing of, *I:549*
 Volume-weighted average price
 (VWAP), *II:117, III:626–627*
 VPRs (voluntary prepayment rates)
 calculation of, *III:76*
 in cash flow calculators, *III:34*
 defined, *III:30*
 impacts of, *III:38*
- W. T. Grant, cash flows of, *II:576*
 Waldrop, Mitchell, *II:699*
 Wal-Mart, *II:569, II:570f*
 Walras, Leon, *II:467, II:468–469,*
II:474
 Waterfalls, development of, *III:8*
 Weak laws of large numbers (WLLN),
III:263
 Wealth, *I:460t, III:130*
 Weather, as chaotic system, *II:653*
 Weibull density, *III:107f*
 Weibull distributions, *III:106–107,*
III:112, III:229, III:262, III:265,
III:267, III:268
 Weighting, efficient, *I:41–42*
 Weights, *II:115, II:185t, II:231–232,*
II:724
 Weirton Steel, cash flows of,
II:577f
 What's the hedge, *I:300, I:303, I:306,*
I:417. See also hedge test
 White noise. *See noise, white*
 Wiener processes, *I:95, I:491, I:497,*
III:534–535, III:579, III:581
 Wilson, Kenneth, *II:480*
 Wind farms, valuation of, *I:563–564*
 Wold representation, *II:506*
 Working capital, *II:551*
 concept of, *II:567*
- XML (eXtensible Markup Language),
 development of, *II:482*
- Yield and bond loss matrix, *III:41f*
 Yield curve risk, *III:307, III:316–317*
- Yield curves
 horizon, *III:585*
 initial consistency with, *III:544*
 issuer par, *I:238t, I:244t*
 nonparallel, *III:309–310*
 parallel shifts in, *III:308–309*
 par-coupon, *III:585*
 reshaping duration, *III:315–316*
 in scenario analysis, *II:290*
 SEDUR/LEDUR, *III:316, III:317*
 shifts in, *III:586*
 slope of, *III:315*
 in term structures, *III:560*
 in valuation, *I:235*
- Yields
 calculation of, *II:613–618*
 comparison across countries, *I:226*
 dividend, *II:4*
 on investments, *II:617–618, II:619*
 loss-adjusted, *III:36, III:40*
 and loss matrix analysis, *III:40–41*
 projected, *III:37f, III:38f*
 real, *I:278–280, I:280f*
 rolling, *I:258–259*
- Yield spreads
 computation of, *I:226*
 determining, *I:373–374*
 for different rating grades, *I:374t*
 in Merton model, *I:305–306*
 over swap and treasury curves,
I:226–227
- Zero-coupon bonds
 assumptions about, *I:261*
 calculations using CIR model, *I:502t*
 calculations using Vasicek model,
I:502t
 defaultable, *I:317, I:335n*
 default-free, *I:318*
 development of valuation model
 for, *III:582–583*
 equations for, *III:554*
 future market price for, *I:492–493*
 lattices for, *I:266f*
 market for, *I:264*
 and martingales, *I:262*
 PDEs of, *I:268–269n*
 pricing of, *I:316*
 term structure model for, *III:584*
 value of, *III:572–573*
 valuing, *I:213, I:499–501, I:499t*
 Zero coupon rates, *III:546–547*
 Zero coupon securities, *I:218*
 Zero one distribution, *III:169–170*
 Zero volatility spread, *III:11–12*
 Zipf's law, *III:263, III:269*
 Z-scores, *II:191, II:240n*



ENCYCLOPEDIA
OF
FINANCIAL MODELS

II

FRANK J. FABOZZI, EDITOR

ENCYCLOPEDIA
OF
FINANCIAL MODELS

Volume II

ENCYCLOPEDIA
OF
FINANCIAL MODELS

Volume II

FRANK J. FABOZZI, EDITOR



WILEY

John Wiley & Sons, Inc.

Copyright © 2013 by Frank J. Fabozzi. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books. For more information about Wiley products, visit our web site at www.wiley.com.

ISBN: 978-1-118-00673-3 (3 v. set : cloth)

ISBN: 978-1-118-010327 (v. 1)

ISBN: 978-1-118-010334 (v. 2)

ISBN: 978-1-118-010341 (v. 3)

ISBN: 978-1-118-182365 (ebk.)

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

About the Editor

Frank J. Fabozzi is Professor of Finance at EDHEC Business School and a member of the EDHEC Risk Institute. Prior to joining EDHEC in August 2011, he held various professorial positions in finance at Yale University's School of Management from 1994 to 2011 and from 1986 to 1992 was a visiting professor of finance and accounting at MIT's Sloan School of Management. From 2008 to 2011, he was an affiliated professor in the Institute of Statistics, Econometrics, and Mathematical Finance at the University of Karlsruhe in Germany. Prior to 1986 he held professorial positions at Lafayette College, Fordham University, Queens College (CUNY), and Hofstra University. From 2003 to 2011, he served on Princeton University's Advisory Council for the Department of Operations Research and Financial Engineering and since then has been a visiting fellow in that department.

Professor Fabozzi is the editor of the *Journal of Portfolio Management*, as well as on the editorial board of the *Journal of Fixed Income*, *Journal of Asset Management*, *Quantitative Finance*, *Review of Futures Markets*, *Journal of Mathematical Finance*, *Journal of Structured Finance*, *Annals of Financial Economics*, and *Theoretical Economic Letters*.

He has authored and edited a number of books in asset management and quantitative finance. His coauthored books in quantitative finance include *A Probability Metrics Approach to Financial Risk Measures* (2011), *Financial Modeling with Lévy Processes and Volatility Clustering* (2011), *Quantitative Equity Investing: Techniques and Strategies* (2010), *Probability and Statistics for Finance* (2010), *Simulation and Optimization Modeling in Finance* (2010), *Bayesian Methods in Finance* (2008), *Advanced Stochastic Models, Risk Assessment, and Portfolio Optimization: The Ideal Risk* (2008), *Financial Econometrics: From Basics to Advanced Modeling Techniques* (2007), *Robust Portfolio Optimization and Management* (2007), and *Mathematics of Financial Modeling and Investment Management* (2004). His books in applied mathematics include *The Methods of Distances in the Theory of Probability and Statistics* (2013) and *Robust and Non-Robust Models in Statistics* (2009). He coauthored three monographs for the Research Foundation of the CFA Institute: *The Impact of the Financial Crisis on the Asset Management Industry* (2010), *Challenges in Quantitative Equity Management* (2008), and *Trends in Quantitative Finance* (2006).

Professor Fabozzi's research papers have appeared in numerous journals, including *Journal of Finance*, *Journal of Finance and Quantitative Analysis*, *Econometric Theory*, *Operations Research*, *Journal of Banking and Finance*, *Journal of Economic Dynamics and Control*, *Studies in Nonlinear Dynamics and Econometrics*, *European Journal of Operational Research*, *Annals of Operations Research*, *Quantitative Finance*, *European Financial Management*, and *The Econometric Journal*. His 2010 article published in *European Financial Management* with Professors Robert Shiller, and Radu Tunaru, "Property Derivatives for Managing European Real-Estate Risk," received the Best Paper Award and his paper with the same coauthors entitled "A Pricing Framework for Real Estate Derivatives" was awarded

Best Research Paper at the 10th Research Conference Campus for Finance held annually at WHU Otto Beisheim School of Management, Vallendar, Germany. An article coauthored with Dr. Sergio Focardi, "An Autoregressive Conditional Duration Model of Credit Risk Contagion," published in 2005 in *Journal of Risk Finance* was the winner of the 2006 Outstanding Paper by Emerald Literati Network.

He has received several awards and honors for his body of work. In 1994 he was awarded an Honorary Doctorate of Humane Letters from Nova Southeastern University. In 2002 he was inducted into the Fixed Income Analysts Society's Hall of Fame, established by the society "to recognize the lifetime achievements of outstanding practitioners in the advancement of the analysis of fixed-income securities and portfolios." In 2007 he was the recipient of the C. Stewart Sheppard Award given by the CFA Institute "in recognition of outstanding contribution to continuing education in the CFA profession." He was the cover story in the July 1999 issue of *Bloomberg Magazine* entitled "The Boswell of Bonds."

Professor Fabozzi was the co-founder of Information Management Network (now a subsidiary of Euromoney), a conference company specializing in financial topics. He is a trustee for the BlackRock family of closed-end funds where he is the chair of the performance committee and a member of the audit committee. He was a director of Guardian Mutual Funds and Guardian Annuity Funds.

He earned both an M.A. and B.A. in economics and statistics in June 1970 from the City College of New York and elected to Phi Beta Kappa in 1969. He earned a Ph.D. in Economics in September 1972 from the City University of New York. Professor Fabozzi holds two professional designations: Chartered Financial Analyst (1977) and Certified Public Accountant (1982).

Contents

Contributors	xi		
Preface	xvii		
Guide to the <i>Encyclopedia of Financial Models</i>	xxxiii		
Index	757		
Volume I			
Asset Allocation	1		
Mean-Variance Model for Portfolio Selection	3		
Principles of Optimization for Portfolio Selection	21		
Asset Allocation and Portfolio Construction Techniques in Designing the Performance-Seeking Portfolio	35		
Asset Pricing Models	47		
General Principles of Asset Pricing	49		
Capital Asset Pricing Models	65		
Modeling Asset Price Dynamics	79		
Arbitrage Pricing: Finite-State Models	99		
Arbitrage Pricing: Continuous-State, Continuous-Time Models	121		
Bayesian Analysis and Financial Modeling Applications	137		
Basic Principles of Bayesian Analysis	139		
Introduction to Bayesian Inference	151		
Bayesian Linear Regression Model	163		
Bayesian Estimation of ARCH-Type Volatility Models	175		
Bayesian Techniques and the Black-Litterman Model	189		
Bond Valuation		207	
Basics of Bond Valuation		209	
Relative Value Analysis of Fixed-Income Products		225	
Yield Curves and Valuation Lattices		235	
Using the Lattice Model to Value Bonds with Embedded Options, Floaters, Option, and Caps/Floors		243	
Understanding the Building Blocks for OAS Models		257	
Quantitative Models to Value Convertible Bonds		271	
Quantitative Approaches to Inflation-Indexed Bonds		277	
Credit Risk Modeling		297	
An Introduction to Credit Risk Models		299	
Default Correlation in Intensity Models for Credit Risk Modeling		313	
Structural Models in Credit Risk Modeling		341	
Modeling Portfolio Credit Risk		361	
Simulating the Credit Loss Distribution		377	
Managing Credit Spread Risk Using Duration Times Spread (DTS)		391	
Credit Spread Decomposition		401	
Credit Derivatives and Hedging Credit Risk		407	
Derivatives Valuation		421	
No-Arbitrage Price Relations for Forwards, Futures, and Swaps		423	
No-Arbitrage Price Relations for Options		437	
Introduction to Contingent Claims Analysis		457	
Black-Scholes Option Pricing Model		465	

Pricing of Futures/Forwards and Options	477	Classification and Regression Trees and Their Use in Financial Modeling	375
Pricing Options on Interest Rate Instruments	489	Applying Cointegration to Problems in Finance	383
Basics of Currency Option Pricing Models	507	Nonlinearity and Nonlinear Econometric Models in Finance	401
Credit Default Swap Valuation	525	Robust Estimates of Betas and Correlations	437
Valuation of Fixed Income Total Return Swaps	541	Working with High-Frequency Data	449
Pricing of Variance, Volatility, Covariance, and Correlation Swaps	545	Financial Modeling Principles	465
Modeling, Pricing, and Risk Management of Assets and Derivatives in Energy and Shipping	555	Milestones in Financial Modeling	467
		From Art to Financial Modeling	479
		Basic Data Description for Financial Modeling and Analysis	485
		Time Series Concepts, Representations, and Models	501
		Extracting Risk-Neutral Density Information from Options Market Prices	521
		Financial Statement Analysis	529
		Financial Statements	531
		Financial Ratio Analysis	545
		Cash-Flow Analysis	565
		Finite Mathematics for Financial Modeling	579
		Important Functions and Their Features	581
		Time Value of Money	595
		Fundamentals of Matrix Algebra	621
		Difference Equations	629
		Differential Equations	643
		Partial Differential Equations in Finance	659
		Model Risk and Selection	689
		Model Risk	691
		Model Selection and Its Pitfalls	699
		Managing the Model Risk with the Methods of the Probabilistic Decision Theory	719
		Fat-Tailed Models for Risk Estimation	731
		Volume III	
		Mortgage-Backed Securities Analysis and Valuation	1
		Valuing Mortgage-Backed and Asset-Backed Securities	3
		The Active-Passive Decomposition Model for MBS	17
		Analysis of Nonagency Mortgage-Backed Securities	29
Volume II			
Equity Models and Valuation	1		
Dividend Discount Models	3		
Discounted Cash Flow Methods for Equity Valuation	15		
Relative Valuation Methods for Equity Analysis	33		
Equity Analysis in a Complex Market	47		
Equity Portfolio Selection Models in Practice	61		
Basics of Quantitative Equity Investing	89		
Quantitative Equity Portfolio Management	107		
Forecasting Stock Returns	121		
Factor Models for Portfolio Construction	135		
Factor Models	137		
Principal Components Analysis and Factor Analysis	153		
Multifactor Equity Risk Models and Their Applications	171		
Factor-Based Equity Portfolio Construction and Analysis	195		
Cross-Sectional Factor-Based Models and Trading Strategies	213		
The Fundamentals of Fundamental Factor Models	243		
Multifactor Equity Risk Models and Their Applications	255		
Multifactor Fixed Income Risk Models and Their Applications	267		
Financial Econometrics	293		
Scope and Methods of Financial Econometrics	295		
Regression Analysis: Theory and Estimation	305		
Categorical and Dummy Variables in Regression Models	333		
Quantile Regression	353		
ARCH/GARCH Models in Applied Financial Econometrics	359		

Measurements of Prepayments for Residential Mortgage-Backed Securities	47	Back-Testing Market Risk Models	361
Prepayments and Factors Influencing the Return of Principal for Residential Mortgage-Backed Securities	65	Estimating Liquidity Risks	371
Operational Risk	79	Estimate of Downside Risk with Fat-Tailed and Skewed Models	381
Operational Risk	81	Moving Average Models for Volatility and Correlation, and Covariance Matrices	395
Operational Risk Models	91	Software for Financial Modeling	415
Modeling Operational Loss Distributions	103	Introduction to Financial Model Building with MATLAB	417
Optimization Tools	121	Introduction to Visual Basic for Applications	449
Introduction to Stochastic Programming and Its Applications to Finance	123	Stochastic Processes and Tools	469
Robust Portfolio Optimization	137	Stochastic Integrals	471
Probability Theory	149	Stochastic Differential Equations	485
Concepts of Probability Theory	151	Stochastic Processes in Continuous Time	495
Discrete Probability Distributions	165	Conditional Expectation and Change of Measure	507
Continuous Probability Distributions	195	Change of Time Methods	519
Continuous Probability Distributions with Appealing Statistical Properties	207	Term Structure Modeling	531
Continuous Probability Distributions Dealing with Extreme Events	227	The Concept and Measures of Interest Rate Volatility	533
Stable and Tempered Stable Distributions	241	Short-Rate Term Structure Models	543
Fat Tails, Scaling, and Stable Laws	259	Static Term Structure Modeling in Discrete and Continuous Time	559
Copulas	283	The Dynamic Term Structure Model	575
Applications of Order Statistics to Risk Management Problems	289	Essential Classes of Interest Rate Models and Their Use	593
Risk Measures	297	A Review of No Arbitrage Interest Rate Models	603
Measuring Interest Rate Risk: Effective Duration and Convexity	299	Trading Cost Models	621
Yield Curve Risk Measures	307	Modeling Market Impact Costs	623
Value-at-Risk	319	Volatility	635
Average Value-at-Risk	331	Monte Carlo Simulation in Finance	637
Risk Measures and Portfolio Selection	349	Stochastic Volatility	653

Contributors

Yves Achdou, PhD

Professor, Lab. Jacques-Louis Lions, University Paris-Diderot, Paris, France

Irene Aldridge

Managing Partner, Able Alpha Trading

Carol Alexander, PhD

Professor of Finance, University of Sussex

Andrew Alford, PhD

Managing Director, Quantitative Investment Strategies, Goldman Sachs Asset Management

Noël Amenc, PhD

Professor of Finance, EDHEC Business School, Director, EDHEC-Risk Institute

Bala Arshanapalli, PhD

Professor of Finance, Indiana University Northwest

David Audley, PhD

Senior Lecturer, The Johns Hopkins University

Jennifer Bender, PhD

Vice President, MSCI

William S. Berliner

Executive Vice President, Manhattan Advisory Services Inc.

Anand K. Bhattacharya, PhD

Professor of Finance Practice, Department of Finance, W. P. Carey School of Business, Arizona State University

Michele Leonardo Bianchi, PhD

Research Analyst, Specialized Intermediaries Supervision Department, Bank of Italy

Olivier Bokanowski

Associate Professor, Lab. Jacques-Louis Lions, University Paris-Diderot, Paris, France

Gerald W. Buetow Jr., PhD, CFA

President and Founder, BFRC Services, LLC

Paul Bukowski, CFA, FCAS

Executive President, Head of Equities, Hartford Investment Management

Joseph A. Cerniglia

Visiting Researcher, Courant Institute of Mathematical Sciences, New York University

Ren-Raw Chen

Professor of Finance, Graduate School of Business, Fordham University

Anna Chernobai, PhD

Assistant Professor of Finance, M. J. Whitman School of Management, Syracuse University

Richard Chin

Investment Manager, New York Life Investments

António Baldaque da Silva
Managing Director, Barclays

Siddhartha G. Dastidar, PhD, CFA
Vice President, Barclays

Arik Ben Dor, PhD
Managing Director, Barclays

Michael Dorigan, PhD
Senior Quantitative Analyst, PNC Capital
Advisors

Kevin Dowd, PhD
Partner, Cobden Partners, London

Pamela P. Drake, PhD, CFA
J. Gray Ferguson Professor of Finance, College
of Business, James Madison University

Lev Dynkin, PhD
Managing Director, Barclays

Brian Eales
Academic Leader (Retired), London Metropolitan
University

Abel Elizalde, PhD
Credit Derivatives Strategy, J.P. Morgan

Robert F. Engle, PhD
Michael Armellino Professorship in the Man-
agement of Financial Services and Director of
the Volatility Institute, Leonard N. Stern School
of Business, New York University

Frank J. Fabozzi, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

Peter Fitton
Manager, Scientific Development, CreditXpert
Inc.

Sergio M. Focardi, PhD
Partner, The Intertek Group

Radu Găbudean, PhD
Vice President, Barclays

Vacslav S. Glukhov, PhD
Head of Quantitative Strategies and Data Ana-
lytics, Liquidnet Europe Ltd, London, United
Kingdom

Felix Goltz, PhD
Head of Applied Research, EDHEC-Risk
Institute

Chris Gowlland, CFA
Senior Quantitative Analyst, Delaware Invest-
ments

Biliana S. Güner
Assistant Professor of Statistics and Economet-
rics, Özyeğin University, Turkey

Francis Gupta, PhD
Director, Index Research & Design, Dow Jones
Indexes

Markus Höchstötter, PhD
Assistant Professor, University of Karlsruhe

John S. J. Hsu, PhD
Professor of Statistics and Applied Probability,
University of California, Santa Barbara

Jay Hyman, PhD
Managing Director, Barclays, Tel Aviv

Bruce I. Jacobs, PhD
Principal, Jacobs Levy Equity Management

Robert R. Johnson, PhD, CFA
Independent Financial Consultant,
Charlottesville, VA

Frank J. Jones, PhD
Professor, Accounting and Finance Depart-
ment, San Jose State University and Chairman,
Investment Committee, Private Ocean Wealth
Management

Robert Jones, CFA
Chairman, Arwen Advisors, and Chairman and
CIO, Systems Two Advisors

Andrew Kalotay, PhD

President, Andrew Kalotay Associates

Young Shin Kim, PhD

Research Assistant Professor, School of Economics and Business Engineering, University of Karlsruhe and KIT

Petter N. Kolm, PhD

Director of the Mathematics in Finance Masters Program and Clinical Associate Professor, Courant Institute of Mathematical Sciences, New York University

Glen A. Larsen Jr., PhD CFA

Professor of Finance, Indiana University Kelley School of Business—Indianapolis

Anthony Lazanas

Managing Director, Barclays

Arturo Leccadito, PhD

Business Administration Department, Università della Calabria

Tony Lelièvre, PhD

Professor, CERMICS, Ecole des Ponts Paristech, Marne-la-Vallée, France

Alexander Levin, PhD

Director, Financial Engineering, Andrew Davidson & Co., Inc.

Kenneth N. Levy, CFA

Principal, Jacobs Levy Equity Management

Terence Lim, PhD, CFA

CEO, Arwen Advisors

Peter C. L. Lin

PhD Candidate, The Johns Hopkins University

Steven V. Mann, PhD

Professor of Finance, Moore School of Business, University of South Carolina

Harry M. Markowitz, PhD

Consultant and Nobel Prize Winner, Economics, 1990

Lionel Martellini, PhD

Professor of Finance, EDHEC Business School, Scientific Director, EDHEC-Risk Institute

James F. McNatt, CFA

Executive Vice President, ValueWealth Services

Christian Menn, Dr Rer Pol

Managing Partner, RIVACON

Ivan Mitov

Head of Quantitative Research, FinAnalytica

Edwin H. Neave

Professor Emeritus, School of Business, Queen's University, Kingston, Ontario

William Nelson, PhD

Professor of Finance, Indiana University Northwest

Frank Nielsen

Managing Director of Quantitative Research, Fidelity Investments - Global Asset Allocation

Philip O. Obazee

Senior Vice President and Head of Derivatives, Delaware Investments

Dominic O'Kane, PhD

Affiliated Professor of Finance, EDHEC Business School, Nice, France

Dessislava A. Pachamanova

Associate Professor of Operations Research, Babson College

Bruce D. Phelps

Managing Director, Barclays

Thomas K. Philips, PhD

Regional Head of Investment Risk and Performance, BNP Paribas Investment Partners

David Philpotts

QEP Global Equities, Schroder Investment Management, Sydney, Australia

Wesley Phoa

Senior Vice President, Capital International Research, Inc.

Svetlozar T. Rachev, PhD Dr Sci

Frey Family Foundation Chair Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

Boryana Racheva-Yotova, PhD

President, FinAnalytica

Shrikant Ramanmurthy

Consultant, New York, NY

Srichander Ramaswamy, PhD

Senior Economist, Bank for International Settlements, Basel, Switzerland

Patrice Retkowsky

Senior Research Engineer, EDHEC-Risk Institute

Paul Sclavounos

Department of Mechanical Engineering, Massachusetts Institute of Technology

Shani Shamah

Consultant, RBC Capital Markets

Koray D. Simsek, PhD

Associate Professor, Sabanci School of Management, Sabanci University

James Sochacki

Professor of Applied Mathematics, James Madison University

Arne D. Staal

Director, Barclays

Maxwell J. Stevenson, PhD

Discipline of Finance, Business School, University of Sydney, Australia

Filippo Stefanini

Head of Hedge Funds and Manager Selection, Eurizon Capital SGR

Stoyan V. Stoyanov, PhD

Professor of Finance at EDHEC Business School and Head of Research for EDHEC Risk Institute-Asia

Anatoliy Swishchuk, PhD

Professor of Mathematics and Statistics, University of Calgary

Ruey S. Tsay, PhD

H.G.B. Alexander Professor of Econometrics and Statistics, University of Chicago Booth School of Business

Radu S. Tunaru

Professor of Quantitative Finance, Business School, University of Kent

Cenk Ural, PhD

Vice President, Barclays

Donald R. Van Deventer, PhD

Chairman and Chief Executive Officer, Kamakura Corporation

Raman Vardharaj

Vice President, Oppenheimer Funds

Robert E. Whaley, PhD

Valere Blair Potter Professor of Management and Co-Director of the Financial Markets Research Center, Owen Graduate School of Management, Vanderbilt University

Mark B. Wickard

Senior Vice President/Corporate Cash
Investment Advisor, Morgan Stanley Smith
Bamey

James X. Xiong, PhD, CFA

Senior Research Consultant, Ibbotson
Associates, A Morningstar Company

Guofu Zhou

Frederick Bierman and James E. Spears Profes-
sor of Finance, Olin Business School, Washing-
ton University in St. Louis

Min Zhu

Business School, Queensland University of
Technology, Australia

Preface

It is often said that investment management is an art, not a science. However, since the early 1990s the market has witnessed a progressive shift toward a more industrial view of the investment management process. There are several reasons for this change. First, with globalization the universe of investable assets has grown many times over. Asset managers might have to choose from among several thousand possible investments from around the globe. Second, institutional investors, often together with their consultants, have encouraged asset management firms to adopt an increasingly structured process with documented steps and measurable results. Pressure from regulators and the media is another factor. Finally, the sheer size of the markets makes it imperative to adopt safe and repeatable methodologies.

In its modern sense, financial modeling is the design (or engineering) of financial instruments and portfolios of financial instruments that result in predetermined cash flows contingent upon different events. Broadly speaking, financial models are employed to manage investment portfolios and risk. The objective is the transfer of risk from one entity to another via appropriate financial arrangements. Though the aggregate risk is a quantity that cannot be altered, risk can be transferred if there is a willing counterparty.

Financial modeling came to the forefront of finance in the 1980s, with the broad diffusion

of derivative instruments. However, the concept and practice of financial modeling are quite old. The notion of the diversification of risk (central to modern risk management) and the quantification of insurance risk (a requisite for pricing insurance policies) were already understood, at least in practical terms, in the 14th century. The rich epistolary of Francesco Datini, a 14th-century merchant, banker, and insurer from Prato (Tuscany, Italy), contains detailed instructions to his agents on how to diversify risk and insure cargo.

What is specific to modern financial modeling is the quantitative management of risk. Both the pricing of contracts and the optimization of investments require some basic capabilities of statistical modeling of financial contingencies. It is the size, diversity, and efficiency of modern competitive markets that makes the use of financial modeling imperative.

This three-volume encyclopedia offers not only coverage of the fundamentals and advances in financial modeling but provides the mathematical and statistical techniques needed to develop and test financial models, as well as the practical issues associated with implementation. The encyclopedia offers the following unique features:

- The entries for the encyclopedia were written by experts from around the world. This diverse collection of expertise has created the most definitive coverage of established and

cutting-edge financial models, applications, and tools in this ever-evolving field.

- The series emphasizes both technical and managerial issues. This approach provides researchers, educators, students, and practitioners with a balanced understanding of the topics and the necessary background to deal with issues related to financial modeling.
- Each entry follows a format that includes the author, entry abstract, introduction, body, listing of key points, notes, and references. This enables readers to pick and choose among various sections of an entry, and creates consistency throughout the entire encyclopedia.
- The numerous illustrations and tables throughout the work highlight complex topics and assist further understanding.
- Each volume includes a complete table of contents and index for easy access to various parts of the encyclopedia.

TOPIC CATEGORIES

As is the practice in the creation of an encyclopedia, the topic categories are presented alphabetically. The topic categories and a brief description of each topic follow.

VOLUME I

Asset Allocation

A major activity in the investment management process is establishing policy guidelines to satisfy the investment objectives. Setting policy begins with the asset allocation decision. That is, a decision must be made as to how the funds to be invested should be distributed among the major asset classes (e.g., equities, fixed income, and alternative asset classes). The term “asset allocation” includes (1) policy asset allocation, (2) dynamic asset allocation, and (3) tactical asset allocation. Policy asset allocation decisions can loosely be characterized as long-term asset allocation decisions, in which the investor seeks to assess an appropriate long-term “normal” asset mix that represents an ideal blend of controlled risk and enhanced return. In dynamic asset allocation the asset mix (i.e., the

allocation among the asset classes) is mechanically shifted in response to changing market conditions. Once the policy asset allocation has been established, the investor can turn his or her attention to the possibility of active departures from the normal asset mix established by policy. If a decision to deviate from this mix is based upon rigorous objective measures of value, it is often called tactical asset allocation. The fundamental model used in establishing the policy asset allocation is the mean-variance portfolio model formulated by Harry Markowitz in 1952, popularly referred to as the theory of portfolio selection and modern portfolio theory.

Asset Pricing Models

Asset pricing models seek to formalize the relationship that should exist between asset returns and risk if investors behave in a hypothesized manner. At its most basic level, asset pricing is mainly about transforming asset payoffs into prices. The two most well-known asset pricing models are the arbitrage pricing theory and the capital asset pricing model. The fundamental theorem of asset pricing asserts the equivalence of three key issues in finance: (1) absence of arbitrage; (2) existence of a positive linear pricing rule; and (3) existence of an investor who prefers more to less and who has maximized his or her utility. There are two types of arbitrage opportunities. The first is paying nothing today and obtaining something in the future, and the second is obtaining something today and with no future obligations. Although the principle of absence of arbitrage is fundamental for understanding asset valuation in a competitive market, there are well-known limits to arbitrage resulting from restrictions imposed on rational traders, and, as a result, pricing inefficiencies may exist for a period of time.

Bayesian Analysis and Financial Modeling Applications

Financial models describe in mathematical terms the relationships between financial random variables through time and/or across assets. The fundamental assumption is that the

model relationship is valid independent of the time period or the asset class under consideration. Financial data contain both meaningful information and random noise. An adequate financial model not only extracts optimally the relevant information from the historical data but also performs well when tested with new data. The uncertainty brought about by the presence of data noise makes imperative the use of statistical analysis as part of the process of financial model building, model evaluation, and model testing. Statistical analysis is employed from the vantage point of either of the two main statistical philosophical traditions—frequentist and Bayesian. An important difference between the two lies with the interpretation of the concept of probability. As the name suggests, advocates of the frequentist approach interpret the probability of an event as the limit of its long-run relative frequency (i.e., the frequency with which it occurs as the amount of data increases without bound). Since the time financial models became a mainstream tool to aid in understanding financial markets and formulating investment strategies, the framework applied in finance has been the frequentist approach. However, strict adherence to this interpretation is not always possible in practice. When studying rare events, for instance, large samples of data may not be available, and in such cases proponents of frequentist statistics resort to theoretical results. The Bayesian view of the world is based on the subjectivist interpretation of probability: Probability is subjective, a degree of belief that is updated as information or data are acquired. Only in the last two decades has Bayesian statistics started to gain greater acceptance in financial modeling, despite its introduction about 250 years ago. It has been the advancements of computing power and the development of new computational methods that have fostered the growing use of Bayesian statistics in financial modeling.

Bond Valuation

The value of any financial asset is the present value of its expected future cash flows. To value

a bond (also referred to as a fixed-income security), one must be able to estimate the bond's remaining cash flows and identify the appropriate discount rate(s) at which to discount the cash flows. The traditional approach to bond valuation is to discount every cash flow with the same discount rate. Simply put, the relevant term structure of interest rate used in valuation is assumed to be flat. This approach, however, permits opportunities for arbitrage. Alternatively, the arbitrage-free valuation approach starts with the premise that a bond should be viewed as a portfolio or package of zero-coupon bonds. Moreover, each of the bond's cash flows is valued using a unique discount rate that depends on the term structure of interest rates and when in time the cash flow is. The relevant set of discount rates (that is, spot rates) is derived from an appropriate term structure of interest rates and when used to value risky bonds augmented with a suitable risk spread or premium. Rather than modeling to calculate the fair value of its price, the market price can be taken as given so as to compute a yield measure or a spread measure. Popular yield measures are the yield to maturity, yield to call, yield to put, and cash flow yield. Nominal spread, static (or zero-volatility) spread, and option-adjusted spread are popular relative value measures quoted in the bond market. Complications in bond valuation arise when a bond has one or more embedded options such as call, put, or conversion features. For bonds with embedded options, the financial modeling draws from options theory, more specifically, the use of the lattice model to value a bond with embedded options.

Credit Risk Modeling

Credit risk is a broad term used to refer to three types of risk: default risk, credit spread risk, and downgrade risk. Default risk is the risk that the counterparty to a transaction will fail to satisfy the terms of the obligation with respect to the timely payment of interest and repayment of the amount borrowed. The counterparty could be the issuer of a debt obligation or an entity on

the other side of a private transaction such as a derivative trade or a collateralized loan agreement (i.e., a repurchase agreement or a securities lending agreement). The default risk of a counterparty is often initially gauged by the credit rating assigned by one of the three rating companies—Standard & Poor’s, Moody’s Investors Service, and Fitch Ratings. Although default risk is the one that most market participants think of when reference is made to credit risk, even in the absence of default, investors are concerned about the decline in the market value of their portfolio bond holdings due to a change in credit spread or the price performance of their holdings relative to a bond index. This risk is due to an adverse change in credit spreads, referred to as credit spread risk, or when it is attributed solely to the downgrade of the credit rating of an entity, it is called downgrade risk. Financial modeling of credit risk is used (1) to measure, monitor, and control a portfolio’s credit risk, and (2) to price credit risky debt instruments. There are two general categories of credit risk models: structural models and reduced-form models. There is considerable debate as to which type of model is the best to employ.

Derivatives Valuation

A derivative instrument is a contract whose value depends on some underlying asset. The term “derivative” is used to describe this product because its value is derived from the value of the underlying asset. The underlying asset, simply referred to as the “underlying,” can be either a commodity, a financial instrument, or some reference entity such as an interest rate or stock index, leading to the classification of commodity derivatives and financial derivatives. Although there are close conceptual relations between derivative instruments and cash market instruments such as debt and equity, the two classes of instruments are used differently: Debt and equity are used primarily for raising funds from investors, while derivatives are primarily

used for dividing up and trading risks. Moreover, debt and equity are direct claims against a firm’s assets, while derivative instruments are usually claims on a third party. A derivative’s value depends on the value of the underlying, but the derivative instrument itself represents a claim on the “counterparty” to the trade. Derivatives instruments are classified in terms of their payoff characteristics: linear and nonlinear payoffs. The former, also referred to as symmetric payoff derivatives, includes forward, futures, and swap contracts while the latter include options. Basically, a linear payoff derivative is a risk-sharing arrangement between the counterparties since both are sharing the risk regarding the price of the underlying. In contrast, nonlinear payoff derivative instruments (also referred to as asymmetric payoff derivatives) are insurance arrangements because one party to the trade is willing to insure the counterparty of a minimum or maximum (depending on the contract) price. The amount received by the insuring party is referred to as the contract price or premium. Derivative instruments are used for controlling risk exposure with respect to the underlying. Hedging is a special case of risk control where a party seeks to eliminate the risk exposure. Derivative valuation or pricing is developed based on no-arbitrage price relations, relying on the assumption that two perfect substitutes must have the same price.

VOLUME II

Difference Equations and Differential Equations

The tools of linear difference equations and differential equations have found many applications in finance. A difference equation is an equation that involves differences between successive values of a function of a discrete variable. A function of such a variable is one that provides a rule for assigning values in sequences to it. The theory of linear difference equations covers three areas: solving difference equations, describing the behavior

of difference equations, and identifying the equilibrium (or critical value) and stability of difference equations. Linear difference equations are important in the context of dynamic econometric models. Stochastic models in finance are expressed as linear difference equations with random disturbances added. Understanding the behavior of solutions of linear difference equations helps develop intuition for the behavior of these models. In nontechnical terms, differential equations are equations that express a relationship between a function and one or more derivatives (or differentials) of that function. The relationship between difference equations and differential equations is that the latter are invaluable for modeling situations in finance where there is a continually changing value. The problem is that not all changes in value occur continuously. If the change in value occurs incrementally rather than continuously, then differential equations have their limitations. Instead, a financial modeler can use difference equations, which are recursively defined sequences. It would be difficult to overemphasize the importance of differential equations in financial modeling where they are used to express laws that govern the evolution of price probability distributions, the solution of economic variational problems (such as intertemporal optimization), and conditions for continuous hedging (such as in the Black-Scholes option pricing model). The two broad types of differential equations are ordinary differential equations and partial differential equations. The former are equations or systems of equations involving only one independent variable. Another way of saying this is that ordinary differential equations involve only total derivatives. Partial differential equations are differential equations or systems of equations involving partial derivatives. When one or more of the variables is a stochastic process, we have the case of stochastic differential equations and the solution is also a stochastic process. An assumption must be made about what is driving noise in a stochastic differential

equation. In most applications, it is assumed that the noise term follows a Gaussian random variable, although other types of random variables can be assumed.

Equity Models and Valuation

Traditional fundamental equity analysis involves the analysis of a company's operations for the purpose of assessing its economic prospects. The analysis begins with the financial statements of the company in order to investigate the earnings, cash flow, profitability, and debt burden. The fundamental analyst will look at the major product lines, the economic outlook for the products (including existing and potential competitors), and the industries in which the company operates. The result of this analysis will be the growth prospects of earnings. Based on the growth prospects of earnings, a fundamental analyst attempts to determine the fair value of the stock using one or more equity valuation models. The two most commonly used approaches for valuing a firm's equity are based on discounted cash flow and relative valuation models. The principal idea underlying discounted cash flow models is that what an investor pays for a share of stock should reflect what is expected to be received from it—return on the investor's investment. What an investor receives are cash dividends in the future. Therefore, the value of a share of stock should be equal to the present value of all the future cash flows an investor expects to receive from that share. To value stock, therefore, an investor must project future cash flows, which, in turn, means projecting future dividends. Popular discounted cash flow models include the basic dividend discount model, which assumes a constant dividend growth, and the multiple-phase models, which include the two-stage dividend growth model and the stochastic dividend discount models. Relative valuation methods use multiples or ratios—such as price/earnings, price/book, or price/free cash flow—to determine whether a stock is trading at higher or lower multiples than its peers.

There are two critical assumptions in using relative valuation: (1) the universe of firms selected to be included in the peer group are in fact comparable, and (2) the average multiple across the universe of firms can be treated as a reasonable approximation of “fair value” for those firms. This second assumption may be problematic during periods of market panic or euphoria. Managers of quantitative equity firms employ techniques that allow them to identify attractive stock candidates, focusing not on a single stock as is done with traditional fundamental analysis but rather on stock characteristics in order to explain why one stock outperforms another stock. They do so by statistically identifying a group of characteristics to create a quantitative selection model. In contrast to the traditional fundamental stock selection, quantitative equity managers create a repeatable process that utilizes the stock selection model to identify attractive stocks. Equity portfolio managers have used various statistical models for forecasting returns and risk. These models, referred to as predictive return models, make conditional forecasts of expected returns using the current information set. Predictive return models include regressive models, linear autoregressive models, dynamic factor models, and hidden-variable models.

Factor Models and Portfolio Construction

Quantitative asset managers typically employ multifactor risk models for the purpose of constructing and rebalancing portfolios and analyzing portfolio performance. A multifactor risk model, or simply factor model, attempts to estimate and characterize the risk of a portfolio, either relative to a benchmark such as a market index or in absolute value. The model allows the decomposition of risk factors into a systematic and an idiosyncratic component. The portfolio’s risk exposure to broad risk factors is captured by the systematic risk. For equity portfolios these are typically fundamental factors (e.g., market capitalization and value

vs. growth), technical (e.g., momentum), and industry/sector/country. For fixed-income portfolios, systematic risk captures a portfolio’s exposure to broad risk factors such as the term structure of interest rates, credit spreads, optionality (call and prepayment), credit, and sectors. The portfolio’s systematic risk depends not only on its exposure to these risk factors but also the volatility of the risk factors and how they correlate with each other. In contrast to systematic risk, idiosyncratic risk captures the uncertainty associated with news affecting the holdings of individual issuers in the portfolio. In equity portfolios, idiosyncratic risk can be easily diversified by reducing the importance of individual issuers in the portfolio. Because of the larger number of issuers in bond indexes, however, this is a difficult task. There are different types of factor models depending on the factors. Factors can be exogenous variables or abstract variables formed by portfolios. Exogenous factors (or known factors) can be identified from traditional fundamental analysis or from economic theory that suggests macroeconomic factors. Abstract factors, also called unidentified or latent factors, can be determined with the statistical tool of factor analysis or principal component analysis. The simplest type of factor models is where the factors are assumed to be known or observable, so that time-series data are those factors that can be used to estimate the model. The four most commonly used approaches for the evaluation of return premiums and risk characteristics to factors are portfolio sorts, factor models, factor portfolios, and information coefficients. Despite its use by quantitative asset managers, the basic building blocks of factor models used by model builders and by traditional fundamental analysts are the same: They both seek to identify the drivers of returns for the asset class being analyzed.

Financial Econometrics

Econometrics is the branch of economics that draws heavily on statistics for testing and

analyzing economic relationships. The economic equivalent of the laws of physics, econometrics represents the quantitative, mathematical laws of economics. Financial econometrics is the econometrics of financial markets. It is a quest for models that describe financial time series such as prices, returns, interest rates, financial ratios, defaults, and so on. Although there are similarities between financial econometric models and models of the physical sciences, there are two important differences. First, the physical sciences aim at finding immutable laws of nature; econometric models model the economy or financial markets—artifacts subject to change. Because the economy and financial markets are artifacts subject to change, econometric models are not unique representations valid throughout time; they must adapt to the changing environment. Second, while basic physical laws are expressed as differential equations, financial econometrics uses both continuous-time and discrete-time models.

Financial Modeling Principles

The origins of financial modeling can be traced back to the development of mathematical equilibrium at the end of the nineteenth century, followed in the beginning of the twentieth century with the introduction of sophisticated mathematical tools for dealing with the uncertainty of prices and returns. In the 1950s and 1960s, financial modelers had tools for dealing with probabilistic models for describing markets, the principles of contingent claims analysis, an optimization framework for portfolio selection based on mean and variance of asset returns, and an equilibrium model for pricing capital assets. The 1970s ushered in models for pricing contingent claims and a new model for pricing capital assets based on arbitrage pricing. Consequently, by the end of the 1970s, the frameworks for financial modeling were well known. It was the advancement of computing power and refinements of the theories to take into account real-world market imperfections and

conventions starting in the 1980s that facilitated implementation and broader acceptance of mathematical modeling of financial decisions. The diffusion of low-cost high-performance computers has allowed the broad use of numerical methods, the landscape of financial modeling. The importance of finding closed-form solutions and the consequent search for simple models has been dramatically reduced. Computationally intensive methods such as Monte Carlo simulations and the numerical solution of differential equations are now widely used. As a consequence, it has become feasible to represent prices and returns with relatively complex models. Nonnormal probability distributions have become commonplace in many sectors of financial modeling. It is fair to say that the key limitation of financial modeling is now the size of available data samples or training sets, not the computations; it is the data that limit the complexity of estimates. Mathematical modeling has also undergone major changes. Techniques such as equivalent martingale methods are being used in derivative pricing, and cointegration, the theory of fat-tailed processes, and state-space modeling (including ARCH/GARCH and stochastic volatility models) are being used in financial modeling.

Financial Statement Analysis

Much of the financial data that are used in constructing financial models for forecasting and valuation purposes draw from the financial statements that companies are required to provide to investors. The four basic financial statements are the balance sheet, the income statement, the statement of cash flows, and the statement of shareholders' equity. It is important to understand these data so that the information conveyed by them is interpreted properly in financial modeling. The financial statements are created using several assumptions that affect how to use and interpret the financial data. The analysis of financial statements involves the selection, evaluation, and

interpretation of financial data and other pertinent information to assist in evaluating the operating performance and financial condition of a company. The operating performance of a company is a measure of how well a company has used its resources—its assets, both tangible and intangible—to produce a return on its investment. The financial condition of a company is a measure of its ability to satisfy its obligations, such as the payment of interest on its debt in a timely manner. There are many tools available in the analysis of financial information. These tools include financial ratio analysis and cash flow analysis. Cash flows are essential ingredients in valuation. Therefore, understanding past and current cash flows may help in forecasting future cash flows and, hence, determine the value of the company. Moreover, understanding cash flow allows the assessment of the ability of a firm to maintain current dividends and its current capital expenditure policy without relying on external financing. Financial modelers must understand how to use these financial ratios and cash flow information in the most effective manner in building models.

Finite Mathematics and Basic Functions for Financial Modeling

The collection of mathematical tools that does not include calculus is often referred to as “finite mathematics.” This includes matrix algebra, probability theory, and statistical analysis. Ordinary algebra deals with operations such as addition and multiplication performed on individual numbers. In financial modeling, it is useful to consider operations performed on ordered arrays of numbers. Ordered arrays of numbers are called vectors and matrices while individual numbers are called scalars. Probability theory is the mathematical approach to formalize the uncertainty of events. Even though a decision maker may not know which one of the set of possible events may finally occur, with probability theory a decision maker has the means of providing each event with

a certain probability. Furthermore, it provides the decision maker with the axioms to compute the probability of a composed event in a unique way. The rather formal environment of probability theory translates in a reasonable manner to the problems related to risk and uncertainty in finance such as, for example, the future price of a financial asset. Today, investors may be aware of the price of a certain asset, but they cannot say for sure what value it might have tomorrow. To make a prudent decision, investors need to assess the possible scenarios for tomorrow’s price and assign to each scenario a probability of occurrence. Only then can investors reasonably determine whether the financial asset satisfies an investment objective included within a portfolio. Probability models are theoretical models of the occurrence of uncertain events. In contrast, statistics is about empirical data and can be broadly defined as a set of methods used to make inferences from a known sample to a larger population that is in general unknown. In finance, a particular important example is making inferences from the past (the known sample) to the future (the unknown population). There are important mathematical functions with which the financial modeler should be acquainted. These include the continuous function, the indicator function, the derivative of a function, the monotonic function, and the integral, as well as special functions such as the characteristic function of random variables and the factorial, the gamma, beta, and Bessel functions.

Liquidity and Trading Costs

In broad terms, liquidity refers to the ability to execute a trade or liquidate a position with little or no cost or inconvenience. Liquidity depends on the market where a financial instrument is traded, the type of position traded, and sometimes the size and trading strategy of an individual trade. Liquidity risks are those associated with the prospect of imperfect market liquidity and can relate to risk of loss or

risk to cash flows. There are two main aspects to liquidity risk measurement: the measurement of liquidity-adjusted measures of market risk and the measurement of liquidity risks per se. Market practitioners often assume that markets are liquid—that is, that they can liquidate or unwind positions at going market prices—usually taken to be the mean of bid and ask prices—without too much difficulty or cost. This assumption is very convenient and provides a justification for the practice of marking positions to market prices. However, it is often empirically questionable, and the failure to allow for liquidity can undermine the measurement of market risk. Because liquidity risk is a major risk factor in its own right, portfolio managers and traders will need to measure this risk in order to formulate effective portfolio and trading strategies. A considerable amount of work has been done in the equity market in estimating liquidity risk. Because transaction costs are incurred when buying or selling stocks, poorly executed trades can adversely impact portfolio returns and therefore relative performance. Transaction costs are classified as explicit costs such as brokerage and taxes, and implicit costs, which include market impact cost, price movement risk, and opportunity cost. Broadly speaking, market impact cost is the price that a trader has to pay for obtaining liquidity in the market and is a key component of trading costs that must be modeled so that effective trading programs for executing trades can be developed. Typical forecasting models for market impact costs are based on a statistical factor approach where the independent variables are trade-based factors or asset-based factors.

VOLUME III

Model Risk and Selection

Model risk is the risk of error in pricing or risk-forecasting models. In practice, model risk arises because (1) any model involves simpli-

fication and calibration, and both of these require subjective judgments that are prone to error, and/or (2) a model is used inappropriately. Although model risk cannot be avoided, there are many ways in which financial modelers can manage this risk. These include (1) recognizing model risk, (2) identifying, evaluating, and checking the model's key assumption, (3) selecting the simplest reasonable model, (4) resisting the temptation to ignore small discrepancies in results, (5) testing the model against known problems, (6) plotting results and employing nonparametric statistics, (7) back-testing and stress-testing the model, (8) estimating model risk quantitatively, and (9) reevaluating models periodically. In financial modeling, model selection requires a blend of theory, creativity, and machine learning. The machine-learning approach starts with a set of empirical data that the financial modeler wants to explain. Data are explained by a family of models that include an unbounded number of parameters and are able to fit data with arbitrary precision. There is a trade-off between model complexity and the size of the data sample. To implement this trade-off, ensuring that models have forecasting power, the fitting of sample data is constrained to avoid fitting noise. Constraints are embodied in criteria such as the Akaike information criterion or the Bayesian information criterion. Economic and financial data are generally scarce given the complexity of their patterns. This scarcity introduces uncertainty as regards statistical estimates obtained by the financial modeler. It means that the data might be compatible with many different models with the same level of statistical confidence. Methods of probabilistic decision theory can be used to deal with model risk due to uncertainty regarding the model's parameters. Probabilistic decision making starts from the Bayesian inference process and involves computer simulations in all realistic situations. Since a risk model is typically a combination of a probability distribution model and a risk measure, a critical assumption is the probability distribution assumed for

the random variable of interest. Too often, the Gaussian distribution is the model of choice. Empirical evidence supports the use of probability distributions that exhibit fat tails such as the Student's t distribution and its asymmetric version and the Pareto stable class of distributions and their tempered extensions. Extreme value theory offers another approach for risk modeling.

Mortgage-Backed Securities Analysis and Valuation

Mortgage-backed securities are fixed-income securities backed by a pool of mortgage loans. Residential mortgage-backed securities (RMBS) are backed by a pool of residential mortgage loans (one-to-four family dwellings). The RMBS market includes agency RMBS and nonagency RMBS. The former are securities issued by the Government National Mortgage Association (Ginnie Mae), Fannie Mae, and Freddie Mac. Agency RMBS include passthrough securities, collateralized mortgage obligations, and stripped mortgage-backed securities (interest-only and principal-only securities). The valuation of RMBS is complicated due to prepayment risk, a form of call risk. In contrast, nonagency RMBS are issued by private entities, have no implicit or explicit government guarantee, and therefore require one or more forms of credit enhancement in order to be assigned a credit rating. The analysis of nonagency RMBS must take into account both prepayment risk and credit risk. The most commonly used method for valuing RMBS is the Monte Carlo method, although other methods have garnered favor, in particular the decomposition method. The analysis of RMBS requires an understanding of the factors that impact prepayments.

Operational Risk

Operational risk has been regarded as a mere part of a financial institution's "other" risks. However, failures of major financial entities

have made regulators and investors aware of the importance of this risk. In general terms, operational risk is the risk of loss resulting from inadequate or failed internal processes, people, or systems or from external events. This risk encompasses legal risks, which includes, but is not limited to, exposure to fines, penalties, or punitive damages resulting from supervisory actions, as well as private settlements. Operational risk can be classified according to several principles: nature of the loss (internally inflicted or externally inflicted), direct losses or indirect losses, degree of expectancy (expected or unexpected), risk type, event type or loss type, and by the magnitude (or severity) of loss and the frequency of loss. Operational risk can be the cause of reputational risk, a risk that can occur when the market reaction to an operational loss event results in reduction in the market value of a financial institution that is greater than the amount of the initial loss. The two principal approaches in modeling operational loss distributions are the nonparametric approach and the parametric approach. It is important to employ a model that captures tail events, and for this reason in operational risk modeling, distributions that are characterized as light-tailed distributions should be used with caution. The models that have been proposed for assessing operational risk can be broadly classified into top-down models and bottom-up models. Top-down models quantify operational risk without attempting to identify the events or causes of losses. Bottom-up models quantify operational risk on a micro level, being based on identified internal events. The obstacle hindering the implementation of these models is the scarcity of available historical operational loss data.

Optimization Tools

Optimization is an area in applied mathematics that, most generally, deals with efficient algorithms for finding an optimal solution among a set of solutions that satisfy given constraints. Mathematical programming, a management

science tool that uses mathematical optimization models to assist in decision making, includes linear programming, integer programming, mixed-integer programming, nonlinear programming, stochastic programming, and goal programming. Unlike other mathematical tools that are available to decision makers such as statistical models (which tell the decision maker what occurred in the past), forecasting models (which tell the decision maker what might happen in the future), and simulation models (which tell the decision maker what will happen under different conditions), mathematical programming models allow the decision maker to identify the “best” solution. Markowitz’s mean-variance model for portfolio selection is an example of an application of one type of mathematical programming (quadratic programming). Traditional optimization modeling assumes that the inputs to the algorithms are certain, but there are also branches of optimization such as robust optimization that study the optimal decision under uncertainty about the parameters of the problem. Stochastic programming deals with both the uncertainty about the parameters and a multiperiod decision-making framework.

Probability Distributions

In financial models where the outcome of interest is a random variable, an assumption must be made about the random variable’s probability distribution. There are two types of probability distributions: discrete and continuous. Discrete probability distributions are needed whenever the random variable is to describe a quantity that can assume values from a countable set, either finite or infinite. A discrete probability distribution (or law) is quite intuitive in that it assigns certain values, positive probabilities, adding up to one, while any other value automatically has zero probability. Continuous probability distributions are needed when the random variable of interest can assume any value inside of one or more

intervals of real numbers such as, for example, any number greater than zero. Asset returns, for example, whether measured monthly, weekly, daily, or at an even higher frequency are commonly modeled as continuous random variables. In contrast to discrete probability distributions that assign positive probability to certain discrete values, continuous probability distributions assign zero probability to any single real number. Instead, only entire intervals of real numbers can have positive probability such as, for example, the event that some asset return is not negative. For each continuous probability distribution, this necessitates the so-called probability density, a function that determines how the entire probability mass of one is distributed. The density often serves as the proxy for the respective probability distribution. To model the behavior of certain financial assets in a stochastic environment, a financial modeler can usually resort to a variety of theoretical distributions. Most commonly, probability distributions are selected that are analytically well known. For example, the normal distribution (a continuous distribution)—also called the Gaussian distribution—is often the distribution of choice when asset returns are modeled. Or the exponential distribution is applied to characterize the randomness of the time between two successive defaults of firms in a bond portfolio. Many other distributions are related to them or built on them in a well-known manner. These distributions often display pleasant features such as stability under summation—meaning that the return of a portfolio of assets whose returns follow a certain distribution again follows the same distribution. However, one has to be careful using these distributions since their advantage of mathematical tractability is often outweighed by the fact that the stochastic behavior of the true asset returns is not well captured by these distributions. For example, although the normal distribution generally renders modeling easy because all moments of the distribution exist, it fails to reflect stylized facts commonly encountered in

asset returns—namely, the possibility of very extreme movements and skewness. To remedy this shortcoming, probability distributions accounting for such extreme price changes have become increasingly popular. Some of these distributions concentrate exclusively on the extreme values while others permit any real number, but in a way capable of reflecting market behavior. Consequently, a financial modeler has available a great selection of probability distributions to realistically reproduce asset price changes. Their common shortcoming is generally that they are mathematically difficult to handle.

Risk Measures

The standard assumption in financial models is that the distribution for the return on financial assets follows a normal (or Gaussian) distribution and therefore the standard deviation (or variance) is an appropriate measure of risk in the portfolio selection process. This is the risk measure that is used in the well-known Markowitz portfolio selection model (that is, mean-variance model), which is the foundation for modern portfolio theory. Mounting evidence since the early 1960s strongly suggests that return distributions do not follow a normal distribution, but instead exhibit heavy tails and, possibly, skewness. The “tails” of the distribution are where the extreme values occur, and these extreme values are more likely than would be predicted by the normal distribution. This means that between periods where the market exhibits relatively modest changes in prices and returns, there will be periods where there are changes that are much higher (that is, crashes and booms) than predicted by the normal distribution. This is of major concern to financial modelers in seeking to generate probability estimates for financial risk assessment. To more effectively implement portfolio selection, researchers have proposed alternative risk measures. These risk measures fall into

two disjointed categories: dispersion measures and safety-first measures. Dispersion measures include mean standard deviation, mean absolute deviation, mean absolute moment, index of dissimilarity, mean entropy, and mean colog. Safety-first risk measures include classical safety first, value-at-risk, average value-at-risk, expected tail loss, MiniMax, lower partial moment, downside risk, probability-weighted function of deviations below a specified target return, and power conditional value-at-risk. Despite these alternative risk measures, the most popular risk measure used in financial modeling is volatility as measured by the standard deviation. There are different types of volatility: historical, implied volatility, level-dependent volatility, local volatility, and stochastic volatility (e.g., jump-diffusion volatility). There are risk measures commonly used for bond portfolio management. These measures include duration, convexity, key rate duration, and spread duration.

Software for Financial Modeling

The development of financial models requires the modeler to be familiar with spreadsheets such as Microsoft Excel and/or a platform to implement concepts and algorithms such as the Palisade Decision Tools Suite and other Excel-based software (mostly @RISK1, Solver2, VBA3), and MATLAB. Financial modelers can choose one or the other, depending on their level of familiarity and comfort with spreadsheet programs and their add-ins versus programming environments such as MATLAB. Some tasks and implementations are easier in one environment than in the other. MATLAB is a modeling environment that allows for input and output processing, statistical analysis, simulation, and other types of model building for the purpose of analysis of a situation. MATLAB uses a number-array-oriented programming language, that is, a programming language in which vectors and matrices

are the basic data structures. Reliable built-in functions, a wide range of specialized toolboxes, easy interface with widespread software like Microsoft Excel, and beautiful graphing capabilities for data visualization make implementation with MATLAB efficient and useful for the financial modeler. Visual Basic for Applications (VBA) is a programming language environment that allows Microsoft Excel users to automate tasks, create their own functions, perform complex calculations, and interact with spreadsheets. VBA shares many of the same concepts as object-oriented programming languages. Despite some important limitations, VBA does add useful capabilities to spreadsheet modeling, and it is a good tool to know because Excel is the platform of choice for many finance professionals.

Stochastic Processes and Tools

Stochastic integration provides a coherent way to represent that instantaneous uncertainty (or volatility) cumulates over time. It is thus fundamental to the representation of financial processes such as interest rates, security prices, or cash flows. Stochastic integration operates on stochastic processes and produces random variables or other stochastic processes. Stochastic integration is a process defined on each path as the limit of a sum. However, these sums are different from the sums of the Riemann-Lebesgue integrals because the paths of stochastic processes are generally not of bounded variation. Stochastic integrals in the sense of Itô are defined through a process of approximation by (1) defining Brownian motion, which is the continuous limit of a random walk, (2) defining stochastic integrals for elementary functions as the sums of the products of the elementary functions multiplied by the increments of the Brownian motion, and (3) extending this definition to any function through approximating sequences. The major application of integration to financial modeling involves stochastic

integrals. An understanding of stochastic integrals is needed to understand an important tool in contingent claims valuation: stochastic differential equations. The dynamic of financial asset returns and prices can be expressed using a deterministic process if there is no uncertainty about its future behavior, or, with a stochastic process, in the more likely case when the value is uncertain. Stochastic processes in continuous time are the most used tool to explain the dynamic of financial assets returns and prices. They are the building blocks to construct financial models for portfolio optimization, derivatives pricing, and risk management. Continuous-time processes allow for more elegant theoretical modeling compared to discrete time models, and many results proven in probability theory can be applied to obtain a simple evaluation method.

Statistics

Probability models are theoretical models of the occurrence of uncertain events. In contrast, statistics is about empirical data and can be broadly defined as a set of methods used to make inferences from a known sample to a larger population that is in general unknown. In finance, a particular important example is making inferences from the past (the known sample) to the future (the unknown population). In statistics, probabilistic models are applied using data so as to estimate the parameters of these models. It is not assumed that all parameter values in the model are known. Instead, the data for the variables in the model to estimate the value of the parameters are used and then applied to test hypotheses or make inferences about their estimated values. In financial modeling, the statistical technique of regression models is the workhorse. However, because regression models are part of the field of financial econometrics, this topic is covered in that topic category. Understanding dependences or functional links between variables is a key theme in

financial modeling. In general terms, functional dependencies are represented by dynamic models. Many important models are linear models whose coefficients are correlation coefficients. In many instances in financial modeling, it is important to arrive at a quantitative measure of the strength of dependencies. The correlation coefficient provides such a measure. In many instances, however, the correlation coefficient might be misleading. In particular, there are cases of nonlinear dependencies that result in a zero correlation coefficient. From the point of view of financial modeling, this situation is particularly dangerous as it leads to substantially underestimated risk. Different measures of dependence have been proposed, in particular copula functions. The copula overcomes the drawbacks of the correlation as a measure of dependency by allowing for a more general measure than linear dependence, allowing for the modeling of dependence for extreme events, and being indifferent to continuously increasing transformations. Another essential tool in financial modeling, because it allows the incorporation of uncertainty in financial models and consideration of additional layers of complexity that are difficult to incorporate in analytical models, is Monte Carlo simulation. The main idea of Monte Carlo simulation is to represent the uncertainty in market variables through scenarios, and to evaluate parameters of interest that depend on these market variables in complex ways. The advantage of such an approach is that it can easily capture the dynamics of underlying processes and the otherwise complex effects of interactions among market variables. A substantial amount of research in recent years has been dedicated to making scenario generation more accurate and efficient, and a number of sophisticated computational techniques are now available to the financial modeler.

Term Structure Modeling

The arbitrage-free valuation approach to the valuation of option-free bonds, bonds with em-

bedded options, and option-type derivative instruments requires that a financial instrument be viewed as a package of zero-coupon bonds. Consequently, in financial modeling, it is essential to be able to discount each expected cash flow by the appropriate interest rate. That rate is referred to as the spot rate. The term structure of interest rates provides the relationship between spot rates and maturity. Because of its role in valuation of cash bonds and option-type derivatives, the estimation of the term structure of interest rates is of critical importance as an input into a financial model. In addition to its role in valuation modeling, term structure models are fundamental to expressing value, risk, and establishing relative value across the spectrum of instruments found in the various interest-rate or bond markets. The term structure is most often specified for a specific market such as the U.S. Treasury market, the bond market for double-A rated financial institutions, the interest rate market for LIBOR, and swaps. Static models of the term structure are characterizations that are devoted to relationships based on a given market and do not serve future scenarios where there is uncertainty. Standard static models include those known as the spot yield curve, discount function, par yield curve, and the implied forward curve. Instantiations of these models may be found in both a discrete- and continuous-time framework. An important consideration is establishing how these term structure models are constructed and how to transform one model into another. In modeling the behavior of interest rates, stochastic differential equations (SDEs) are commonly used. The SDEs used to model interest rates must capture the market properties of interest rates such as mean reversion and/or a volatility that depends on the level of interest rates. For a one-factor model, the SDE is used to model the behavior of the short-term rate, referred to as simply the "short rate." The addition of another factor (i.e., a two-factor model) involves extending the SDE to represent the behavior of the short rate and a long-term rate (i.e., long rate).

The entries can serve as material for a wide spectrum of courses, such as the following:

- Financial engineering
- Financial mathematics
- Financial econometrics
- Statistics with applications in finance
- Quantitative asset management
- Asset and derivative pricing
- Risk management

Frank J. Fabozzi
Editor, *Encyclopedia of Financial Models*

Guide to the *Encyclopedia of Financial Models*

The *Encyclopedia of Financial Models* provides comprehensive coverage of the field of financial modeling. This reference work consists of three separate volumes and 127 entries. Each entry provides coverage of the selected topic intended to inform a broad spectrum of readers ranging from finance professionals to academicians to students to fiduciaries. To derive the greatest possible benefit from the *Encyclopedia of Financial Models*, we have provided this guide. It explains how the information within the encyclopedia can be located.

ORGANIZATION

The *Encyclopedia of Financial Models* is organized to provide maximum ease of use for its readers.

Table of Contents

A complete table of contents for the entire encyclopedia appears in the front of each volume. This list of titles represents topics that have been carefully selected by the editor, Frank J. Fabozzi. The Preface includes a more detailed description of the volumes and the topic categories that the entries are grouped under.

Index

A Subject Index for the entire encyclopedia is located at the end of each volume. The sub-

jects in the index are listed alphabetically and indicate the volume and page number where information on this topic can be found.

Entries

Each entry in the *Encyclopedia of Financial Models* begins on a new page, so that the reader may quickly locate it. The author's name and affiliation are displayed at the beginning of the entry. All entries in the encyclopedia are organized according to a standard format, as follows:

- Title and author
- Abstract
- Introduction
- Body
- Key points
- Notes
- References

Abstract

The abstract for each entry gives an overview of the topic, but not necessarily the content of the entry. This is designed to put the topic in the context of the entire *Encyclopedia*, rather than give an overview of the specific entry content.

Introduction

The text of each entry begins with an introductory section that defines the topic under

discussion and summarizes the content. By reading this section, the reader gets a general idea about the content of a specific entry.

Body

The body of each entry explains the purpose, theory, and math behind each model.

Key Points

The key points section provides in bullet point format a review of the materials discussed in

each entry. It imparts to the reader the most important issues and concepts discussed.

Notes

The notes provide more detailed information and citations of further readings.

References

The references section lists the publications cited in the entry.

ENCYCLOPEDIA
OF
FINANCIAL MODELS

Volume II

Equity Models and Valuation

Dividend Discount Models

PAMELA P. DRAKE, PhD, CFA

J. Gray Ferguson Professor of Finance, College of Business, James Madison University

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Dividends are cash payments made by a corporation to its owners. Though cash dividends are paid to both preferred and common shareholders, most of the focus of the attention is on the dividends paid to the residual owners of the corporation, the common shareholders. Dividends paid to common and preferred shareholders are not legal obligations of a corporation, and some corporations do not pay cash dividends. But for those companies that pay dividends, changes in dividends are noticed by investors—increases in dividends are viewed favorably and are associated with increases in the company’s stock price, whereas decreases in dividends are viewed quite unfavorably and are associated with decreases in the company’s stock price. Most models that use dividends in the estimation of stock value use current dividends, some measure of historical or projected dividend growth, and an estimate of the required rate of return. Popular models include the basic dividend discount model that assumes a constant dividend growth, and the multiple-phase models, which include the two-stage dividend growth model and the stochastic dividend discount models.

In this entry, we discuss dividend discount models and their limitations. We begin with a review of the various ways to measure dividends and then take a look at how dividends and stock prices are related.

DIVIDEND MEASURES

Dividends are measured using three different measures:

- Dividends per share
- Dividend yield
- Dividend payout

The value of a share of stock today is the investors’ assessment of today’s worth of future cash flows for each share. Because future cash flows to shareholders are dividends, we need a measure of dividends for each share of stock to estimate future cash flows per share. The *dividends per share* is the dollar amount of dividends paid out during the period per share of common stock:

$$\begin{aligned} &\text{Dividends per share} \\ &= \frac{\text{Dividends}}{\text{Number of shares outstanding}} \end{aligned}$$

If a company has paid \$600,000 in dividends during the period and there are 1.5 million shares of common stock outstanding, then

$$\begin{aligned}\text{Dividends per share} &= \frac{\$600,000}{1,500,000 \text{ shares}} \\ &= \$0.40 \text{ per share}\end{aligned}$$

The company paid out 40 cents in dividends per common share during this period.

The *dividend yield*, the ratio of dividends to price, is

$$\begin{aligned}\text{Dividend yield} \\ &= \frac{\text{Annual cash dividends per share}}{\text{Market price per share}}\end{aligned}$$

The dividend yield is also referred to as the dividend-price ratio. Historically, the dividend yield for U.S. stocks has been a little less than 5%, according to a study by Campbell and Shiller (1998). In an exhaustive study of the relation between dividend yield and stock prices, Campbell and Shiller find that:

- There is a weak relation between the dividend yield and subsequent 10-year dividend growth.
- The dividend yield does not forecast future dividend growth.
- The dividend yield predicts future price changes.

The weak relation between the dividend yield and future dividends may be attributed to the effects of the business cycle on dividend growth. The tendency for the dividend yield to revert to its historical mean has been observed by researchers.

Another way of describing dividends paid out during a period is to state the dividends as a portion of earnings for the period. This is referred to as the *dividend payout ratio*:

$$\begin{aligned}\text{Dividend payout ratio} \\ &= \frac{\text{Dividends}}{\text{Earnings available to common shareholders}}\end{aligned}$$

If a company pays \$360,000 in dividends and has earnings available to common shareholders

of \$1.2 million, the payout ratio is 30%:

$$\begin{aligned}\text{Dividend payout ratio} &= \frac{\$360,000}{\$1,200,000} \\ &= 0.30 \text{ or } 30\%\end{aligned}$$

This means that the company paid out 30% of its earnings to shareholders.

The proportion of earnings paid out in dividends varies by company and industry. For example, the companies in the steel industry typically pay out 25% of their earnings in dividends, whereas the electric utility companies pay out approximately 75% of their earnings in dividends.

If companies focus on dividends per share in establishing their dividends (e.g., a constant dividends per share), the dividend payout will fluctuate along with earnings. We generally observe that companies set the dividend policy such that dividends per share grow at a relatively constant rate, resulting in dividend payouts that fluctuate.

DIVIDENDS AND STOCK PRICES

If an investor buys a common stock, he or she has bought shares that represent an ownership interest in the corporation. Shares of common perpetual security—there is no maturity. The investor who owns shares of common stock has the right to receive a certain portion of any dividends—but dividends are not a sure thing. Whether or not a corporation pays dividends is up to its board of directors—the representatives of the common shareholders. Typically, we see some pattern in the dividends companies pay: Dividends are either constant or grow at a constant rate. But there is no guarantee that dividends will be paid in the future.

Preferred shareholders are in a similar situation as the common shareholders. They expect to receive cash dividends in the future, but the payment of these dividends is up to the board of directors. But there are three major differences

between the dividends of preferred and common shares. First, the dividends on preferred stock usually are specified at a fixed rate or dollar amount, whereas the amount of dividends is not specified for common shares. Second, preferred shareholders are given preference: their dividends must be paid before any dividends are paid on common stock. Third, if the preferred stock has a cumulative feature, dividends not paid in one period accumulate and are carried over to the next period. Therefore, the dividends on preferred stock are more certain than those on common shares.

It is reasonable to figure that what an investor pays for a share of stock should reflect what he or she expects to receive from it—return on the investor’s investment. What an investor receives are cash dividends in the future. How can we relate that return to what a share of common stock is worth? Well, the value of a share of stock should be equal to the present value of all the future cash flows an investor expects to receive from that share. To value stock, therefore, an investor must project future cash flows, which, in turn, means projecting future dividends. This approach to the valuation of common stock is referred to as the discounted cash flow approach and the models used are referred to as *dividend discount models*.

Dividend discount models are not the only approach to valuing common stock. There are fundamental factor models, also referred to as multifactor equity models.

BASIC DIVIDEND DISCOUNT MODELS

As discussed above, the basis for the dividend discount model (DDM) is simply the application of present value analysis, which asserts that the fair price of an asset is the present value of the expected cash flows. This model was first suggested by Williams (1938). In the case of common stock, the cash flows are the expected

dividend payouts. The basic DDM model can be expressed mathematically as:

$$P = \frac{D_1}{(1 + r_1)^1} + \frac{D_2}{(1 + r_2)^2} + \dots \quad (1)$$

where

P = the fair value or theoretical value of the common stock

D_t = the expected dividend for period t

r_t = the appropriate discount or capitalization rate for period t

The dividends are expected to be received forever.

Practitioners rarely use the dividend discount model given by equation (1). Instead, one of the DDMs discussed below is typically used.

THE FINITE LIFE GENERAL DIVIDEND DISCOUNT MODEL

The DDM given by equation (1) can be modified by assuming a finite life for the expected cash flows. In this case, the expected cash flows are the expected dividend payouts and the expected sale price of the stock at some future date. The expected sale price is also called the terminal price and is intended to capture the future value of all subsequent dividend payouts. This model is called the *finite life general DDM* and is expressed mathematically as:

$$P = \frac{D_1}{(1 + r_1)^1} + \frac{D_2}{(1 + r_2)^2} + \dots + \frac{D_N}{(1 + r_N)^N} + \frac{P_N}{(1 + r_N)^N} \quad (2)$$

where

P_N = the expected sale price (or terminal price) at the horizon period N

N = the number of periods in the horizon

and P , D_t , and r_t are the same as defined above.

Assuming a Constant Discount Rate

A special case of the finite life general DDM that is more commonly used in practice is one in which it is assumed that the discount rate is constant. That is, it is assumed each r_t is the same for all t . Denoting this constant discount rate by r , equation (2) becomes:

$$P = \frac{D_1}{(1+r)^1} + \frac{D_2}{(1+r)^2} + \cdots + \frac{D_N}{(1+r)^N} + \frac{P_N}{(1+r)^N} \quad (3)$$

Equation (3) is called the constant discount rate version of the finite life general DDM. When practitioners use any of the DDM models presented in this entry, typically the constant discount rate version form is used.

Let's illustrate the finite life general DDM assuming a constant discount rate assuming each period is a year. Suppose that the following data are determined for stock XYZ by a financial analyst:

$$\begin{aligned} D_1 &= \$2.00 & D_2 &= \$2.20 & D_3 &= \$2.30 \\ D_4 &= \$2.55 & D_5 &= \$2.65 \\ P_5 &= \$26 & N &= 5 & r &= 0.10 \end{aligned}$$

Based on these data, the fair price of stock XYZ is

$$\begin{aligned} P &= \frac{\$2.00}{(1.10)^1} + \frac{\$2.20}{(1.10)^2} + \frac{\$2.30}{(1.10)^3} + \frac{\$2.55}{(1.10)^4} \\ &\quad + \frac{\$2.65}{(1.10)^5} + \frac{\$26.00}{(1.10)^5} = \$24.895 \end{aligned}$$

Required Inputs

The finite life general DDM requires three forecasts as inputs to calculate the fair value of a stock:

1. The expected terminal price (P_N)
2. The dividends up to the assumed horizon (D_1 to D_N)
3. The discount rates (r_1 to r_N) or r (in the case of the constant discount rate version)

Thus the relevant question is, How accurately can these inputs be forecasted?

The terminal price is the most difficult of the three forecasts. According to theory, P_N is the present value of all future dividends after N ; that is, D_{N+1} , D_{N+2} , \dots , D_{infinity} . Also, the future discount rate (r_t) must be forecasted. In practice, forecasts are made of either dividends (D_N) or earnings (E_N) first, and then the price P_N is estimated by assigning an "appropriate" requirement for yield, price-earnings ratio, or capitalization rate. Note that the present value of the expected terminal price $P_N/(1+r)^N$ becomes very small if N is very large.

The forecasting of dividends is "somewhat" easier. Usually, past history is available, management can be queried, and cash flows can be projected for a given scenario. The discount rate r is the required rate of return. Forecasting r is more complex than forecasting dividends, although not nearly as difficult as forecasting the terminal price (which requires a forecast of future discount rates as well). As noted above, in practice for a given company r is assumed to be constant for all periods and typically generated from the capital asset pricing model (CAPM). The CAPM provides the expected return for a company based on its systematic risk (beta).

Assessing Fair Value

Given the fair price derived from a dividend discount model, the assessment of the stock proceeds along the following lines. If the market price is below the fair price derived from the model, the stock is undervalued or cheap. The opposite holds for a stock whose market price is greater than the model-derived price. In this case, the stock is said to be overvalued or expensive. A stock trading equal to or close to its fair price is said to be fairly valued.

The DDM tells us the fair price but does not tell us when the price of the stock should be expected to move to this fair price. That is, the model says that based on the inputs generated by the analyst, the stock may be cheap, expensive, or priced appropriately. However, it does

not tell us that if it is mispriced how long it will take before the market recognizes the mispricing and corrects it. As a result, an investor may hold on to a stock perceived to be cheap for an extended period of time and may underperform a benchmark during that period.

While a stock may be mispriced, an investor must also consider how mispriced it is in order to take the appropriate action (buy a cheap stock and sell or sell short an expensive stock). This will depend on the degree of mispricing and transaction costs.

CONSTANT GROWTH DIVIDEND DISCOUNT MODEL

If future dividends are assumed to grow at a constant rate (g) and a single discount rate (r) is used, then the finite life general DDM assuming a constant growth rate given by equation (3) becomes

$$P = \frac{D_0(1+g)^1}{(1+r)^1} + \frac{D_0(1+g)^2}{(1+r)^2} + \frac{D_0(1+g)^3}{(1+r)^3} + \dots + \frac{D_0(1+g)^N}{(1+r)^N} + \frac{P_N}{(1+r)^N} \quad (4)$$

and it can be shown that if N is assumed to approach infinity, equation (4) is equal to:

$$P = \frac{D_0(1+g)}{r-g} \quad (5)$$

Equation (5) is the *constant growth dividend discount model* (Gordon and Shapiro, 1956). An equivalent formulation for the constant growth DDM is

$$P = \frac{D_1}{r-g} \quad (6)$$

where D_1 is equal to $D_0(1+g)$.

Consider a company that currently pays dividends of \$3.00 per share. If the dividend is expected to grow at a rate of 3% per year and the discount rate is 12%, what is the value of a share of stock of this company? Using equation (5),

$$P = \frac{\$3.00(1+0.03)}{0.12-0.03} = \frac{\$3.09}{0.09} = \$34.33$$

If the growth rate for this company's dividends is 5%, instead of 3%, the current value is \$45.00:

$$P = \frac{\$3.00(1+0.05)}{0.12-0.05} = \frac{\$3.15}{0.07} = \$45.00$$

Therefore, the greater the expected growth rate of dividends, the greater the value of a share of stock.

In this last example, if the discount rate is 14% instead of 12% and the growth rate of dividends is 3%, the value of a share of stock is:

$$P = \frac{\$3.00(1+0.03)}{0.14-0.03} = \frac{\$3.09}{0.11} = \$28.09$$

Therefore, the *greater* the discount rate, the *lower* the current value of a share of stock.

Let's apply the model as given by equation (5) to estimate the price of three companies: Eli Lilly, Schering-Plough, and Wyeth Laboratories. The discount rate for each company was estimated using the capital asset pricing model assuming (1) a market risk premium of 5% and (2) a risk-free rate of 4.63%. The market risk premium is based on the historical spread between the return on the market (often proxied with the return on the S&P 500 Index) and the risk-free rate. Historically, this spread has been approximately 5%. The risk-free rate is often estimated by the yield on U.S. Treasury securities. At the end of 2006, 10-year Treasury securities were yielding approximately 4.625%. We use 4.63% as an estimate for the purposes of this illustration. The beta estimate for each company was obtained from the Value Line Investment Survey: 0.9 for Eli Lilly, 1.0 for Schering-Plough and Wyeth. The discount rate, r , for each company based on the CAPM is:

Eli Lilly	$r = 0.0463 + 0.9(0.05) = 9.125\%$
Schering-Plough	$r = 0.0463 + 1.0(0.05) = 9.625\%$
Wyeth	$r = 0.0463 + 1.0(0.05) = 9.625\%$

The dividend growth rate can be estimated by using the compounded rate of growth of historical dividends.

The compound growth rate, g , is found using the following formula:

$$g = \left(\frac{\text{Last dividend}}{\text{Starting dividend}} \right)^{1/\text{no. of years}} - 1$$

This formula is equivalent to calculating the geometric mean of 1 plus the percentage change over the number of years. Using time value of money math, the 2006 dividend is the future value, the starting dividend is the present value, the number of years is the number of periods; solving for the interest rate produces the growth rate.

Substituting the values for the starting and ending dividend amounts and the number of periods into the formula, we get:

Company	1991 dividend	2006 dividend	Estimated annual growth rate
Eli-Lilly	\$0.50	\$1.60	8.063%
Schering-Plough	\$0.16	\$0.22	2.146%
Wyeth	\$0.60	\$1.01	3.533%

The value of D_0 , the estimate for g , and the discount rate r for each company are summarized below:

Company	Current dividend D_0	Estimated annual growth rate g	Required rate of return r
Eli-Lilly	\$1.60	8.063%	9.125%
Schering-Plough	\$0.22	2.146%	9.625%
Wyeth	\$1.01	3.533%	9.625%

Substituting these values into equation (5), we obtain:

$$\begin{aligned} \text{Eli Lilly estimated price} &= \frac{\$1.60(1 + 0.08063)}{0.09125 - 0.08063} \\ &= \frac{\$1.729}{0.0162} = \$162.80 \end{aligned}$$

Schering-Plough estimated price

$$\begin{aligned} &= \frac{\$0.22(1 + 0.02146)}{0.09625 - 0.02146} \\ &= \frac{\$0.225}{0.07479} = \$3.00 \end{aligned}$$

$$\begin{aligned} \text{Wyeth estimated price} &= \frac{\$1.01(1 + 0.03533)}{0.09625 - 0.03533} \\ &= \frac{\$1.046}{0.06092} = \$17.16 \end{aligned}$$

Comparing the estimated price with the actual price, we see that this model does not do a good job of pricing these stocks:

Company	Estimated price at the end of 2006	Actual price at the end of 2006
Eli Lilly	\$162.80	\$49.87
Schering-Plough	\$3.00	\$23.44
Wyeth	\$17.16	\$50.52

Notice that the constant growth DDM is considerably off the mark for all three companies. The reasons include: (1) the dividend growth pattern for none of the three companies appears to suggest a constant growth rate, and (2) the growth rate of dividends in recent years has been much slower than earlier years (and, in fact, negative for Schering-Plough after 2003), causing growth rates estimated from the long time periods to overstate future growth. And this pattern is not unique to these companies.

Another problem that arises in using the constant growth rate model is that the growth rate of dividends may exceed the discount rate, r . Consider the following three companies and their dividend growth over the 16-year period from 1991 through 2006, with the estimated required rates of return:

Company	1991 dividend	2006 dividend	Estimated growth rate g	Estimated required rate of return
Coca Cola	\$0.24	\$1.24	11.70%	7.625%
Hershey	\$0.24	\$1.03	10.198%	7.875%
Tootsie Roll	\$0.04	\$0.31	14.627%	8.625%

For these three companies, the growth rate of dividends over the prior 16 years is greater than the discount rate. If we substitute the D_0 (the 2006 dividends), the g , and the r into equation

(5), the estimated price at the end of 2006 is negative, which doesn't make sense. Therefore, there are some cases in which it is inappropriate to use the constant rate DDM.

The potential for misvaluation using the constant rate DDM is highlighted by Fogler (1988) in his illustration using ABC prior to its being taken over by Capital Cities in 1985. He estimated the value of ABC stock to be \$53.88, which was less than its market price at the time (of \$64) and less than the \$121 paid per share by Capital Cities.

MULTIPHASE DIVIDEND DISCOUNT MODELS

The assumption of constant growth is unrealistic and can even be misleading. Instead, most practitioners modify the constant growth DDM by assuming that companies will go through different growth phases. Within a given phase, dividends are assumed to grow at a constant rate. Molodovsky, May, and Chattiner (1965) were some of the pioneers in modifying the DDM to accommodate different growth rates.

Two-Stage Growth Model

The simplest form of multi-phase DDM is the two-stage growth model. A simple extension of equation (4) uses two different values of g . Referring to the first growth rate as g_1 and the second growth rate as g_2 and assuming that the first growth rate pertains to the next four years and the second growth rate refers to all years following, equation (4) can be modified as:

$$P = \frac{D_0(1+g_1)^1}{(1+r)^1} + \frac{D_0(1+g_1)^2}{(1+r)^2} + \frac{D_0(1+g_1)^3}{(1+r)^3} + \frac{D_0(1+g_1)^4}{(1+r)^4} + \frac{D_0(1+g_1)^5}{(1+r)^5} + \frac{D_0(1+g_1)^6}{(1+r)^6} + \dots$$

which simplifies to:

$$P = \frac{D_0(1+g_1)^1}{(1+r)^1} + \frac{D_0(1+g_1)^2}{(1+r)^2} + \frac{D_0(1+g_1)^3}{(1+r)^3} + \frac{D_0(1+g_1)^4}{(1+r)^4} + P_4$$

Because dividends following the fourth year are presumed to grow at a constant rate g_2 forever, the value of a share at the end of the fourth year (that is, P_4) is determined by using equation (5), substituting $D_0(1+g_1)^4$ for D_0 (because period 4 is the base period for the value at end of the fourth year) and g_2 for the constant rate g :

$$P = \frac{D_0(1+g_1)^1}{(1+r)^1} + \frac{D_0(1+g_1)^2}{(1+r)^2} + \frac{D_0(1+g_1)^3}{(1+r)^3} + \frac{D_0(1+g_1)^4}{(1+r)^4} + \left[\frac{1}{(1+r)^4} \left(\frac{D_0(1+g_1)^4(1+g_2)}{r-g_2} \right) \right] \quad (7)$$

Suppose a company's dividends are expected to grow at 4% rate for the next four years and then 8% thereafter. If the current dividend is \$2.00 and the discount rate is 12%,

$$P = \frac{\$2.08}{(1+0.12)^1} + \frac{\$2.16}{(1+0.12)^2} + \frac{\$2.25}{(1+0.12)^3} + \frac{\$2.34}{(1+0.12)^4} + \left[\frac{1}{(1+0.12)^4} \left(\frac{\$2.53}{0.12-0.08} \right) \right] = \$46.87$$

If this company's dividends are expected to grow at the rate of 4% forever, the value of a share is \$26.00; if this company's dividends are expected to grow at the rate of 8% forever, the value of a share is \$52.00. But because the growth rate of dividends is expected to increase from 4% to 8% in four years, the value of a share is between those two values, or \$46.87.

As you can see from this example, the basic valuation model can be modified to accommodate different patterns of expected dividend growth.

Three-Stage Growth Model

The most popular multiphase model employed by practitioners appears to be the three-stage DDM. (The formula for this model is derived in Sorensen and Williamson [1985].) This model assumes that all companies go through three phases, analogous to the concept of the product life cycle. In the growth phase, a company experiences rapid earnings growth as it produces new products and expands market share. In the transition phase the company's earnings begin

to mature and decelerate to the rate of growth of the economy as a whole. At this point, the company is in the maturity phase in which earnings continue to grow at the rate of the general economy.

Different companies are assumed to be at different phases in the three-phase model. An emerging growth company would have a longer growth phase than a more mature company. Some companies are considered to have higher initial growth rates and hence longer growth and transition phases. Other companies may be considered to have lower current growth rates and hence shorter growth and transition phases.

In the typical investment management organization, analysts supply the projected earnings, dividends, growth rates for earnings, and dividend and payout ratios using fundamental security analysis. The growth rate at maturity for the entire economy is applied to all companies. As a generalization, approximately 25% of the expected return from a company (projected by the DDM) comes from the growth phase, 25% from the transition phase, and 50% from the maturity phase. However, a company with high growth and low dividend payouts shifts the relative contribution toward the maturity phase, while a company with low growth and a high payout shifts the relative contribution toward the growth and transition phases.

STOCHASTIC DIVIDEND DISCOUNT MODELS

As we noted in our discussion and illustration of the constant growth DDM, an erratic dividend pattern such as that of Wyeth can lead to quite a difference between the estimated price and the actual price. In the case of the pharmaceutical companies, the estimated price overstated the actual price for Eli Lilly, but understated the price of Schering-Plough and Wyeth.

Hurley and Johnson (1998a, 1998b) have suggested a new family of valuation model. Their

model allows for a more realistic pattern of dividend payments. The basic model generates dividend payments based on a model that assumes that either the firm will increase dividends for the period by a constant amount or keep dividends the same. The model is referred to as a *stochastic DDM* because the dividend can increase or be constant based on some estimated probability of each possibility occurring. The dividend stream used in the stochastic DDM is called the stochastic dividend stream.

There are two versions of the stochastic DDM. One assumes that dividends either increase or decrease at a constant growth rate. This version is referred to as a binomial stochastic DDM because there are two possibilities for dividends. The second version is called a trinomial stochastic DDM because it allows for an increase in dividends, no change in dividends, and a cut in dividends. We discuss each version below.

Binomial Stochastic Model

For both the binomial and trinomial stochastic DDM, there are two versions of the model—the additive growth model and the geometric growth model. The former model assumes that dividend growth is additive rather than geometric. This means that dividends are assumed to grow by a constant dollar amount. So, for example, if dividends are \$2.00 today and the additive growth rate is assumed to be \$0.25 per year, then next year dividends will grow to \$2.25, in two years dividends will grow to \$2.50, and so on. The second model assumes a geometric rate of dividend growth. This is the same growth rate assumption used in the earlier DDMs presented in this entry.

Binomial Additive Stochastic Model

This formulation of the model is expressed as follows:

$$D_{t+1} = \begin{cases} D_t + C & \text{with probability } p \\ D_t & \text{with probability } 1 - p \end{cases} \quad \text{for } t = 1, 2, \dots$$

where

- D_t = dividend in period t
- D_{t+1} = dividend in period $t+1$
- C = dollar amount of the dividend increase
- p = probability that the dividend will increase

Hurley and Johnson (1998a) have shown that the theoretical value of the stock based on the additive stochastic DDM assuming a constant discount rate is equal to:

$$P = \frac{D_0}{r} + \left[\frac{1}{r} + \frac{1}{r^2} \right] Cp \tag{8}$$

For example, consider once again Wyeth. In the illustration of the constant growth model, we used D_0 of \$1.01 and a g of 3.533%. We estimate C by calculating the dollar increase in dividends for each year that had a dividend increase and then taking the average dollar dividend increase. The average of the increases is \$0.0373.

In the 15-year span 1991 through 2006, dividends increased 11 of the 14 year-to-year differences. Therefore, $p = 11/15 = 73.3333\%$. Substituting these values into equation (8), we find the estimated price to be:

$$P = \frac{\$1.01}{0.09625} + \left[\left(\frac{1}{0.09125} + \frac{1}{0.09125^2} \right) (\$0.03727) \left(\frac{11}{15} \right) \right]$$

$$P = \$10.49351 + [(118.336) (\$0.3727) (0.73333)]$$

$$P = \$10.49351 + \$3.23682 = \$13.73033$$

Applying this model to the other two pharmaceutical companies, we see that the model produces an estimated price that is closer to the actual price than the fair value based on the constant growth model:

Company	Actual price at the end of 2006	Estimated price at the end of 2006 using a constant growth model	Estimated price at the end of 2006 using the binomial additive stochastic model
Eli Lilly	\$49.87	\$162.79	\$29.94
Schering-Plough	\$23.44	\$3.00	\$11.04
Wyeth	\$50.52	\$17.16	\$13.73

Binomial Geometric Stochastic Model

Letting g be the growth rate of dividends, then the geometric dividend stream is

$$D_{t+1} = \begin{cases} D_t(1+g) & \text{with probability } p \\ D_t & \text{with probability } 1-p \end{cases} \text{ for } t = 1, 2, \dots$$

Hurley and Johnson (1998b) show that the price of the stock in this case is:

$$P = \frac{D_0(1+pg)}{r-pg} \tag{9}$$

Equation (9) is the binomial stochastic DDM assuming a geometric growth rate and a constant discount rate.

Trinomial Stochastic Models

The trinomial stochastic DDM allows for dividend cuts. Within the Hurley-Johnson stochastic DDM framework, Yao (1997) derived this model that allows for a cut in dividends. He notes that is not uncommon for a firm to cut dividends temporarily. In fact, an examination of the dividend record of the electric utilities industry as published in *Value Line Industry Review* found that in the aggregate firms cut dividends three times over a 15-year period.

Trinomial Additive Stochastic Model

The additive stochastic DDM can be extended to allow for dividend cuts as follow:

$$D_{t+1} = \begin{cases} D_t + C & \text{with probability } p_U \\ D_t - C & \text{with probability } p_D \\ D_t & \text{with probability } p_C \\ 1 - p_C = 1 - p_U - p_D \end{cases} \text{ for } t = 1, 2, \dots$$

where

- p_U = probability that the dividend will increase
- p_D = probability that the dividend will decrease
- p_C = probability that the dividend will be unchanged

The theoretical value of the stock based on the trinomial additive stochastic DDM then

becomes:

$$P = \frac{D_0}{r} + \left[\frac{1}{r} + \frac{1}{r^2} \right] C(p_U - p_D) \quad (10)$$

Notice that when p_D is zero (that is, there is no possibility for a cut in dividends), equation (10) reduces to equation (8).

Trinomial Geometric Stochastic Model

For the trinomial geometric stochastic DDM allowing for a possibility of cuts, we have:

$$D_{t+1} = \begin{cases} D_t(1+g) & \text{with probability } p_U \\ D_t(1-g) & \text{with probability } p_D \\ D_t & \text{with probability } 1 - p_C \\ & = 1 - p_U - p_D \end{cases} \quad \text{for } t = 1, 2, \dots$$

and the theoretical price is:

$$P = \frac{D_0[1 + (p_U + p_D)]}{r - (p_U - p_D)g} \quad (11)$$

Once again, substituting zero for p_D , equation (11) reduces to equation (9)—the binomial geometric stochastic DDM.

Applications of the Stochastic DDM

Yao (1997) applied the stochastic DDMs to five electric utility stocks that had regular dividends from 1979 to 1994 and found that the models fit the various utility stocks differently.

We see similar results in an updated example using five electric utilities, as shown in Table 1. For three of the five utilities, the binomial model provides an estimate closest to the actual stock price, whereas for the other two utilities, the trinomial model offers the closest estimate. In

none of the cases, however, did the constant dividend growth model offer the closest approximation to the actual stock price.

Advantages of the Stochastic DDM

The stochastic DDM developed by Hurley and Johnson is a powerful tool for the analyst because it allows the analyst to generate a probability distribution for a stock's value. The probability distribution can be used by an analyst to assess whether a stock is sufficiently mispriced to justify a buy or sell recommendation. For example, suppose that a three-phase DDM indicates that the value of a stock trading at \$35 is \$42. According to the model, the stock is underpriced and the analyst would recommend the purchase of this stock. However, the analyst cannot express his or her confidence as to the degree to which the stock is undervalued.

Hurley and Johnson show how the stochastic DDM can be used to overcome this limitation of traditional DDMs. An analyst can use the derived probability distribution from the stochastic DDM to assess the probability that the stock is undervalued. For example, an analyst may find from a probability distribution that the probability that the stock is greater than \$35 (the market price) is 90%.

To employ a stochastic DDM an analyst must be prepared to make subjective assumptions about the uncertain nature of future dividends. Monte Carlo simulation available on a spreadsheet (@RISK in Excel, for example) can then be used to generate the probability distribution.

Table 1 Fit of the Different Dividend Models Applied to Five Electric Utilities

Company	Consolidated Edison	Dominion Resources	FPL Group	PPL	TECO Energy
Actual stock price, end of 2006	\$45.82	\$40.73	\$52.98	\$34.89	\$16.46
Estimated stock price given the . . .					
Constant dividend growth model	\$33.57	\$19.36	\$22.14	\$16.54	\$7.46
Binomial stochastic dividend model	\$43.59	\$30.51	\$36.12	\$28.30	\$23.02
Trinomial stochastic dividend model	\$63.12	\$25.84	\$41.23	\$23.71	\$14.45

EXPECTED RETURNS AND DIVIDEND DISCOUNT MODELS

Thus far, we have seen how to calculate the fair price of a stock given the estimates of dividends, discount rates, terminal prices, and growth rates. The model-derived price is then compared to the actual price and the appropriate action is taken.

The analysis can be recast in terms of expected return. This is found by calculating the return that will make the present value of the expected cash flows equal to the actual price. Mathematically, this is expressed as follows:

$$P_A = \frac{D_1}{(1 + ER)^1} + \frac{D_2}{(1 + ER)^2} + \dots + \frac{D_N}{(1 + ER)^N} + \frac{P_N}{(1 + ER)^N} \quad (12)$$

where

P_A = actual price of the stock
 ER = expected return

The expected return (ER) in equation (12). For example, consider the following inputs used at the outset of this entry to illustrate the finite life general DDM as given by equation (3). For stock XYZ, the inputs assumed are:

$$D_1 = \$2.00 \quad D_2 = \$2.20 \quad D_3 = \$2.30 \\ D_4 = \$2.55 \quad D_5 = \$2.65 \quad P_5 = \$26 \quad N = 5$$

We calculated a fair price based on equation (3) to be \$24.90. Suppose that the actual price is \$25.89. Then the expected return is found by solving the following equation for ER :

$$\begin{aligned} \$25.89 = & \frac{\$2.00}{(1 + ER)} + \frac{\$2.20}{(1 + ER)^2} + \frac{\$2.30}{(1 + ER)^3} \\ & + \frac{\$2.55}{(1 + ER)^4} + \frac{\$2.65}{(1 + ER)^5} + \frac{\$26.00}{(1 + ER)^5} \end{aligned}$$

The expected return is 9%.

The expected return is the discount rate that equates the present value of the expected future cash flows with the present value of the stock. The higher the expected return—for a given set of future cash flows—the lower the

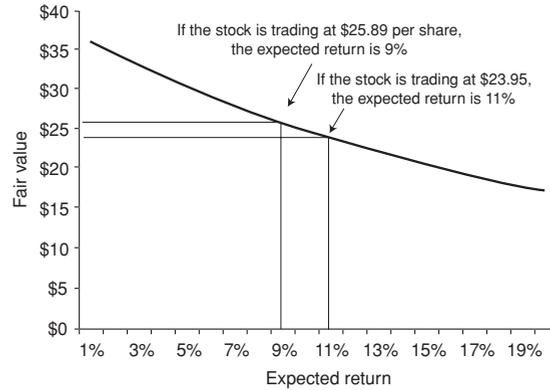


Figure 1 The Relation between the Fair Value of a Stock and the Stock’s Expected Return

current value. The relation between the fair value of a stock and the expected return of a stock is shown in Figure 1.

Given the expected return and the required return (that is, the value for r), any mispricing can be identified. If the expected return exceeds the required return, then the stock is undervalued; if it is less than the required return then the stock is overvalued. A stock is fairly valued if the expected return is equal to the required return. In our illustration, the expected return (9%) is less than the required return (10%); therefore, stock XYZ is overvalued.

With the same set of inputs, the identification of a stock being mispriced or fairly valued will be the same regardless of whether the fair value is determined and compared to the market price or the expected return is calculated and compared to the required return. In the case of XYZ stock, the fair value is \$24.90. If the stock is trading at \$25.89, it is overvalued. The expected return if the stock is trading at \$25.89 is 9%, which is less than the required return of 10%. If, instead, the stock price is \$24.90, it is fairly valued. The expected return can be shown to be 10%, which is the same as the required return. At a price of \$23.95, it can be shown that the expected return is 11%. Since the required return is 10%, stock XYZ would be undervalued.

While the illustration above uses the basic DDM, the expected return can be computed for

any of the models. In some cases, the calculation of the expected return is simple since a formula can be derived that specifies the expected return in terms of the other variables. For example, for the constant growth DDM given by equation (5), the expected return (r) can be easily solved to give:

$$r = \frac{D_1}{P} + g$$

Rearranging the constant growth model to solve for the expected return, we see that the required rate of return can be specified as the sum of the dividend yield and the expected growth rate of dividends.

KEY POINTS

- Dividends are measured in a number of ways, including dividends per share, dividend yield, and dividend payout.
- The discounted cash flow approach to valuing common stock requires projecting future dividends. Hence, the model used to value common stock is called a dividend discount model.
- The simplest dividend discount model is the constant growth model. More complex models include the *multiphase* model and stochastic models.
- Stock valuation using a dividend discount model is highly dependent on the inputs used.
- A dividend discount model does not indicate when the current market price will reach its fair value.
- The output of a dividend discount model is the fair price. However, the model can be used to generate the expected return.
- The expected return is the interest rate that will make the present value of the expected dividends plus terminal price equal to the stock's market price. The expected return is then compared to the required return to assess whether a stock is fairly priced in the market.

REFERENCES

- Campbell, J. Y., and Shiller, R. J. (1998). Valuation ratios and the long-run stock market outlook. *Journal of Portfolio Management* 24 (Winter): 11–26.
- Fogler, R. H. (1988). Security analysis, DDMs, and probability. In *Equity Markets and Valuation Methods* (pp. 51–52). Charlottesville, VA: The Institute of Chartered Financial Analysts.
- Gordon, M., and Shapiro, E. (1956). Capital equipment analysis: The required rate of profit. *Management Science* 3: 102–110.
- Hurley, W. J., and Johnson, L. (1994). A realistic dividend valuation model. *Financial Analysts Journal* 50 (July–August): 50–54.
- Hurley, W. J., and Johnson, L. (1998a). Generalized Markov dividend discount models. *Journal of Portfolio Management* 25 (Fall): 27–31.
- Hurley, W. J., and Johnson, L. (1998b). *The Theory and Application of Stochastic Dividend Models*. Monograph 7, Clarica Financial Services Research Centre, School of Business and Economics, Wilfrid Laurier University.
- Molodovsky, N., May, C., and Chattiner, S. (1965). Common stock valuation: Principles, tables, and applications. *Financial Analysts Journal* 21 (November–December): 111–117.
- Sorensen, E., and Williamson, E. (1985). Some evidence on the value of dividend discount models. *Financial Analysts Journal* 41 (November–December): 60–69.
- Williams, J. B. (1938). *The Theory of Investment Value*. Cambridge, MA: Harvard University Press.
- Yao, Y. (1997). A trinomial dividend valuation model. *Journal of Portfolio Management* 21 (Summer): 99–103.

Discounted Cash Flow Methods for Equity Valuation

GLEN A. LARSEN Jr., PhD, CFA

Professor of Finance, Indiana University Kelley School of Business—Indianapolis

Abstract: Most applied methods of valuing a firm's equity are based on discounted cash flow and relative valuation models. Although stock and firm valuation is very strongly tilted toward the use of *discounted cash flow* methods, it is impossible to ignore the fact that many analysts use other methods to value equity and entire firms. The primary alternative valuation method is *relative valuation*. Both discounted cash flow and relative valuation methods require strong assumptions and expectations about the future. No one single valuation model or method is perfect. All valuation estimates are subject to *model error* and *estimation error*.

Sound investing requires that an investor does not pay more for an asset than its worth. There are those who argue that value is in the eyes of the beholder, which is simply not true when it comes to financial assets. Perceptions may be all that matter when the asset is an art object or antique automobile, but investors should not buy financial assets for aesthetic or emotional reasons; financial assets are acquired for the cash flows expected from them in future periods. Consequently, perceptions of value have to be backed up by reality, which implies that the price paid for any financial asset should reflect the cash flows that it is expected to generate.

Realize that at the end of the most careful and detailed valuation, there will be uncertainty about the final numbers, biased as they are by the assumptions that we make about the future of the company and the economy. It is unrealistic to expect or demand absolute certainty in valuation, since cash flows and discount rates

are estimated with error. This also means that you have to give yourself a reasonable margin for error in making recommendations on the basis of valuations. Most importantly, realize that the degree of precision in valuations is likely to vary widely across investments. For example, the valuation of a large and mature company, with a long financial history, will usually be much more precise than the valuation of a young company or of a company that is in a sector that is in turmoil.

Implicit often in the act of valuation is the assumption that markets make mistakes and that we can find these mistakes, often using information that tens of thousands of other investors can access. Thus, the argument that those who believe that markets are inefficient should spend their time and resources on valuation whereas those who believe that markets are efficient should take the market price as the best estimate of value, seems to be reasonable.

This statement, though, does not reflect the internal contradictions in both positions. Those who believe that markets are efficient may still feel that valuation has something to contribute, especially when they are called upon to value the effect of a change in the way a firm is run or to understand why market prices change over time.

Furthermore, it is not clear how markets would become efficient in the first place, if investors did not attempt to find under- and over-valued stocks and trade on these valuations. In other words, a precondition for market efficiency seems to be the existence of millions of investors who believe that markets are not.

Stock-pricing models are not physical or chemical laws of nature. There is, however, a strong principle of investing that must eventually hold true for all firms over time if they are to have a positive value. This principle is that you should always be able, in your mind, to construct some sort of logical connection between a positive stock price today and a stream of future cash flows to the investor. The logical chain might be long. You might assume that years of start-up losses (earnings are zero or negative) will be followed by more years of all profits being reinvested. But you should be able to envision some connection between today's positive stock price and a stream of cash flows that will commence someday in the future.

In this entry, we discuss practical methods of valuing a firm's equity based on discounted cash flow (DCF) models. Although stock and firm valuation is very strongly tilted toward the use of DCF methods, it is impossible to ignore the fact that many analysts use other methods to value equity and entire firms. The DCF model is the subject of this entry. The primary alternative valuation method is relative valuation (RV). Both DCF and RV valuation methods require strong assumptions and expectations about the future. No one single valuation model or method is perfect. All valuation estimates are subject to model error and estimation error. Nevertheless, investors use these models to help form their expectations about a *fair market*

price. Markets then generate an observable market clearing price based on investor expectations, and this market clearing price constantly changes along with investor expectations.

DIVIDEND DISCOUNT MODEL

The *dividend discount model* (DDM) is the most basic DCF stock approach to equity valuation, originally formulated by Williams (1938). It states that the stock price should equal the present value of all expected future dividends into perpetuity under the assumption that a firm has an infinite life. But you may also have ignored the DDM once you recognized how difficult it is to apply in the real world. The next several paragraphs simply review the basic concepts in order to highlight the complexities that surround implementing the DDM in practice.

Consider an investor who buys a share of stock, planning to hold it for one year. As you know from previous studies, the *intrinsic value* of the share is the present value, $P(0)$, of the expected dividend to be received at the end of the first year, $ED(1)$, and the expected sales price, $EP(1)$.

$$P(0) = [ED(1) + EP(1)]/(1 + R) \quad (1)$$

Keep in mind that since we live in a world of uncertainty and no human can perfectly forecast the future, future prices and dividends are unknown. Specifically, we are dealing with expected values, not certain values. Under the assumption that dividends can be predictable, given a company's dividend history, the expected future dividend in the next period, $ED(1)$, can be estimated based on historical trends. You might ask how we can estimate $EP(1)$, the expected year-end price.

According to equation (1), the year-end intrinsic value, $P(1)$, will be

$$P(1) = [ED(2) + EP(2)]/(1 + R) \quad (2)$$

If we assume the stock will be selling for its intrinsic value next year, then $P(1) = EP(1)$, and we can substitute equation (2) into equation (1),

which gives

$$P(0) = ED(1)/(1 + R) + [ED(2) + EP(2)]/(1 + R)^2 \quad (3)$$

Equation (3) may be interpreted as the present value of dividends plus the expected sales price at the end of a two-year holding period. Of course, now we need to come up with a forecast of $EP(2)$. Continuing in the same way, we can replace the expected price at the end of two years by the intrinsic value at the end of two years. That is, replace $EP(2)$ by $[ED(3) + EP(3)]/(1 + R)$, which relates $P(0)$ to the value of dividends over three years plus the expected sales price at the end of a three-year holding period.

More generally, for a holding period of T years, we can write the stock value as the present value of dividends over the T years discounted at an appropriate discount rate, R , that is assumed to remain constant, plus the present value of the ultimate sales price, $EP(T)$:

$$P(0) = ED(1)/(1 + R) + ED(2)/(1 + R)^2 + \dots + [ED(T) + EP(T)]/(1 + R)^T \quad (4)$$

In short, the intrinsic price of a share of stock is the present value of a stream of payments (dividends in the case of stocks) and a final payment (the sales price of the stock at time T).

The key problems with implementing this model are the uncertainty of future dividends, the lack of a fixed maturity date, and the unknown sales price at the horizon date and the appropriate discount rate. Indeed, one can continue to substitute for a terminal price on out to infinity (INF):

$$P(0) = ED(1)/(1 + R) + ED(2)/(1 + R)^2 + \dots + ED(INF)/(1 + R)^{INF} \quad (5)$$

Equation (5) states that the stock price should equal the present value of all expected future dividends in perpetuity. This formula is the DDM in mathematical form. It is tempting, but incorrect, to conclude from the equation that the DDM focuses exclusively on dividends and ignores capital gains as a motive for investing in stock. Indeed, we assume explicitly in equation

(4), the finite version of the DDM, that capital gains (as reflected in the expected sales price, $EP(T)$) are part of the stock's value. $EP(T)$ is the present value at time T of all dividends expected to be paid after the horizon date. That value is then discounted back to today, time $T = 0$. The DDM asserts that stock prices are determined ultimately by the cash flows accruing to stockholders, and those are dividends.

Stocks That Currently Pay No Dividend

If investors never expected a dividend to be paid, then this model implies that the stock would have no value. To reconcile the fact that stocks not paying a current dividend do have a positive market value with this model, one must assume that investors expect that someday, at some time T , the firm must pay out some cash, even if only a liquidating dividend.

CONSTANT-GROWTH DDM

The general form of the DDM, as it stands, is still not very useful in valuing a stock because it requires dividend forecasts for every year into the indefinite future. To make the DDM practical, we need to introduce some simplifying assumptions. One useful and common first pass at the problem is to assume that dividends are trending upward at a stable or constant growth rate, g .

For example, if $g = 0.05$ and the most recently paid dividend was $D(0) = 3.81$, expected future dividends are

$$ED(1) = D(0)(1 + g) = (3.81)(1.05) = 4.00$$

$$ED(2) = D(0)(1 + g)^2 = (3.81)(1.05)^2 = 4.20$$

$$ED(3) = D(0)(1 + g)^3 = (3.81)(1.05)^3 = 4.41$$

and so on. Using these dividend forecasts, we can solve for intrinsic value as

$$P(0) = ED(1)/(1 + R) + ED(2)/(1 + R)^2 + ED(3)/(1 + R)^3 + \dots$$

Since the basic form of this equation stretches to infinity, basic algebra allows this equation to

be written as

$$P(0) = ED(1)/(R - g) \quad (6)$$

Equation (6) is called the constant-growth DDM, or the Gordon-Shapiro model, after Myron Gordon and Eli Shapiro, who popularized the model [see Gordon (1962) and Gordon and Shapiro (1956)].

Equation (6) should remind you of the formula for the present value of perpetuity. If dividends were expected not to grow, $g = 0$, then the dividend stream would be a simple perpetuity, and the valuation formula would be

$$P(0) = ED(1)/R$$

$P(0) = ED(1)/(R - g)$ is a generalization of the perpetuity formula to cover the case of a perpetuity growing at a constant rate, g . As g increases, for a given value of $ED(1)$, the stock price rises. The constant-growth DDM is valid only when g is less than R . If dividends were expected to grow forever (to infinity) at a rate faster than R , the value of the stock would be infinite. Further, in all of the DDM equations presented, R is also assumed to be constant forever.

NONCONSTANT-GROWTH DDM

If you feel that you know the future growth rates in each period for a firm, then you can certainly use unique growth rates, $g(T)$ and required rates of return, $R(T)$, in the present value equation and discount all unique dividends and future selling price back to the present. The problem becomes one of time, effort, and estimation risk. At some future point in time, what you believe to be a better unique estimate of a future dividend or a future discount rate will in reality be no better than an assumption of constant growth and constant discount rate.

INTUITION BEHIND THE DDM

In a market economy, common sense dictates that you should go into business only if you expect to make money. In a sole proprietorship, everything left over from the revenue you earned, minus expenses, is yours. In other forms of a business organization, you need to be a bit more formal because there are other owners. In a partnership, partners draw money out of the business. And shareholders get money out of a corporation by receiving dividends. Using the corporate form as an example, the value per share is determined by the value of the dividends distributed to each shareholder. That is, the value per share is determined by the present value of each shareholder's expected share of the profits.

Here is a simple example that illustrates several of the uncertainties involved with the basic DCF valuation process for a share of common stock. Let's say you consider buying shares of a corporation. How much will you pay if the expected annual dividend forever is \$10 per share? That depends on how much of an annual "return" you want. If you want a 10% return, you'll offer \$100 (that is, a \$10 dividend divided by a \$100 investment equals a return of 10%). But just because you offer to pay \$100 doesn't mean someone will sell to you at that price.

Financial capital is subject to principles of market supply and demand, just like commodities. Suppose market conditions are such that prevailing rates of return for corporate shares in this particular risk class are in the 5% range. If I'm selling stock that commands a \$10 per share dividend I can demand a price of \$200, and someone will give it to me. Suppose this corporation is a bit riskier than most others. A buyer may say, "If I'm willing to accept the prevailing 5% return, there are hundreds upon hundreds of better-quality corporations I can invest in. So if you want me to buy your shares, you need to give me incentive to bypass all the others. The buyer and seller may settle on a 7%

return, which is equivalent to a price of about \$143. The appropriate required rate of return, R , is therefore critical, and R can vary with market conditions.

In all cases, assuming that the life of the corporation is infinite, the current price, $P(0)$, is computed as the constant dividend in perpetuity, D , divided by the required rate of return, R , that is, the present value of all future constant dividends. Often, though, investors use return, R , as the basis for comparing and pricing investments. R is often estimated from observable information as D (dividend) divided by current price $P(0)$. Mathematically, it looks like this:

$$R = D/P(0)$$

You've seen this before. It is the dividend yield.

COMPLICATIONS IN IMPLEMENTING THE DDM IN THE REAL WORLD

As you can see by now, there are essentially four major issues that complicate finding the present value of all future dividends and, therefore, in implementing the DDM.

Expected Growth of Dividends

As profits grow over time (as we hope they will), dividends can be expected to grow and not remain constant forever. If profits and dividends are growing by 10% every year, the dividend this year may be \$10, but by next year, it will be \$11. If we divide \$11 by today's \$200 purchase price, next year's yield will be 5.5% ($11/200$). The year after, assuming further 10% growth, the dividend will be \$12.10. Dividing that by the \$200 purchase price produces a yield of 6.05%. The buyer might smile, but the seller won't accept it. The seller wants a price that truly is consistent with the prevailing 5% yield. At \$200, the buyer gets too much of a good deal. If the latter holds the stock over time, he'll wind up with an annual return well in excess of 5%.

Appropriate Expected Required Rate of Return

Simply stated, present value is a tool for computing today's equivalent of a cash payment to be made tomorrow. As stated earlier, this is often referred to as DCF valuation. If I offer you \$10 today or \$10 a year from now, you'll probably choose \$10 today. But if the choice is \$10 today or \$11.50 a year from now, you have to pause. If you can invest today's \$10 payment for one year at 5%, at the end of the year you'll have \$10.50. But if you bypass the \$10 for now and wait, you can get \$11.50 a year hence. That's a better deal. The way to decide if you should wait is to do some mathematics that helps you decide how much you must receive today to allow you to invest and wind up with \$11.50 a year hence. In this example, the "present value" of \$11.50 one year from now, assuming a 5% return, is \$10.95. If I take \$10.95 and invest it for one year at 5%, I'll wind up with \$11.50 at the end of the year. If interest rates rise, to say 8%, it'll take less money today to generate \$11.50 a year hence (\$10.65 will be sufficient). So as interest rates rise, present values fall, and vice versa.

Expected Future Selling Price

Thus far, we have thought about a stream of dividends stretching into the infinite future. Even long-term investors prefer a holding period that's something short of infinity. So we need to account for the fact that someday you'll want to sell your shares. As such, the proceeds you expect to get when you sell are included, along with dividends, in the stream of cash you expect to get, and that goes into the present value calculation.

Let's think about a projection of the future sale price. If you think you may sell in two years, imagine how a prospective buyer, two years into the future, will value the dividend stream that he'll get. Continuing with the preceding example, he'll be looking at an initial payout of

\$12.10 and a 5% return. So a price of \$244 seems a reasonable starting point. Of course, you'll need to make adjustments for probable growth beyond year two. And perhaps 5% won't be appropriate as a rate of return. Market rates may rise or fall, and/or the quality of the corporation may improve or deteriorate relative to alternative investments. And two years hence, the growth forecast may change. But in any case, we do have a \$244 starting point. The changes may bring it up, perhaps to \$275, or down, possibly to \$175. But if an exuberant analyst publishes a target price of \$1,000, you ought to raise an eyebrow and insist that the analyst get serious about justifying his presumably bold assumptions about market rates, growth, or company quality.

Reinvestment of Profits/Internal Financing that Support Growth

It is standard for corporations to refrain from paying out all annual profit as dividend. Some money is held in the business for a rainy day. And some money is simply reinvested for future growth. Either way, profits not paid out as dividends are known as retained earnings. Reinvestment is more desirable than dividend payments if the corporation can earn a higher return on the money than the shareholder could get (by reinvesting the dividends). If all goes well, the reinvestment will enable the corporation to pay a higher dividend in the future than would otherwise have been the case. Going back to the preceding example, if reinvestment gives the corporation the ability to set a year-five payout at \$18 rather than \$12.10, that raises the starting-point target price to \$360. A shareholder who accepts a forecast like that would likely forgo all or some immediate dividend payments in order to get that bigger future reward. As you can see, even if a corporation currently pays little or no dividend, we still have to acknowledge dividends as a major factor in our thoughts about share pricing.

For better or worse, many corporations now see themselves as "growth" companies. And many shareholders have accepted a situation where these publicly traded growth companies pay out very little of their profits, if anything, as dividends, and reinvest most or all profits back into the business. Many companies do not deliver nearly as well on the growth dream as everybody hopes. But the growth culture remains alive and well, and the dividend payout ratio has declined.

ADAPTING TO THE COMPLICATIONS: THE EARNINGS PER SHARE APPROACH

As a result of the four complications listed, modern stock prices have become uncoupled from dividends. So, in the real world, it is difficult to compute a fair price through the basic dividend formulas presented.

Here is one solution. It involves substituting earnings per share (EPS) for dividends. This doesn't really work in a theoretical DDM sense, but it does work within the context of a growth culture. Shareholders have so thoroughly accepted and adopted growth that they act as if all corporate EPS (whether paid as dividends or reinvested back into the business) is in their hands. So, instead of working with a dividend yield as presented earlier, we can substitute an earnings (E) yield, which is computed as follows:

$$\text{Earnings Yield} = E/P$$

Does the E/P ratio look familiar? It should. Turn it upside down and we get something you see all the time: the P/E (price/earnings) ratio.

It is important to emphasize that P/E ratios are not just one of those things we use for the heck of it. They have a serious and solid intellectual underpinning. They are equivalent to earnings yields, which are the modern-day

substitute for dividend yields—the true basis for valuing ownership of corporate stock. So when somebody states that P/E ratios are no longer relevant, you'd best turn away. Buying any stock without addressing the P/E ratio is not sensible.

When we flip P/E back over and think of earnings yield, we can understand, from the prior discussion of dividend yield, that a bad company's stock will have to offer a higher yield to attract buyers. Similarly, the yield for a great company will be low (otherwise, there would be too many would-be buyers). Let's see how this works when we flip the earnings yields back to P/Es.

If EPS equals \$3.00 and the earnings yield is 5%, the price will be \$60. If it's a bad company and the yield is higher, at 8%, the stock price will be \$37.50. If it's a good company and the yield is lower, say 3%, the stock price will be \$100. The starting number translates to a P/E as follows: a \$60 price divided by \$3.00 EPS gives us a P/E of 20. A bad-company stock price of \$37.50 divided by EPS of \$3.00 produces a P/E of 12.5. A good-company stock price of \$100 divided by EPS of \$3.00 produces a P/E of 33.3.

That's the basis for the generally recognized phenomenon of good stocks having higher P/Es and bad stocks generally having lower P/Es. So, once again, this isn't just one of those things. It's an inevitable result of the basic principles of finance and math. When evaluating companies, good or bad is usually determined based on growth prospects and risk.

We handled the complicating factors by treating EPS as if it were the same as a dividend. But notwithstanding, we still have a reasonably rational basis for stock prices. We can argue over what the growth prospects are and what the market return ought to be (based on differing assessments of market conditions and company-quality issues). So there will always be disagreement on what, exactly, a fair stock price ought to be. But all rational investors should be somewhere in the same ballpark. We may have a big ballpark and debate if a stock

that commands \$25 today is worth \$15 or \$35. But we are unlikely to seriously consider a price of, say, \$350.

FREE CASH FLOW DCF MODEL—TOTAL FIRM VALUATION

While estimating future cash flows to an individual share of stock can seem daunting, some investors prefer to estimate the free cash flow to the entire firm. Doing this allows investors to estimate the value of the entire firm and then "back out" an estimated value of a share of stock. This is called the *free cash flow* (FCF) model. While legitimate accounting rules do enable managers and auditors some range of choices, at the end of the day, good companies wind up looking good and bad companies wind up looking bad. In short, there's no one number in an income report that truly gives you the necessary information to value a firm from a discounted expected future cash flow viewpoint. You still have to select which type of cash flow you're going to look at. But the choice becomes very easy once you ask yourself the following question: What's my specific purpose for wanting to know how a company is doing?

There are many different types of users of financial information, and each is best served by concentrating on the information most relevant to him/her. Let's look at various kinds of numbers and consider what they say, and what types of investors will find them most useful.

Generally accepted accounting principles (GAAP) is a set of formal rules that produces what most of us have come to accept as the most official, or standard, version of income that a public corporation can report. Novices often believe this is the only valid number and are perplexed to learn otherwise. Essentially, GAAP is simple: Revenues minus costs equal profits. But the world is a complex place. For our convenience, we divide our activities into time periods. In a simple world, all costs would

be incurred in the same period as the revenues with which they are associated. But that is often not the case, so accountants have to find ways to identify which expenses should be matched against which revenues. One example is depreciation, a concept used to allocate multiperiod costs of a given expense to all the periods in which the expense generates revenue (e.g., if a factory can produce revenue for 10 years, charge one-tenth of the cost to build it against revenue in each year).

Observers correctly note that depreciation rules are artificial, and advocate use of other performance measures that are supposedly more “real.” We’ll touch on this later. But for now, it’s important to understand that depreciation rules are motivated by good purpose. They, and other GAAP rules, are designed to paint a picture of the “economic” performance of the business, something that is not necessarily the same as a running tally of physical dollars coming in and going out within a specific period of time.

If you are looking to see how a company is doing because you want to form an opinion as to whether or not it has a track record of “success” (defined however you wish), GAAP income is very important to you.

As noted, many investors do not like GAAP because of the artificial nature of depreciation. Their objection is valid. GAAP is, indeed, imperfect. Companies have latitude to determine how to calculate it. They don’t always use an equal allocation for each year. It’s difficult, if not impossible, to reliably estimate useful life, especially since assets are usually enhanced (that is, factories modernized) as time passes, thereby giving rise to extended life and additional depreciable expenses tacked on. An assumption that at the end of the depreciation period the asset will be worth zero, or some predetermined salvage value, is often untrue in the real world. And besides, there are other kinds of “artificial” revenue-expense matching formulations to cover other situations. But depreciation is usually the biggest objection.

Difference between Cash Flow and Free Cash Flow

The response is often to add depreciation back to net income to calculate cash flow. This can be a trap for the unwary. The phrase “cash flow” sounds comforting. After all, how much more reliable a gauge of performance can you seek than cash in minus cash out? Read the warning label closely. Is the cash flow you’re seeing truly computed by adding depreciation back to net income? If that’s what’s happening, be very careful. Companies spend money to enhance their assets every year. Because it is understood that the benefits of these expenditures will span many years, they are not put on the income statement in any single year. So, in truth, simple cash flow understates a company’s true cash-in minus cash-out situation. The solution lies in the firm’s free cash flow. To arrive at a firm’s FCF, we start with net income, add back the noncash depreciation charge, and then subtract the year’s capital-spending outlays. (There are other adjustments, such as those relating to dividends and changes in net working capital; but for now, these simple adjustments will suffice.)

Once you hone in on FCF, you aren’t likely to be misled regarding liquidity. But that does not mean you are learning about general corporate success or failure. Capital-spending programs aren’t “smooth.” In some years, expenditures are very large as major programs ramp up. In other years, capital spending shrinks as these programs wind down toward completion. If we’re in a heavy-spending year, FCF could be negative, even though the company may be having a great year.

DCF valuation depends on the construction of pro forma financial statements in order to estimate a firm’s future cash flows. Pro forma is Latin for “as if.” This measure shows how a company might perform in the future “as if” it performs as it has in the past and other assumptions that are made by the analyst. In any event, it is necessary to construct pro forma financial

statements in order to estimate future free cash flows that are the basis for total firm valuation.

CALCULATING FCF

Operating cash flow (OCF) is defined as being equal to earnings before interest and taxes (EBIT) minus taxes plus depreciation. Note, though, that cash flows cannot be maintained over time unless depreciating fixed assets are replaced. That is, the firm must reinvest in those assets that are depreciating (wearing out) so that it can stay alive. Interest paid or any other financing costs such as dividends or principal repaid are not subtracted because we are interested in the cash flow generated by the assets of the firm. The particular mixture of debt and equity a firm actually chooses to use is a managerial decision and determines how the OCF is distributed between owners (equity holders) and creditors (debt holders). The mixture also determines the firm's weighted average cost of capital (WACC), which impacts the firm's value through the discount rate.

$$\text{OCF} = \text{EBIT} - \text{Taxes} + \text{Depreciation}$$

Net operating profit after tax (NOPAT) is defined as EBIT minus taxes.

$$\text{NOPAT} = \text{EBIT} - \text{Taxes} = \text{EBIT} \times (1 - \text{Tax rate})$$

As a result, OCF can also be written as NOPAT plus any noncash adjustments. Where depreciation is the only noncash adjustment:

$$\text{OCF} = \text{NOPAT} + \text{Depreciation}$$

Free cash flow is defined as being the cash flow actually available for distribution to investors after the company has made all the investments in fixed assets and working capital necessary to sustain ongoing operations. To be more specific, the value of a company's operations depends on all the future expected FCFs, defined as OCF minus the amount of investment in working capital and fixed assets necessary to sustain the business. Thus, FCF represents the cash that is actually available for distribution to investors.

Therefore, the way for managers to make their companies more valuable is to create a sustainable increase in the firm's FCF.

$$\begin{aligned} \text{FCF} &= \text{OCF} - \text{Change in NWC} \\ &\quad - \text{Gross investment in operating capital} \end{aligned}$$

Let's illustrate this. Assume a firm has NOPAT of \$170.3 million. Its OCF is NOPAT plus any noncash adjustments as shown on the statement of cash flows. Where depreciation is the only noncash charge, the operating cash flow is:

$$\begin{aligned} \text{OCF} &= \text{NOPAT} + \text{Depreciation} \\ &= \$170.3 + \$100 = \$270.3 \text{ million} \end{aligned}$$

Further, assume the firm had \$1,455 million of operating assets, or operating capital, at the end of the year, but \$1,800 million at the end of the next year. Therefore, during the year:

$$\begin{aligned} \text{Net investment in operating capital} \\ &= \$1,800 - \$1,455 = \$345 \text{ million} \end{aligned}$$

However, the firm took \$100 million of depreciation. We find the gross investment in operating capital as follows:

$$\begin{aligned} \text{Gross investment in operating capital} \\ &= \text{Net investment} + \text{Depreciation} \\ &= \$345 + \$100 = \$445 \text{ million} \end{aligned}$$

FCF in the year is:

$$\begin{aligned} \text{CF} &= \text{OCF} - \text{Gross investment in operating capital} \\ &= \$270.3 - \$445 = -\$174.7 \text{ million (Negative FCF)} \end{aligned}$$

Even though the firm had a positive OCF, its very high investment in operating capital resulted in a negative FCF. Since FCF is what is available for distribution to investors, not only was there nothing for investors, but investors actually had to provide more money to the firm to keep the business going.

Is a negative FCF always bad? It depends on why the FCF was negative. If FCF was negative because NOPAT was negative, this is a bad sign, because the company probably is experiencing operating problems. Exceptions to this might be start-up companies, or companies

Table 1 Free Cash Flow Statement: Indirect Method

Net Income (Net Earnings)	
+ Depreciation	Depreciation is a noncash expense, and therefore is added back to calculate cash flows.
– Increase in accounts receivable (A/R)	The increase in A/R represents sales that have not yet been collected, and therefore did not produce a cash inflow.
– Increase in inventories	The increase in inventory has not been recognized as part of cost of goods sold (COGS) but was fully paid for, and therefore is deducted from the cash flow.
+ Increase in accounts payable (A/P)	The increase in A/P represents costs that have not yet been paid, and therefore is added back to the cash flow.
+ Increase in taxes payable	Like the increase in A/P, these taxes have not yet been paid.
+ After-tax interest expense	We want to evaluate the operating side of the business and its financial side separately. The interest payment is a financial expense, and therefore we add back the “net interest cost.”
= Operating cash flow (OCF)	
– Gross investment in property, plant, and equipment (PP&E), at cost	Some of the cash from operations must be used to buy the assets, such as equipment and plants that will allow the firm to generate future income. This is cash that cannot be freely used to pay dividends, to buy back shares, to repay loans, and the like, and therefore is deducted from the OCF to arrive at the FCF.
= Free cash flow (FCF)	This is the cash that the firm can use to distribute to any and all of its suppliers of capital, such as stockholders, debt holders, and warrant holders.

that are incurring significant current expenses to launch a new product line. Also, many high-growth companies have positive NOPAT but negative FCF due to new investment in operating assets needed to support growth. There is nothing wrong with profitable growth, but at some point in time FCF must turn positive in order for a firm to have value. We will see this later in a firm valuation example.

USING THE CASH-FLOW STATEMENT TO ARRIVE AT OCF AND FCF

As stated earlier, FCF is a concept that defines the amount of cash that the firm can distribute to security holders. There are two principal techniques to calculate the FCF—the indirect method and the direct method. Tables 1 and 2

Table 2 Free Cash-Flow Statement: Direct Method

Sales	As recorded on the Income Statement
– COGS+SG&A	Cost of goods sold (COGS) + Selling, general and administrative expenses (SG&A)
– Increase in accounts receivable (A/R)	Credit sales are recorded as income but do not generate a cash inflow. Thus, to adjust “sales” to cash basis, we deduct the increase in A/R.
– Increase in inventory	Inventory was paid for and thus represents a cash drain.
+ Increase in accounts payable (A/P)	A/P are expenses not yet paid.
+ Depreciation	Depreciation is not a cash expense and is netted out.
– Tax on operating income	The difference between taxes on operating income and the increase in taxes payable is the tax shield on interest, which we don’t want to include in the OCF
+ Increase in taxes payable	
= Operating cash flow (OCF)	
– Gross investment in property, plant & equipment (PP&E) at cost	
= Free cash flow (FCF)	

illustrate the direct and the indirect methods of converting accounting earnings into FCFs. The indirect approach first converts the net income (NI) to OCF then to FCF. The direct approach converts each item in the income statement to cash basis.

The indirect method of calculating cash flows starts with the firm's NI and makes appropriate adjustments to arrive at a number that shows how much cash the firm has taken in over the period. The adjustments that have to be made to NI are of two types—operational adjustments and financial adjustments. When a firm pays interest, net income is defined as

$$\text{NI} = \text{EBIT} - \text{Interest} - \text{Taxes}$$

$$\text{NI} = \text{EBT} - \text{Taxes}$$

The following adjustments must be made in order to present the results of the business activity of the firm on a cash basis as explained later in this entry.

Adjustments for Changes in Net Working Capital

Adjustments for changes in net working capital (ANWC) are made because not all the sales are made in cash and because not all the firm's expenses are paid out in cash. The term and notation are somewhat misleading: Not all the firm's working capital items are operationally related; since we are interested in cash derived from the ongoing business activity of the firm, we ignore all other current items in our ANWC. Cash and marketable securities are the best example of working capital items that we exclude from our definition of ANWC, as they are the firm's stock of excess liquidity. Another working capital item that we exclude from the adjustment is notes payable or short-term borrowing. Since our aim in the FCF statement is to calculate the cash available to the firm from its business activities, we exclude from the FCF statement any cash flows relating to the

firm's financing activities—short term or long term.

Adjustments for Investment in New Fixed Assets

When investment in these assets is necessary for the ongoing business activity of the firm, it cannot be used to pay security holders and thus must be deducted to calculate the FCF.

Adjustments for Depreciation and Other Noncash Expenses

Although depreciation is an expense for tax and financial reporting purposes (thus lowering earnings before taxes [EBT] and hence profits after taxes—[NI]), it is by itself not a cash expense. In the FCF statement, we thus add the depreciation back to NI. The remaining effect of depreciation and other noncash expenses on the FCF is the tax savings they entail.

Financial Adjustments

Financial adjustments are adjustments for financial items included in NI. Since FCF is a concept that relates to the ongoing business (as opposed to financial) activities of the firm, we want to neutralize financial items when converting NI into FCF. Thus, for example, although NI includes interest as an expense, we will add back the after-tax interest expenses to obtain the FCF.

The concept of FCF is of cash flows that are generated by the business activities of the firm and are available (that is, "free") for distribution to all suppliers of capital, such as equity holders, bondholders, convertible holders, and preferred stockholders. The calculation of accounting earnings (net income), however, is done from the point of view of shareholders, which is only one group of capital suppliers.

After calculating the FCFs, we consider their uses. The FCFs can be paid to any security holder of the firm, such as debt holders,

Table 3 Cash Flow Statement

Periodic payments	Interest Preferred dividend Regular dividend And so on	These periodic payments to the capital suppliers of the firm are after tax! (The free cash flows [FCFs] from which we pay these financial flows are also after-tax cash flows!)
Capital market transactions	Retirement of securities Debt retirement Preferred stock retirement Share repurchase And so on New financing New bank loans New bond flotation Stock sale Exercise of warrants And so on	These sums represent cash paid when old securities are retired or represent cash received when new securities are affiliated (privately or publicly).
Change in cash	=FCF – financial cash flows	

stockholders, warrant holders, and convertible bondholders.

The cash flows paid to the security holders are the financial cash flows, which include interest, dividends, principal repayment, share repurchases, and funds received upon the issuance of new securities. Obviously, when the FCF is negative (e.g., because growth opportunities necessitate large investments in fixed assets), the financial cash flows must be a net inflow of funds net new financing (of, say, the needed investments).

The difference between the funds generated by the firm's business, the FCF, and the funds distributed to the security holders of the firm, the financial cash flows (see Table 3), is the change in cash over the period.

Thus, the bottom line of the cash flow statement is the closing link of the three accounting statements of financial performance:

- The income statement's bottom line-retained earnings feeds into the closing balance sheet as the increase in accumulated retained earnings.
- The income statement and the beginning and closing balance sheets are the basis for the computation of the cash flow statements.

- The last line of the cash flow statement—change in cash (and cash equivalents)—feeds back into the end-of-period balance sheet's cash account.

The cross-reference of the three accounting statements means that we can use accounting methods to ensure that models of projected financial performance are internally consistent. The firm's income statement and its cash flow statement are often the basis for predictions of its future FCFs. Note, however, that these statements reflect the past performance of the firm and are not, in themselves, necessarily predictive of future firm performance.

VALUING THE TOTAL FIRM

Earlier we introduced several equations for valuing a firm's common stock. For example, review the constant growth dividend discount model and the nonconstant growth dividend discount model. These models (equations) have one common element: They all assume that the firm is currently paying a dividend. However, consider the situation of a start-up company formed to develop and market a new product. Such a company generally expects to have

low sales during its first few years as it develops and begins to market its product. Then, if the product catches on, sales will grow rapidly for several years. Growing sales require additional assets. A company cannot grow without increasing its assets. Moreover, increasing a liability and/or equity account must finance asset growth.

Small firms can often obtain some bank credit, but they must maintain a reasonable balance between debt and equity. Thus, additional bank borrowings require increases in equity, but small firms have limited access to the stock market. Moreover, even if they can sell stock, their owners are often reluctant to do so for fear of losing voting control. Therefore, the best source of equity for most small businesses is from retaining earnings, so most small firms pay no dividends during their rapid-growth years. Eventually, most successful firms do pay dividends, with dividends growing rapidly at first but then slowing down as the firm approaches maturity.

Although most larger firms do pay a dividend, some firms, even highly profitable ones, have never paid a dividend. How can the value of such a company be determined? Similarly, suppose you start a business and someone offers to buy it from you. How could you determine its value, or that of any privately held business? Alternatively, suppose you work for a company with a number of divisions. How could you determine the value of one particular division that the company wants to sell? In none of these cases could you use the dividend growth model. However, you could use the FCF model to estimate total firm value, then back out the value of equity.

ESTIMATING TOTAL FIRM VALUE USING THE FCF MODEL

Tables 4 and 5 contain the actual 20X8 and projected 20X9 to 20Y2 financial statements for

XYZ Inc. The negative FCF in the early years is typical for young, high-growth companies. Even though NOPAT is positive in all years, FCF is negative because of the need to invest in operating assets. The negative FCF means the company will have to obtain new funds from investors, and the balance sheets in Table 5 show that notes payable, long-term bonds, and preferred stock all increase from 20X8 to 20X9.

Assume that XYZ's cost of capital is 10.84%. To find its going-concern value, we use an approach similar to the nonconstant dividend growth model, proceeding as follows:

1. Assume that the firm will experience nonconstant growth for N years, after which it will grow at some constant rate.
2. Calculate the expected FCF for each of the N nonconstant growth years, and find the present value (PV) of these cash flows.
3. Recognize that after Year N growth will be constant, so we can use the constant growth formula to find the firm's value at Year N . This "terminal value" is the value of the PVs for $N + 1$ and all subsequent years (to infinity), discounted back to Year N . Then, the Year N value must be discounted back to the present to find its PV at Year 0.
4. Now sum all the PVs, those of the annual free cash flows during the nonconstant period plus the PV of the terminal value, to find the firm's value of operations. This going-concern value, when added to the value of the nonoperating assets, is the total value of the firm.

Stockholders will also help fund XYZ's growth. They will receive no dividends until 20Y1, so all of the net income from 20X8 to 20Y1 will be reinvested. However, as growth slows, FCF will become positive, and XYZ plans to use some of its FCF to pay dividends beginning in 20Y1. A variant of the constant growth dividend model can be used to find the value of XYZ's operations once its FCF stabilize and begin to grow at a constant rate:

Table 4 XYZ Inc.: Income Statements (in millions except for per-share data)

	Actual 20X8	Projected			
		20X9	20Y0	20Y1	20Y2
Net sales	\$700.00	\$850.00	\$1,000	\$1,100	\$1,500
Costs (except depreciation)	(599)	(734)	(911)	(935)	(982)
Depreciation	(28)	(31)	(34)	(36)	(38)
Total operating costs	(627)	(765)	(945)	(971)	(1,020)
Earnings before interest and taxes (EBIT)	73	85	55	129	135
Less "net interest"	(13)	(15)	(16)	(17)	(19)
Earnings before taxes	60	70	39	112	116
Taxes (40%)	(24)	(28)	(15.6)	(44.8)	(46.4)
Net income before preferred dividends	36	42	23.4	67.2	69.6
Preferred dividends	(6)	(7)	(7.4)	(8)	(8.3)
Net income available for common dividends	30	35	16	59.2	61.3
Common dividends	—	—	—	44.2	45.3
Addition to retained earnings	30	35	16	15	16
Number of shares	100	100	100	100	100
Dividends per share	—	—	—	0.442	0.453

Notes:

1. "Net interest" is interest paid on debt less interest earned on marketable securities. Both items could be shown separately on the income statements, but for this example we combine them and show net interest.
2. Net income is projected to decline in 20Y0. This is due to a projected cost for a one-time marketing program in that year.
3. Growth has been rapid in the past, but the market is becoming saturated, so the sales growth rate is expected to decline from 21% in 20X9 to a sustainable rate of 5% in 20Y2 and beyond (forever). Further, the entire economy has seldom grown more than a 4% to 6% rate on an average annual basis. If one firm were to grow faster than 6% forever, it would most likely become the only firm in the economy! Therefore, a 5% growth rate beyond year 20Y2 is a reasonable assumption. Firms cannot grow faster than the overall economy forever. Growth must slow down at some point in the future to a more sustainable average rate.
4. Profit margins are expected to improve as the production process becomes more efficient and because XYZ will no longer be incurring marketing costs associated with the introduction of a major product.
5. All items on the financial statements are projected to grow at a 5% rate after the year 20Y2. Notice that the company does not pay a dividend, but it is expected to start paying out about 75% of its earnings beginning in 20Y1.
6. A firm's value is determined by its ability to generate cash flow, both now and in the future. Therefore, XYZ's value can be calculated as the present value of its expected future FCFs from operations, discounted at its cost of capital, k , plus the value of nonoperating assets. Here is the equation for the value of operations, or the firm's value as a going concern:

$$\text{Value of operations} = \text{Present value of expected future FCF} + \text{Present value of nonoperating assets}$$

Based on a 10.84% cost of capital, a \$49 million FCF in 20Y2, and a 5% growth rate, the value of XYZ's operations as of December 31, 20Y2 (terminal value) is forecasted to be \$880.99 million:

$$\begin{aligned} \text{Terminal value} &= \frac{\$49(1 + 0.05)}{(0.1084 - 0.050)} \\ &= \frac{\$51.45}{(0.1084 - 0.05)} = \$880.99 \end{aligned}$$

This \$880.99 million figure is called the company's terminal or horizon value, because it is the value at the end of the forecast period. Moreover, this is the amount that XYZ could expect to receive if it sold its operating assets on December 31, 20Y2.

Table 6 shows the free cash flow for each year during the nonconstant growth period, along with the value of operations in 20Y2, at the end of the nonconstant growth period. To find the

Table 5 XYZ Inc.: Balance Sheets (millions of dollars)

	Actual 20X8	Projected			
		20X9	20Y0	20Y1	20Y2
Cash	\$17	\$20	\$22	\$23	\$24
Marketable securities (1)	63	70	80	84	88
Accounts receivable	85	100	110	116	121
Inventories	170	200	220	231	243
Total current assets	335	390	432	454	476
Net plant and equipment	279	310	341	358	376
Total assets	614	700	773	812	852
<i>Liabilities and Equity</i>					
Accounts payable	17	20	22	23	24
Notes payable	123	140	160	168	176
Accruals	43	50	55	58	61
Total current liabilities	183	210	237	249	261
Long-term bonds	124	140	160	168	176
Preferred stock	62	70	80	84	88
Common stock (2)	200	200	200	200	200
Retained earnings	45	80	96	111	127
Common equity	245	280	296	311	327
Total liabilities and equity	614	700	773	812	852

Notes:

1. All assets except marketable securities are operating assets required to support sales. The marketable securities are financial assets not required in operations.
2. Common equity is shown at par plus paid-in capital. Present value of nonoperating assets.

value of operations as of “today,” December 31, 20X8, we find the PV of each annual cash flow in Table 7, discounting at the 10.84% cost of capital.

The sum of the PVs (all FCFs and the terminal value discounted at 10.84%) is approximately \$615 million. The \$615.27 represents an estimate of the price XYZ could expect to receive if it sold its operating assets today, December 31, 20X8. The total value of any company is the value of its operations plus the value of its nonoperating assets. As the December 31, 20X8, balance sheet in Table 5 shows, XYZ had \$63 million of marketable securities on that date. Unlike operating assets, we do not have to calculate a present value for marketable securities because short-term financial assets as reported on the balance sheet are at, or close to, their market value.

Therefore, XYZ’s total value on December 31, 20X8, is $\$615.27 + \$63.00 = \$678.27$ million. If the company’s total value on December 31,

20X8, is \$678.27 million, what is the value of its common equity?

First, Table 5 shows that notes payable and long-term debt total $\$123 + \$124 = \$247$ million, and these securities have the first claim on assets and income. (Accounts payable and accruals were netted out earlier.) Next, the preferred stock has a claim of \$62 million, and it ranks above the common.

Therefore, the value left for common stockholders is $\$678.27 - \$247 - \$62 = \369.27 million.

Table 8 summarizes the calculations used to find XYZ’s stock value per share. There are 100 million shares outstanding, and their total value is \$369.27 million. Therefore, the value of a single share is \$3.69 ($\$369.27/100 = \3.69).

Much can be learned from the total firm valuation model, so many analysts today use it for all types of valuations. The process of projecting the future financial statements can reveal quite a bit about the company’s operations and

Table 6 Calculating XYZ's Pro Forma Expected Free Cash Flow (in millions)

	Actual 20X8	Projected			
		20X9	20Y0	20Y1	20Y2
Calculation of free cash flow					
Required net operating working capital	\$212	\$250	\$275	\$289	\$303
Required net plant and equipment	279	310	341	358	376
Required net operating assets	\$491	\$560	\$616	\$647	\$679
Required net new investment in operating assets = change in net operating assets from previous year	69	56	31	32	
NOPAT (Net operating profit after taxes)	EBIT \times (1 - Tax rate)	\$51	\$33	\$77.40	\$81
Less: Required investment in operating assets	69	56	31	32	
Free cash flow (FCF)		(\$18)	(\$23)	\$46.40	\$49

Notes:

- NOPAT declines in 20Y0 because of a marketing expenditure projected for that year.
- Table 4 calculates free cash flow for each year. Line 1, with data for 20X8 from the balance sheets in Table 5, shows the required net operating working capital, or operating current assets minus operating current liabilities, for 20X8:

$$\begin{aligned} \text{Required net operating working capital} &= (\text{Cash} + \text{Accounts receivable} + \text{Inventories}) \\ &\quad - (\text{Accounts payable} + \text{Accruals}) \\ &= (\$17.00 + \$85.00 + \$170.00) - (\$17.00 - \$43.00) = \$212.00. \end{aligned}$$

- Line 2 shows required net plant and equipment, and Line 3, which is the sum of Lines 1 and 2, shows the required net operating assets, sometimes called net operating capital. For 20X8, net operating capital is $\$212 + \$279 = \$491$ million.
- Line 4 shows the required net annual addition to operating assets, found as the change in net operating assets from the previous year. For 20X9, the required net investment in operating assets is $\$560 - \$491 = \$69$ million.
- Line 5 shows NOPAT, or net operating profit after taxes. Note that EBIT is operating earnings before taxes, while NOPAT is operating earnings after taxes. Therefore, $\text{NOPAT} = \text{EBIT} (1 - T)$. With 20X9 EBIT of \$85 as shown in Table 5 and a tax rate of 40%, NOPAT as projected for 20X9 is \$51 million:

$$\text{NOPAT} = \text{EBIT}(1 - T) = \$85(1.0 - 0.4) = \$51 \text{ million.}$$

- Although XYZ's operating assets are projected to produce \$51 million of after-tax profits in 20X9, the company must invest \$69 million in new assets in 20X9. Therefore, the FCF for 20X9, shown on Line 7, is a negative \$18 million:

$$\begin{aligned} \text{FCF in 20X9} &= \$51 - \$69 = -\$18.00 \text{ million (negative)} \\ &\quad \text{Present value of nonoperating assets} \end{aligned}$$

Table 7 Process for Finding the Value of Operations Assumes $g = 5\%$ (constant) for Years 12/31/Y2 in Perpetuity

Year	12/31/X8	12/31/X9	12/31/Y0	12/31/Y1	12/31/Y2
FCF		(18.00)	(23.00)	46.40	49.00
Terminal value (TV)					880.99
Total		(18.00)	(23.00)	46.40	929.99
Present value of FCF and TV @10.84% = \$615.27					
\$615.27 = Value of operating assets as of 12/31/X8					

Table 8 Finding the Value of XYZ's Stock (in millions except for per-share data)

1. Value of operations (net of payables and accruals)	\$615.27
2. Plus value of nonoperating assets	\$63.00
3. Total market value of the firm	\$678.27
4. Less: Value of debt	\$247.00
Value of preferred stock	\$62.00
5. Value of common equity	\$369.27
6. Divide by number of shares	100
7. Estimated value per share	\$3.69

financing needs. Also, such an analysis can provide insights into actions that might be taken to increase the company's value.

KEY POINTS

- The two most commonly used approaches for equity valuation are the discounted cash flow and relative valuation models.
- Despite the fact that equity valuation is very strongly tilted toward the use of discounted cash flow models, it is impossible to ignore the fact that many financial modelers employ relative valuation techniques.
- Expected future cash flow is the true basis for financial value. Take the firms that look

attractive based on “fundamentals” and attempt to estimate their current fair value based on the present value of all expected future cash flows (dividends and future selling price).

- The basic source of estimation risk when using discounted cash flow models in calculating the value of any financial asset is that the present value depends on expected future cash flows and the appropriate discount rates that reflect the risk of the future cash flows. Cash flow valuation models, therefore, rely on assumptions (often extreme).
- With cash flow valuation, the main problem is estimation risk. No financial modeler can correctly and consistently forecast the future. Estimation risk comes from not being able to perfectly forecast future cash flows and discount rates.

REFERENCES

- Gordon, M. J. (1956). Capital equipment analysts: The required rate of profit. *Management Science* 3, 1: 102–110.
- Gordon, M. J. (1962). *The Investment, Financing, and Valuation of the Corporation*. Homewood, IL: Richard D. Irwin.
- Williams, J. B. (1938). *The Theory of Investment Value*. Cambridge: Harvard University Press.

Relative Valuation Methods for Equity Analysis

GLEN A. LARSEN Jr., PhD, CFA

Professor of Finance, Indiana University, Kelley School of Business–Indianapolis

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

CHRIS GOWLLAND, CFA*

Senior Quantitative Analyst, Delaware Investments

Abstract: Relative valuation methods use multiples or ratios, such as price/earnings, price/book, or price/free cash flow, to determine whether a particular firm is trading at higher or lower multiples than its peers. Such methods require the user to choose a suitable universe of firms that are more or less comparable, though this can become difficult for firms with unusual characteristics in terms of product mix or geographical exposure. Relative valuation methods can be useful for portfolio managers who expect to be fully invested at all times, as they provide a practical tool for attempting to capture the “*value premium*” by which firms trading at lower multiples tend to outperform those trading at higher multiples. Implicitly, relative valuation methods assume that the average multiple across the universe of firms can be treated as a reasonable approximation of “*fair value*” for those firms; this may be problematic during periods of *market panic* or *euphoria*.

Much research in corporate finance and similar academic disciplines is tilted toward the use of discounted cash flow (DCF) methods. However, many analysts also make use of *relative valuation* methods, which compare several firms by using multiples or ratios. Multiples that are commonly used for such purposes include *price/earnings*, *price/book*, and *price/free cash flow*.

Relative valuation methods implicitly assume that “*similar*” firms are likely to be valued similarly by investors. Therefore, on average, we would expect that firms that are generally comparable are likely to trade at similar multiples, in terms of price/earnings, price/book, or various other metrics. If this assumption is approximately correct, then relative valuation methods can be used to identify firms

*The material discussed here does not necessarily represent the opinions, methods, or views of Delaware Investments.

that look “cheap” or “expensive” relative to their peers. When a particular firm’s multiples are extremely different from the rest of the universe, this may indicate a potential investment opportunity—though further analysis will likely be required to determine whether there are reasons why such a firm is valued differently from other companies that otherwise appear comparable.

The basis of relative valuation methods is to use one or several ratios to determine whether a firm looks “cheap” or “expensive” by comparison with generally similar firms. Relative valuation methods do not attempt to explain why a particular firm is trading at a particular price; instead, they seek to measure how the market is currently valuing multiple companies, with the underlying assumption that the average multiple for a group of companies is probably a reasonable approximation to overall market sentiment toward that particular industry. In other words, relative valuation work assumes that on average, the share prices of companies in a particular universe are likely to trade at similar multiples relative to their own financial or operating performance. Baker and Ruback (1999) provide a more formal presentation of these concepts. However, it is important to realize that at any particular time, some firms are likely to be trading at higher or lower multiples than would be justified under “fair value.”

Making effective use of relative valuation methods does require careful selection of “similar” companies. Sometimes this is relatively simple, for instance when an analyst is dealing with industries where there are a large number of roughly homogeneous firms providing goods or services that are approximately equivalent. However, sometimes there can be considerable difficulties in identifying “similar” companies, particularly if the firms under consideration are unusually idiosyncratic in terms of their product mix, geographical focus, or market position. In this entry, we will provide some tentative guidance about how to build a universe of *comparable companies*. However, ultimately this part of the process will depend on

the skill and knowledge of the individual analyst; two different experts may pick different sets of “similar” firms, and thus generate different values from their relative valuation analysis.

BASIC PRINCIPLES OF RELATIVE VALUATION

Analysis based on relative valuation requires the analyst to choose a suitable universe of companies that are more or less comparable with one another. There is no standardized approach concerning how to choose such a universe of similar firms, and the process relies to some extent on an analyst’s personal judgment concerning the particular industry and geography involved. However, it is possible to lay out some general principles that combine practitioners’ insights with the results of academic inquiry.

Sources of Data

Relative valuation approaches can only be employed if there is sufficient information, produced on an approximately consistent footing, about the various companies that are the subjects of analysis. In most countries, companies that are publicly listed on stock exchanges are required by law and regulation to report their historical results publicly in a timely manner, or risk being delisted from the exchange. (There may be occasional exceptions to this general pattern, particularly for entities that are majority owned or controlled by their home country government. But such anomalies are not frequently observed except during crisis periods.) Consequently, it is almost always possible to obtain information about listed companies’ historical results. However, multiples based solely on historical data may not provide a complete picture, as most analysts would probably agree that forward-looking estimates are likely to provide more useful insights into the market’s opinion of a particular company (Valentine, 2011, p. 261).

Investment banks, rating agencies, and other firms can provide estimates of a firm’s future

earnings, revenues, and other metrics, typically over the next two or three years. Various data providers such as Bloomberg or Thomson Reuters collect such information and use it as the basis for “consensus” estimates, which can be viewed as representing the market’s general opinion of a company’s future prospects. It is also possible to use a firm’s own in-house estimates for the companies under coverage, as these may incorporate insights that are not yet reflected in current pricing. However, for precisely this reason, in-house estimates should be used as a supplement rather than as a replacement for consensus figures.

It is conventional to consider more than one year of data, as there may be disparities in how the market is valuing results in the immediate future and in the slightly longer term. However, it is often difficult or impossible to obtain consensus estimates more than two or three years into the future. Consequently, relative valuation approaches generally focus on relatively short periods into the future, rather than seeking to gauge how the market is valuing expected performance five or ten years hence. (In this respect, relative valuation analysis can be viewed as somewhat limited by comparison with DCF approaches, which typically give considerably more attention to the relatively distant future.)

Number of Comparable Firms

In general, an analyst would like to use data from other firms that are as similar as possible. However, if the criteria for “similarity” are specified too stringently, then there may be too few firms included in the universe. And if the sample is too small, then the idiosyncrasies of individual firms may exert an excessive influence on the average multiple, even if the analyst focuses on the median rather than the mean when calculating the “average” multiple.

Generally speaking, we believe that it is desirable to have at least five or six comparable companies, in order to begin drawing conclusions about relative valuation for a particular indus-

try. Conversely, there may be few benefits from considering more than 12 companies, particularly if the larger universe contains firms that resemble less closely the particular company that is the focus of the analyst’s attention.¹ For most practical purposes, a group of between six and 12 comparable firms should be sufficiently large to produce usable results.

Basis for Selecting Comparable Firms

In an ideal situation, a universe of comparable companies would be similar in terms of size, industry focus, and geography. This tends to be easier when considering small or mid-sized firms—say, with market caps between \$100 million and \$10 billion (based on 2010 U.S. dollars). Firms that are below this size limit, in other words microcap stocks, may be more difficult to use for relative valuation purposes. Even if these firms are public, they may receive less coverage from research analysts, who typically are more interested in companies that are large, liquid, and already owned by institutional investors (see Bhushan, 1989).

Conversely, it can also be difficult to perform relative value analysis on companies with relatively high market capitalization. Many large firms are dominant players in their particular market niches, in which case they may be more likely to trade at a premium reflecting their higher degree of market power. Alternatively, large firms may be effectively a conglomerate of numerous smaller entities, each engaged in a specific activity, and there may be no other large or small firm that produces an approximately equivalent blend of goods and/or services.

When attempting to assess the relative value of firms that are large and/or complex, it can often be useful to assess “relative value” using two separate approaches. The first approach is to consider the firm as a complete entity and try to find other firms that are at least somewhat comparable in terms of size and complexity, even if their business mix is not precisely identical. In such cases, it can often be useful

to consider similar firms that may be located in other countries, even though their different geographical positioning may affect their level of risk and thus the multiples at which they trade. The second approach is to use a sum-of-the-parts valuation method, which will be discussed in more detail later in this entry.

Geography and Clientele

Differences related to geographic location can affect the extent to which companies can be viewed as broadly similar. For instance, in the United States public utilities are predominantly regulated at the state level, and the public utility commissions in one state may operate quite differently from their counterparts elsewhere. Consequently, a public utility operating in one state may not be directly comparable with a public utility located in another state. In recent decades, there has been a wave of acquisition activity in the U.S. utility industry, so that now some utilities have operations in multiple states. In such instances, the valuation placed on a utility will presumably incorporate investors' perceptions of the regulatory environment affecting each of its state-level operations. For relative value purposes, a group of multistate public utilities may not be very similar to a public utility that is operating in only one state.

Regional differences in regulatory regimes may only affect a subset of companies. However, firms in the same industry may well have quite different client bases and geographic exposures. For instance, one retailer may aim to sell a wide range of goods to a mass-market client base at the regional or national level, while another retailer might instead focus on selling a limited number of luxury products to the most affluent members of the global population. These two firms are likely to have substantially different product quality, cost bases, profit margins, and sensitivity to macroeconomic conditions. In particular, retailers of luxury goods to a global client base may have developed brands that transcend national borders, and a high proportion of their current and future

revenues and profits may come from outside their home country. Under such conditions, it is possible that a suitable universe of comparable companies might include at least a few foreign firms, particularly if they have similarly broad geographic reach.

In past decades, analysts focusing on U.S. firms would probably have only rarely used foreign firms in their analysis of "comparable companies." However, as both U.S. and foreign firms have become increasingly globalized, and as accounting standards around the world have gradually started to become more similar, we believe that for some types of relative value analysis, there may be benefits to including firms that are generally comparable in terms of size and product mix, even if their legal headquarters are not located in the United States. For more insights into these issues, see Copeland, Koller, and Murrin (2000, Chapter 18).

Many companies have "depository receipts" in other markets, such as ADRs. Consensus estimates may be available for a firm's local results and/or its depository receipts. The estimates for the depository receipts may be affected by actual or expected movements between the currencies of the two countries, which may bias the analysis. We therefore recommend that when calculating figures for companies that are listed in different countries, all multiples should be consistently calculated in terms of local currency throughout, in order to ensure that anticipated or historical currency fluctuations will not affect the results. A substantial number of non-US companies have a share price quoted in one currency, but report their financial results in another currency; to avoid potential mismatch-related errors in such cases, it may be prudent to convert all numbers into a single numéraire such as the US dollar.

Sector and Industry Characteristics

Some academic research has examined different ways of selecting a universe of comparable firms. Bhojraj, Lee, and Oler (2003) compared the effect of using four different industry

classification methods, and concluded that at least for a universe of U.S. securities, the Global Industry Classification Standard (jointly developed and maintained by Standard & Poor's and Morgan Stanley Capital International) appeared to do the best job of identifying firms with similar *valuation multiples* and stock price movements. Chan, Lakonishok, and Swaminathan (2007) compared the effect of using industry classification schemes with statistically based clustering approaches, and found that examining stocks in terms of industry membership seemed to give better explanatory power than working in terms of either sectors or subindustries. To our knowledge, there have not been any parallel investigations into the effectiveness of different industry classification schemes for cross-national analysis. The results of Phylaktis and Xia (2006) suggest that the importance of sector-level effects has been increasing in recent years, while the influence of country-level effects has waned slightly.

Technology and Intraindustry Diversity

As discussed above, some academic research has suggested that firms from similar industries tend to trade at similar multiples and to experience similar stock price movements. Industry membership therefore would seem to be a useful starting point for analysis. Thus, for instance, trucking companies and railroad companies both provide transportation services, but railroads will generally trade at different multiples from trucking companies because their cost structure and balance sheets tend to be quite different.

In some cases, there can be substantial variation even within a particular subindustry. For instance, "publishing" covers a wide variety of different business models, including daily newspapers, weekly magazines, publishers of textbooks and professional journals, printers of fiction or nonfiction books, and suppliers of financial data. Each of these individual industries is likely to have different sources of

revenue, different technological requirements, different cost structures, and different rates of expected growth. Admittedly, the larger publishing houses may have operations spanning several different fields, but the relative contributions of each division to the firm's overall revenues and profits may differ substantially. In such instances, relative value analysis may result in a wide range of valuation multiples, possibly with several different clusters reflecting each firm's competitive position. We consider such difficulties in the next section.

There are also some industries in which technological differences are the principal basis on which relative values are assigned. For instance, small companies in the field of biotechnology may have only a handful of products, each of which could potentially be a great success or a dismal failure. Some companies of this type may be still at the prerevenue stage when they go public, so that their valuation is entirely based on the market's expectations about the ultimate value of technology that has not yet generated actual sales. In such instances, relative value analysis might require particularly careful selection of companies that are truly comparable in terms of the market's perception of their stage of development and the likelihood that their key products will ultimately be successful. Arguably, relative value analysis in such cases may not generate particularly useful results, because the spread of potential outcomes is so broad.

Bimodal and Multimodal Patterns

Sometimes the outcome of a relative value analysis will show that the valuation multiples are not evenly spread between low and high, but instead are bimodal or multimodal—in other words, there seem to be two or more clusters of results. We show an example of this in our hypothetical example below, which suggests that in a universe of seven firms, two are expected to achieve a return on equity (ROE) of 11% to 12% in FY0 and FY1, whereas the other companies are generally projected to deliver an

ROE of 8% to 9%. Such differences may appear relatively minor, but if the market really does expect these outcomes, then the two companies with higher profitability may legitimately be expected to trade at a premium to their peers.

When a relative valuation table appears to have bimodal or multimodal characteristics, an analyst will generally be well advised to investigate further. In any given sector or industry, there may well be some firms that are truly capable of producing higher returns than their peers, perhaps as a result of better management, a stronger market position, or a more supportive regulatory environment. Relative valuation methods can identify potential outliers of this type, but cannot test whether the estimates themselves are reasonable.

One potentially useful approach is to extend the analysis further back into the past, using historical prices for valuation purposes, and if possible also using as-was projections for the relevant period. Such projections are now widely available from various different data vendors, including Bloomberg, FactSet, and Thomson Reuters. Consider the companies that are currently trading at a premium or a discount to their peers—did they also trade at a discount to their peers in the past? A logical extension of relative value analysis based on a single period is to gauge whether a particular firm persistently tends to trade at a lower or higher multiple than its peers, and then assess whether its current multiple is above or below what would be expected on the basis of prior periods. Damodaran (2006, Chapter 7, p. 244) notes that relative valuations frequently have low persistence over time. For industries in which this is the case, then relative valuation methods may indeed provide useful investment signals.

Choice of Valuation Multiples

Many relative valuation methods compare a company's share price with some measure of its performance, such as earnings per share (EPS) or free cash flow per share. Other relative val-

uation methods compare a company's share price with some measure of its size, such as book value per share. Block (1999) has reported that the majority of practitioners consider that when analyzing securities, measures of earnings and cash flow are somewhat more important than measures of book value or dividends. However, many practitioners will make use of various metrics in their work, in the expectation that the different multiples will provide varying perspectives. Liu, Nissim, and Thomas (2002) compared the efficacy of six different metrics for relative valuations of U.S. firms on a universe-wide basis. Liu, Nissim, and Thomas (2007) extended the analysis to seven different metrics applied to 10 different countries and multiple industries. Hooke (2010, Chapter 15) presents an example using eight different metrics applied to the single industry of temporary staffing companies. In a hypothetical example below, we use three different metrics for relative valuation analysis, and we believe that most practitioners would consider that between three and six different metrics is probably justifiable. It is certainly possible to have a much larger number of metrics (see Damodaran, 2006, p. 650), but the results may be harder to interpret.

A ratio such as price/earnings can be calculated in terms of share price/EPS, or alternatively can be interpreted as market cap/net income. For most purposes, these two ratios will be the same. However, share issuance or buyback activity may impair the comparability of figures expressed in terms of EPS. If there is any possibility of ambiguity, then we would generally recommend using market cap/net income.

For instance, a company may currently have 100 million shares outstanding, a current share price of \$40, and expected earnings of \$2 in FY0 and \$3 in FY1. If the P/E ratio is calculated in terms of price/EPS, then the FY0 ratio is 20 and the FY1 ratio is 13.3. However, analysts may be expecting that the company will buy back and cancel 20% of its shares during FY1. If so,

then the projected net income in FY1 would presumably be \$240 million rather than \$300 million. If the P/E ratio is calculated using market cap and net income, then the FY1 ratio would be 16.7 rather than 13.3. This hypothetical example indicates the importance of ensuring that the denominator is being calculated on a basis that reflects the historical or projected situation for the relevant period. (An investor might consider that if a firm's management is indeed strongly committed to buying back its own shares, then this might indicate that the firm's management views the shares as being undervalued. However, such considerations would presumably be included as a qualitative overlay to the relative valuation analysis.)

Choice of Numerator: Market Cap versus Firm Value

In some instances, the choice of numerator may have a significant impact on the multiple. For instance, many analysts will use price/sales ratios for valuation purposes. However, a firm's revenues are generated from the total of its capital base, comprising both equity and debt.

Consider two companies, A and B, which both have a current market cap of \$300 million and projected annual revenues of \$600 million in FY0, so that they both have a current price/sales ratio of 2. But suppose that Company A has no outstanding borrowings, whereas Company B has net debt of \$300 million. One could argue that Company B is actually rather less attractive than Company A, as apparently it requires twice as much capital to generate the same volume of sales. In effect, analyzing the company in terms of "firm value/sales" rather than price/sales would reveal that Company B is actually making less efficient use of its capital than Company A.

There is no single definition of "firm value" that is generally accepted by all practitioners. In an ideal world, one would want to have the market value of the firm's equity capital and of the firm's debt capital. However, because

corporate bonds and bank loans typically are not traded in liquid markets, there may not be any reliable indicator of the market value of debt capital. Consequently, it is conventional to use market capitalization to estimate how investors are valuing the firm's equity capital, but then to use figures from the firm's most recent balance sheet together with the notes to the financial statements as a proxy for net debt. The broadest definition of which we are aware is the following:

$$\begin{aligned} \text{Net Debt} &= \text{Total Short-Term Debt} \\ &+ \text{Total Long-Term Debt} + \text{Minority Interest} \\ &+ \text{Unfunded Pension Liabilities} \\ &- \text{Cash and Equivalents} \end{aligned}$$

In practice, for most firms, the biggest components of net debt are likely to be total short-term debt, total long-term debt, and cash and equivalents. In most cases, using an alternative definition of firm value will often have only a small impact on the calculated multiple.

Conceptually, it is possible to divide the income statement between the line items that are generated on the basis of total capital, and those that pertain solely to equity capital. For most firms, the separator between these two categories is Net Interest Expense or Net Interest Income. Analyzing relative valuation for banks and insurance companies can be somewhat more complex, as discussed in Copeland, Koller, and Murrin (2000, Chapters 21 and 22). Generally speaking, it is usually desirable that the numerator and denominator of a valuation metric should be consistent with each other (Damodaran, 2006, pp. 239–240).

Industry-Specific Multiples

Analysts covering some industries may make use of information specific to that industry, such as paid miles flown for airlines, same-store sales for retailers, or revenue per available room for hotel chains. Such data can provide insights into how the market is valuing individual

firms' historical or expected operating performance. However, we consider that they should be viewed as a supplement to other multiples, rather than as a replacement for them, for two reasons: because it can be difficult to reconcile a company's operating performance with its financial results, and also because there may be little or no intuition about what would be a "reasonable" estimate for long-run valuation levels (Damodaran, 2006, Chapter 7, pp. 237–238). Natural resource producers tend to be valued in terms of both their operating efficiency and the resources that they control, so it may be useful to include some measure of their reserves in the analysis (Hooke, 2010, Chapter 21). Many practitioners make use of efficiency metrics when using relative valuation approaches to assess some types of banks and other lending institutions (Hooke, 2010, Chapter 22).

HYPOTHETICAL EXAMPLE

Suppose that an analyst is seeking to gauge whether Company A is attractive or unattractive on the basis of relative valuation methods. Suppose that the analyst has determined that there are six other listed companies in the same

industry which are approximately the same size, and which are also comparable in terms of product mix, client base, and geographical focus.² Based on this information, the analyst can calculate some potentially useful multiples for all seven companies. A hypothetical table of such results is shown in Table 1. (For the purposes of this simple hypothetical example, we are assuming that all the firms have the same fiscal year. We will consider calendarization later in this entry.)

In this hypothetical scenario, Company A is being compared to Companies B through G, and therefore Company A should be excluded from the calculation of median and standard deviation, which would otherwise lead to double-counting. The median is used because it tends to be less influenced by outliers than the statistical mean, so it is likely to be a better estimate for the central tendency. (Similarly, the standard deviation can be strongly influenced by outliers, and it would be possible to use "median absolute deviation" as a more robust way of gauging the spread around the central tendency. Such approaches may be particularly appropriate when the data contain one or a handful of extreme outliers for certain metrics, which might be associated with company-specific idiosyncrasies.) The table has been arranged in terms of market

Table 1 Hypothetical Relative Valuation Results

Company	Share Price (\$)	Market Cap (\$m)	P/E		P/FCF		P/B	
			FY0	FY1	FY0	FY1	FY0	FY1
A	20.00	400	12.0	10.0	8.5	7.0	1.30	1.20
B	16.00	550	11.5	11.5	5.0	6.0	1.00	0.95
C	40.00	500	13.0	12.0	8.0	7.5	1.50	1.40
D	15.00	450	12.5	12.0	8.0	7.0	1.10	1.05
E	13.00	350	14.5	13.0	9.0	8.0	1.25	1.15
F	30.00	350	12.5	12.5	7.0	4.5	1.15	1.15
G	15.00	300	15.0	14.0	7.0	6.0	1.20	1.15
Median		400	12.75	12.25	7.50	6.50	1.18	1.15
Std Dev		98.3	1.33	0.89	1.37	1.26	0.17	0.15
<i>A versus Median</i>		0%	−6%	−18%	13%	8%	11%	4%

Notes: P/E refers to price/earnings before extraordinary items; P/B refers to price / book value; P/FCF refers to price/free cash flow (defined as earnings before extraordinary items plus noncash items taken from the cash flow statement); FY0 refers to the current fiscal year; FY1 refers to the next fiscal year; figures for FY0 and FY1 could have been derived from consensus sell-side estimates or other sources.

cap, from largest to smallest, which can sometimes reveal patterns associated with larger or smaller firms, though there don't appear to be any particularly obvious trends in this particular set of hypothetical numbers.

The table suggests that the chosen universe of comparable companies may be reasonably similar to Company A in several important respects. In terms of size, Companies B, C, and D are slightly larger, while Companies E, F, and G are slightly smaller, but the median market cap across the six firms is the same as Company A's current valuation. In terms of P/E ratios, Company A looks slightly cheap in terms of FY0 earnings and somewhat cheaper in terms of FY1 earnings. In terms of P/FCF ratios, Company A looks somewhat expensive in terms of FY0 free cash flow, but only slightly expensive in terms of FY1 free cash flow. And finally, in terms of P/B ratios, Company A looks somewhat expensive in terms of FY0 book value, but roughly in line with its peers in terms of FY1 book value.

Analysis of the Hypothetical Example

So what are the implications of these results? First, Company A looks relatively cheap compared to its peer group in terms of P/E ratios, particularly in terms of its FY1 multiples. Second, Company A looks rather expensive compared to its peer group in terms of P/FCF and P/B ratios, particularly in terms of FY0 figures. If an analyst were focusing solely on P/E, then Company A would look cheap compared with the peer group, and this might suggest that Company A could be an attractive investment opportunity.

However, the analyst might be concerned that Company A looks comparatively cheap in terms of P/E, but somewhat expensive in terms of price/book. One way to investigate this apparent anomaly is to focus on ROE, which is defined as earnings/book value. Using the data in the table, it is possible to calculate the ROE

for Company A and for the other six companies by dividing the P/B ratio by the P/E ratio—because this effectively cancels out the “price” components, and thus will generate an estimated value for EPS divided by book value per share, which is one way to calculate ROE.

The results suggest that Company A is expected to deliver an ROE of 10.8% in FY0 and 12% in FY1, whereas the median ROE of the other six firms is 8.7% in FY0 and 8.8% in FY1. Most of the comparable companies are expected to achieve an ROE of between 8% and 9% in both FY0 and FY1, though apparently Company C is expected to achieve an ROE of 11.5% in FY0 and 11.7% in FY1. (A similar analysis can be conducted using “free cash flow to equity,” which involves dividing the P/B ratio by the P/FCF ratio. This indicates that Company A is slightly below the median of Companies B through G in FY0, but in line with its six peers during FY1.)

These results suggest that Company A is expected to deliver an ROE that is substantially higher than most of its peers. Suppose that an analyst is skeptical that Company A really can deliver such a strong performance, and instead hypothesizes that Company A's ROE during FY0 and FY1 may only be in line with the median ROE for the peer group in each year. Based on the figures in Table 1, Company A's book value in FY0 is expected to be \$15.38, and the company is projected to deliver \$1.67 of earnings. Now suppose that Company A's book value remains the same, but that its ROE during FY0 is only 8.7%, which is equal to the median for its peers. Then the implied earnings during FY0 would only be \$1.35, and the “true” P/E for Company A in FY0 would be 14.9, well above the peer median of 12.75.

The analysis can be extended a little further, from FY0 to FY1. The figures in the table above suggest that Company A's book value in FY1 will be \$16.67, and that the company will generate \$2.00 of earnings during FY1. But if Company A only produced \$1.35 of earnings during FY0, rather than the table's expectation of \$1.67, then the projected FY1 book value may be too

high. A quick way to estimate Company A's book value in FY1 is to use a *clean surplus* analysis, using the following equation:

$$\text{Book}_{\text{FY1}} = \text{Book}_{\text{FY0}} + \text{Net Income}_{\text{FY1}} \\ - \text{Dividends}_{\text{FY1}}$$

Based on the figures in the table above, Company A is expected to have earnings of \$1.67 during FY0, and \$2.00 during FY1. The implied book value per share is \$15.38 in FY0, and \$16.67 during FY1. According to the clean surplus formula, Company A is expected to pay a dividend of \$0.38 per share in FY1.

Assuming that the true earnings in FY0 are indeed \$1.35 rather than \$1.67, and that the dividend payable in FY1 is still \$0.38, then the expected book value for Company A in FY1 would be \$16.35 rather than \$16.67. Taking this figure and applying the median FY1 peer ROE, the expected FY1 earnings for Company A would be \$1.42 rather than \$2.00, and consequently the "true" P/E for FY1 would be 13.9 instead of the figure of 10.0 shown in the table. At those levels, the stock would presumably no longer appear cheap by comparison with its peer group. Indeed, Company A's FY1 P/E multiple would be roughly in line with Company G, which has the highest FY1 P/E multiple among the comparable companies.

This quick analysis therefore suggests that the analyst may want to focus on why Company A is expected to deliver FY0 and FY1 ROE that is at or close to the top of its peer group. As noted previously, Company A and Company C are apparently expected to have an ROE that is substantially stronger than those of the other comparable companies. Is there something special about Companies A and C that would justify such an expectation? Conversely, is it possible that the estimates for Companies A and C are reasonable, but that the projected ROE for the other companies is too pessimistic? If the latter scenario is valid, then it's possible that the P/E ratios for some of the other companies in the comparable universe are too high, and thus

that those firms could be attractively valued at current levels.

Other Potential Issues

Multiples Involving Low or Negative Numbers

It is conventional to calculate valuation multiples with the market valuation as the numerator and the firms' financial or operating data as the denominator. If the denominator is close to zero, or negative, then the valuation multiple may be very large or negative. The simplest example of such problems might involve a company's earnings. Consider a company with a share price of \$10 and projected earnings of \$0.10 for next year. Such a company is effectively trading at a P/E of 100. If consensus estimates turn more bearish, and the company's earnings next year are expected to be minus \$0.05, the company will now be trading at a P/E of -200.

It is also possible for a firm to have negative shareholders' equity, which would indicate that the total value of its liabilities exceeds the value of its assets. According to a normal understanding of accounting data, this would indicate that the company is insolvent. However, some companies have been able to continue operating under such circumstances and even to retain a stock exchange listing. Firms with negative shareholders' equity will also have a negative price/book multiple. (In principle, a firm can even report negative net revenues during a particular period, though this would require some rather unusual circumstances. One would normally expect few firms to report negative revenues for more than a single quarter.)

As noted previously, averages and standard deviations tend to be rather sensitive to outliers, which is one reason to favor using the median and the median absolute deviation instead. But during economic recessions at the national or global level, many companies may have low or negative earnings. Similarly, firms in cyclical industries will often go through

periods when sales or profits are unusually low, by comparison with their average levels through a complete business cycle. Under such circumstances, an analyst may prefer not to focus on conventional metrics such as Price/Earnings, but instead to use line items from higher up the income statement that typically will be less likely to generate negative numbers.

Calendarization

Some of the firms involved in the relative valuation analysis may have fiscal years that end in different months. Most analyst estimates are based on a firm's own reporting cycle. It is usually desirable to ensure that all valuation multiples are being calculated on a consistent basis, so that calendar-based effects are not driving the analysis.

One way to ensure that all valuation multiples are directly comparable is to calendarize the figures. Consider a situation where at the start of January, an analyst is creating a valuation analysis for one firm whose fiscal year ends in June, while the other firms in the universe have fiscal years that end in December. Calendarizing the results for the June-end firm will require taking half of the projected number for FY0 and adding half of the projected number for FY1. (If quarter-by-quarter estimates are available, then more precise adjustments can be implemented by combining 3QFY0, 4QFY0, 1QFY1, and 2QFY1.)

Calendarization is conceptually simple, but may require some care in implementation during the course of a year. One would expect that after a company has reported results for a full fiscal year, the year defined as "FY0" would immediately shift forward 12 months. However, analysts and data aggregators may not change the definitions of "FY0" and "FY1" for a few days or weeks. In case of doubt, it may be worth looking at individual estimates in order to double-check that the correct set of numbers is being used.

Sum-of-the-Parts Analysis

When attempting to use relative valuation methods on firms with multiple lines of business, the analyst may not be able to identify any company that is directly similar on all dimensions. In such instances, relative valuation methods can be extended to encompass "sum-of-the-parts" analysis, which considers each part of a business separately and attempts to value them individually by reference to companies that are mainly or solely in one particular line of business (see Hooke, 2010, Chapter 18).

Relative valuation analysis based on sum-of-the-parts approaches will involve the same challenges as were described above—identifying a suitable universe of companies engaged in each particular industry, collecting and collating the necessary data, and then using the results to gauge what might be a "fair value" for each of the individual lines of business. But in addition to these considerations, there is an additional difficulty, which is specific to sum-of-the-parts analysis. This problem is whether to apply a *conglomerate discount*, and if so, how much.

Much financial theory assumes that all else equal, investors are likely to prefer to invest in companies that are engaged in a single line of business, rather than to invest in conglomerates that have operations across multiple industries. Investing in a conglomerate effectively means being exposed to all of that conglomerate's operations, and the overall mix of industry exposures might not mimic the portfolio that the investor would have chosen if it were possible instead to put money into individual companies.

A possible counterargument might be that a conglomerate with strong and decisive central control may achieve *synergies* with regard to revenues, costs, or taxation that would not be available to individual free-standing firms dealing at arms' length with one another. A skeptical investor might wonder, on the other hand, about whether the potential positive impact of such synergies may be partly or wholly

undermined by the negative impacts of centralized decision making, transfer pricing, and regulatory or reputational risk.

For these reasons, an analyst might consider that it is reasonable to apply a discount to the overall value that emerges from the “sum of the parts.” Some practitioners favor a discount of somewhere between 5% and 15%, for the reasons given above. Academic research on spinoffs has suggested that the combined value of the surviving entity and the spun-off firm tends to rise by an average of around 6%, though with a wide range of variation (see Burch and Nanda, 2003). (Some analysts have suggested that in some particular contexts, for instance in markets where competent managers are very scarce, then investors should be willing to pay a premium for being able to invest in a conglomerate that is fortunate enough to have such executives. However, this appears not to be a mainstream view.)

Relative Valuation versus DCF: A Comparison

Relative valuation methods can generally be implemented fairly fast, and the underlying information necessary to calculate can also be updated quickly. Even with the various complexities discussed above, an experienced analyst can usually create a relative valuation table within an hour or two. And the calculated valuation multiples can adjust as market conditions and relative prices change. In both respects, relative valuation methods have an advantage over DCF models, which may require hours or days of work to build or update, and which require the analyst to provide multiple judgment-based inputs about unknowable future events. Moreover, as noted by Baker and Ruback (1999), if a DCF model is extended to encompass multiple possible scenarios, it may end up generating a range of “fair value” prices that is too wide to provide much insight into whether the potential investment is attractive at its current valuation.

Relative valuation methods focus on how much a company is worth to a minority shareholder, in other words an investor who will have limited or zero ability to influence the company’s management or its strategy. Such an approach is suitable for investors who intend to purchase only a small percentage of the company’s shares and to hold those shares until the valuation multiple moves from being “cheap” to being “in line” or “expensive” compared with the peer group. As noted above, relative valuation methods make no attempt to determine what is the “correct” price for a company’s shares, but instead focus on trying to determine whether a company looks attractive or unattractive by comparison with other firms that appear to be approximately similar in terms of size, geography, industry, and other parameters.

DCF methods attempt to determine how much a company is worth in terms of “fair value” over a long time horizon. DCF methods can readily incorporate a range of assumptions about decisions in the near future or the distant future, and therefore can provide a range of different scenarios. For this reason, most academics and practitioners consider that DCF methods are likely to produce greater insight than relative valuation methods into the various forces that may affect the fair value for a business. More specifically, DCF methods can be more applicable to situations where an investor will seek to influence a company’s future direction—perhaps as an activist investor pushing management in new directions, or possibly as a bidder for a controlling stake in the firm. In such situations, relative valuation analysis is unlikely to provide much insight because the investor will actually be seeking to affect the company’s valuation multiples directly, by affecting the value of the denominator.

Nevertheless, even where an analyst favors the use of DCF approaches, we consider that relative valuation methods can still be valuable as a “sanity check” on the output from a DCF-based valuation. An analyst can take the

expected valuation from the DCF model and compare it with the projected values for net income, shareholders' equity, operating cash flow, and similar metrics. These ratios drawn from the DCF modeling process can then be compared with the multiples for a universe of similar firms. If the multiples generated by the analyst's DCF model are approximately comparable with the multiples that can be derived for similar companies that are already being publicly traded, then the analyst may conclude that the DCF model's assumptions appear to be reasonable. However, if the multiples from the analyst's model appear to diverge considerably from the available information concerning valuation multiples for apparently similar firms, then it may be a good idea to reexamine the model, rechecking whether the underlying assumptions are truly justifiable.

Relative valuation methods can also be useful in another way when constructing DCF models. Most DCF models include a "terminal value," which represents the expected future value of the business, discounted back to the present, from all periods subsequent to the ones for which the analyst has developed explicit estimates. One way to calculate this terminal value is in terms of a perpetual growth rate, but the choice of a particular growth rate can be difficult to justify on the basis of the firm's current characteristics. An alternative approach is to take current valuation multiples for similar firms and use those values as multiples for terminal value (see Damodaran, 2006, Chapter 4, pp. 143–144).

KEY POINTS

- Relative valuation methods tend to receive less attention from academics than DCF approaches, but such methods are widely used by practitioners. If relative valuation approaches suggest that a company is cheap on some metrics but expensive on others, this may indicate that the market views that com-

pany as being an outlier for some reason, and an analyst will probably want to investigate further.

- Choosing an appropriate group of comparable companies is perhaps the most challenging aspect of relative valuation analysis. Where possible, an analyst should seek to identify six to 12 companies that are similar in terms of size, geography, and industry. If this is not possible, then an analyst should feel free to relax one or more of these parameters in order to obtain a usable universe.
- Determining an appropriate set of valuation multiples is also important. Calculating a single set of multiples is likely to provide fewer insights than using several different metrics that span multiple time periods. It is conventional to use consensus estimates of future financial and operating performance, as these presumably represent the market's collective opinion of each firm's prospects.
- Most relative valuation analysis is performed using standard multiples such as price/earnings or firm value/sales. Under some conditions, using industry-specific multiples can be valuable, though there may be fewer consensus estimates for such data, and there may also be less intuition about what is the "fair" price for such ratios.
- Relative valuation methods are particularly useful for investors who aim to take minority stakes in individual companies when they are "cheap" relative to their peers, and then sell those stakes when the companies become "expensive." Such methods are likely to be less directly useful for investors who will seek to influence a company's management, or who aim to take a controlling stake in a company. For such investors, DCF methods are likely to be more applicable.

NOTES

1. By contrast, in an example of how to assess a small wine producer, the proposed universe of comparables consisted of 15 "beverage

firms," including both small and large caps, and covering specialists in beer, wine, and soft drink production. Arguably, some of these are unlikely to be very similar to the proposed target of analysis. See Chapter 7 in Damodaran (2006, pp. 249–252).

2. For further examples using real firms and actual figures, see Damodaran (2006, Chapters 7 and 8) or Hooke (2010, Chapter 15).

REFERENCES

- Baker, M., and Ruback, R. (1999). *Estimating Industry Multiples*. Cambridge, MA: Harvard Business School.
- Bhojraj, S., Lee, C. M. C., and Oler, D. K. (2003). What's my line? A comparison of industry classification schemes for capital market research. *Journal of Accounting Research* 41, 5: 745–774.
- Bhushan, R. (1989). Firm characteristics and analyst following. *Journal of Accounting and Economics* 11, 2–3: 255–274.
- Block, S. B. (1999). A study of financial analysts: Practice and theory. *Financial Analysts Journal* 55, 4: 86–95.
- Burch, T. R., and Nanda, V. (2003). Divisional diversity and the conglomerate discount: Evidence from spinoffs. *Journal of Financial Economics* 70, 1: 69–98.
- Chan, L. K. C., Lakonishok, J., and Swaminathan, B. (2007). Industry classifications and return comovement. *Financial Analysts Journal* 63, 6: 56–70.
- Copeland, T., Koller, T., and Murrin, J. (2000). *Valuation: Measuring and Managing the Value of Companies*, 3rd edition. New York: Wiley.
- Damodaran, A. (2006). *Damodaran on Valuation: Security Analysis for Investment and Corporate Finance*, 2nd edition. New York: Wiley.
- Hooke, J. (2010). *Security Analysis on Wall Street: A Comprehensive Guide to Today's Valuation Methods*, 2nd edition. New York: Wiley.
- Liu, J., Nissim, D., and Thomas, J. (2002). Equity valuation using multiples. *Journal of Accounting Research* 40, 1: 135–172.
- Liu, J., Nissim, D., and Thomas, J. (2007). Is cash flow king in valuations? *Financial Analysts Journal* 63, 2: 56–68.
- Phylaktis, K., and Xia, L. (2006). The changing roles of industry and country effects in the global equity markets. *European Journal of Finance* 12, 8: 627–648.
- Valentine, J. J. (2011). *Best Practices for Equity Research Analysts*. New York: McGraw-Hill.

Equity Analysis in a Complex Market

BRUCE I. JACOBS, PhD

Principal, Jacobs Levy Equity Management

KENNETH N. LEVY, CFA

Principal, Jacobs Levy Equity Management

Abstract: Investment approaches are determined by investors' views of the market. For investors who believe the market is basically efficient, so that price changes are essentially random and unpredictable, the reasonable approach is passive investing, or indexing, which makes no attempt to outperform the underlying market. Investors who believe there are clear-cut patterns discernible in stock price movements may aim for above-market returns by using fairly simple approaches, such as buying stocks with low price/earning ratios or buying small-capitalization stocks. But what if the market is not totally efficient, but there are no simple patterns that can be exploited for consistent excess returns? Such a complex market requires an investment approach capable of dealing with that complexity.

Scientists classify systems into three types—ordered, random, and complex. Ordered systems, such as the structure of diamond crystals or the dynamics of pendulums, are definable and predictable by relatively simple rules and can be modeled using a relatively small number of variables. Random systems like the Brownian motion of gas molecules or white noise (static) are unordered; they are the product of a large number of variables. Their behavior cannot be modeled and is inherently unpredictable.

Complex systems like the weather and the workings of DNA fall somewhere between the domains of order and randomness. Their behavior can be at least partly comprehended and modeled, but only with great difficulty. The number of variables that must be modeled and

their interactions are beyond the capacity of the human mind alone. Only with the aid of advanced computational science can the mysteries of complex systems be unraveled.¹

The stock market is a complex system.² Stock prices are not completely random, as the efficient market hypothesis and random walk theory would have it. Some price movements can be predicted, and with some consistency. But stock price behavior is not ordered. It cannot be successfully modeled by simple rules or screens such as low price-to-earnings ratios (P/Es) or even by elegant theories such as the capital asset pricing model or arbitrage pricing theory. Rather, stock price behavior is permeated by a complex web of *interrelated return effects*. A model of the market that is complex enough to

disentangle these effects provides opportunities for modeling price behavior and predicting returns.

This entry describes our approach to *investing* and its application to the *stock selection*, *portfolio construction*, and performance evaluation problems. We begin with the very basic question of how one should approach the *equity market*. Should one attempt to cover the broadest possible range of stocks, or can greater analytical insights be garnered by focusing on a particular subset of the market or a limited number of stocks? Each approach has its advantages and disadvantages. However, combining the two may offer the best promise of finding the key to unlocking investment opportunity in a complex market.

While covering the broadest possible range of stocks, a complex approach recognizes that there are significant differences in the ways different types of stocks respond to changes in both fundamentals and investor behavior. This requires taking into account the interrelationships between numerous potential sources of price behavior. *Multivariate analysis* disentangles the web of return-predictor relationships that constitutes a *complex market* and provides independent, additive return predictions that are more robust than the predictions from univariate analyses.

AN INTEGRATED APPROACH TO A SEGMENTED MARKET

While one might think that U.S. equity markets are fluid and fully integrated, in reality there are barriers to the free flow of capital. Some of these barriers are self-imposed by investors. Others are imposed by regulatory and tax authorities or by client guidelines.

Some funds, for example, are prohibited by regulation or internal policy guidelines from buying certain types of stock—non-dividend-paying stock, or stock below a given capitaliza-

tion level. Tax laws, too, may effectively lock investors into positions they would otherwise trade. Such barriers to the free flow of capital foster market segmentation.

Other barriers are self-imposed. Traditionally, for example, managers have focused (whether by design or default) on distinct approaches to stock selection. Value managers have concentrated on buying stocks selling at prices perceived to be low relative to the company's assets or earnings. Growth managers have sought stocks with above-average earnings growth not fully reflected in price. Small-capitalization managers have searched for opportunity in stocks that have been overlooked by most investors. The stocks that constitute the natural selection pools for these managers tend to group into distinct market segments.

Client preferences encourage this balkanization of the market. Some investors, for example, prefer to buy *value stocks*, while others seek *growth stocks*; some invest in both, but hire separate managers for each segment. Both institutional and individual investors generally demonstrate a reluctance to upset the apple cart by changing allocations to previously selected style managers. Several periods of underperformance, however, may undermine this loyalty and motivate a flow of capital from one segment of the market to another (often just as the out-of-favor segment begins to benefit from a reversion of returns back up to their historical mean).

The actions of investment consultants have formalized a market segmented into style groupings. Consultants design style indexes that define the constituent stocks of these segments and define managers in terms of their proclivity for one segment or another. As a manager's performance is measured against the given style index, managers who stray too far from index territory are taking on extra risk. Consequently, managers tend to stick close to their style homes, reinforcing market segmentation.

An investment approach that focuses on individual market segments can have its advantages. Such an approach recognizes, for example, that the U.S. equity market is neither entirely homogeneous nor entirely heterogeneous. All stocks do not react alike to a given impetus, but nor does each stock exhibit its own, totally idiosyncratic price behavior. Rather, stocks within a given style, or sector, or industry tend to behave similarly to each other and somewhat differently from stocks outside their group.

An approach to stock selection that specializes in one market segment can optimize the application of talent and maximize the potential for outperformance. This is most likely true for traditional, fundamental analysis. The in-depth, labor-intensive research undertaken by traditional analysts can become positively ungainly without some focusing lens.

An investment approach that focuses on the individual segments of the market, however, presents some theoretical and practical problems. Such an approach may be especially disadvantaged when it ignores the many forces that work to integrate, rather than segment, the market.

Many managers, for example, do not specialize in a particular market segment but are free to choose the most attractive securities from a broad universe of stocks. Others, such as style rotators, may focus on a particular type of stock, given current economic conditions, but be poised to change their focus should conditions change. Such managers make for capital flows and price arbitrage across the boundaries of particular segments.

Furthermore, all stocks can be defined by the same fundamental parameters—by market capitalization, P/E, dividend discount model ranking, and so on. All stocks can be found at some level on the continuum of values for each parameter. Thus, growth and value stocks inhabit the opposite ends of the continuums of P/E and

dividend yield, and small and large stocks the opposite ends of the continuums of firm capitalization and analyst coverage.

As the values of the parameters for any individual stock change, so too does the stock's position on the continuum. An out-of-favor growth stock may slip into value territory. A small-cap company may grow into the large-cap range.

Finally, while the values of these parameters vary across stocks belonging to different market segments—different styles, sectors, and industries—and while investors may favor certain values—low P/E, say, in preference to high P/E—arbitrage tends to counterbalance too pronounced a predilection on the part of investors for any one set of values. In equilibrium, all stocks must be owned. If too many investors want low P/E, low-P/E stocks will be bid up to higher P/E levels, and some investors will step in to sell them and buy other stocks deserving of higher P/Es. Arbitrage works toward market integration and a single pricing mechanism.

A market that is neither completely segmented nor completely integrated is a complex market. A complex market calls for an investment approach that is 180 degrees removed from the narrow, segment-oriented focus of traditional management. It requires a complex, *unified approach* that takes into account the behavior of stocks across the broadest possible selection universe, without losing sight of the significant differences in price behavior that distinguish particular market segments.

Such an approach offers three major advantages. First, it provides a coherent evaluation framework. Second, it can benefit from all the insights to be garnered from a wide and diverse range of securities. Third, because it has both breadth of coverage and depth of analysis, it is poised to take advantage of more profit opportunities than a more narrowly defined, segmented approach proffers.

A Coherent Framework

To the extent that the market is integrated, an investment approach that models each industry or style segment as if it were a universe unto itself is not the best approach. Consider, for example, a firm that offers both core and value strategies. Suppose the firm runs a model on its total universe of, say, 3,000 stocks. It then runs the same model or a different, segment-specific model on a 500-stock subset of large-cap value stocks.

If different models are used for each strategy, the results will differ. Even if the same model is estimated separately for each strategy, its results will differ because the model coefficients are bound to differ between the broader universe and the narrower segment. What if the core model predicts GM will outperform Ford, while the value model shows the reverse? Should the investor start the day with multiple estimates of one stock's alpha? This would violate what we call the *law of one alpha*.³

Of course, the firm could ensure coherence by using separate models for each market segment—growth, value, small-cap, linking the results via a single, overarching model that relates all the subsets. But the firm then runs into a second problem with segmented investment approaches: To the extent that the market is integrated, the pricing of securities in one segment may contain information relevant to pricing in other segments.

For example, within a generally well-integrated national economy, labor market conditions in the United States differ region by region. An economist attempting to model employment in the Northeast would probably consider economic expansion in the Southeast. Similarly, the investor who wants to model growth stocks should not ignore value stocks. The effects of inflation, say, on value stocks may have repercussions for growth stocks; after all, the two segments represent opposite ends of the same P/E continuum.

An investment approach that concentrates on a single market segment does not make use of

all available information. A complex, unified approach considers all the stocks in the universe, value and growth, large and small. It thus benefits from all the information to be gleaned from a broad range of stock price behavior.

Of course, an increase in breadth of inquiry will not benefit the investor if it comes at the sacrifice of depth of inquiry. A complex approach does not ignore the significant differences across different types of stock, differences exploitable by specialized investing. What's more, in examining similarities and differences across market segments, it considers numerous variables that may be considered to be defining.

For value, say, a complex approach does not confine itself to a dividend discount model measure of value, but examines also earnings, cash flow, sales, and yield value, among other attributes. Growth measurements to be considered include historical, expected, and sustainable growth, as well as the momentum and stability of earnings. Share price, volatility, and analyst coverage are among the elements to be considered along with market capitalization as measures of size.

At a deeper level of analysis, one must also consider alternative ways of specifying such fundamental variables as earnings or cash flow. Over what period does one measure earnings? If using analyst earnings expectations, which measure provides the best estimate of future real earnings? The consensus of all available estimates made over the past six months, or only the very latest earnings estimates? Are some analysts more accurate or more influential? What if a recent estimate is not available for a given company?⁴

Predictor variables are often closely correlated with each other. *Small-cap stocks*, for example, tend to have low P/Es; low P/E is correlated with high yield; both low P/E and high yield are correlated with dividend discount model (DDM) estimates of value. Furthermore, they may be correlated with a stock's industry affiliation. A simple low-P/E screen, for example, will tend to select a large number

of bank and utility stocks. Such correlations can distort naïve attempts to relate returns to potentially relevant predictors. A true picture of the return-predictor relationship emerges only after *disentangling* the predictors.

DISENTANGLING

The effects of different sources of stock return can overlap. In Figure 1, the lines represent connections documented by academic studies; they may appear like a ball of yarn after the cat got to it. To unravel the connections between predictor variables and return, it is necessary to examine all the variables simultaneously.

For instance, the low-P/E effect is widely recognized, as is the small-size effect. But stocks with low P/Es also tend to be of small size. Are P/E and size merely two ways of looking at the same effect? Or does each variable matter? Perhaps the excess returns to small-cap stocks are merely a January effect, reflecting the tendency of taxable investors to sell depressed stocks at year-end. Answering these questions requires disentangling return effects via multivariate regression.⁵

Common methods of measuring return effects (such as quintiling or univariate, single-variable, regression) are naïve because they assume, naïvely, that prices are responding only to the single variable under consideration, low

P/E, say. But a number of related variables may be affecting returns. As we have noted, small-cap stocks and banking and utility industry stocks tend to have low P/Es. A univariate regression of return on low P/E will capture, along with the effect of P/E, a great deal of noise related to firm size, industry affiliation, and other variables.

Simultaneous analysis of all relevant variables via multivariate regression takes into account and adjusts for such interrelationships. The result is the return to each variable separately, controlling for all related variables. A multivariate analysis for low P/E, for example, will provide a measure of the excess return to a portfolio that is market-like in all respects except for having a lower-than-average P/E ratio. Disentangled returns are *pure returns*.

Noise Reduction

Figure 2 plots naïve and pure cumulative monthly excess (relative to a 3,000-stock universe) returns to high book-to-price ratio (B/P). (Conceptually, naïve and pure returns come from a portfolio having a B/P that is one standard deviation above the universe mean B/P; for the pure returns, the portfolio is also constrained to have universe-average exposures to all the other variables in the model, including fundamental characteristics and industry affiliations.) The *naïve returns* show a great deal of volatility; the pure returns, by contrast, follow a much smoother path. There is a lot of noise in the naïve returns. What causes it?

Notice the divergence between the naïve and pure return series for the 12 months starting in March 1979. This date coincides with the crisis at Three Mile Island nuclear power plant. Utilities such as GPU, operator of the Three Mile Island power plant, tend to have high B/Ps, and naïve B/P measures will reflect the performance of these utilities along with the performance of other high-B/P stocks. Electric utility prices plummeted 24% after the Three Mile Island crisis. The naïve B/P measure reflects this decline.

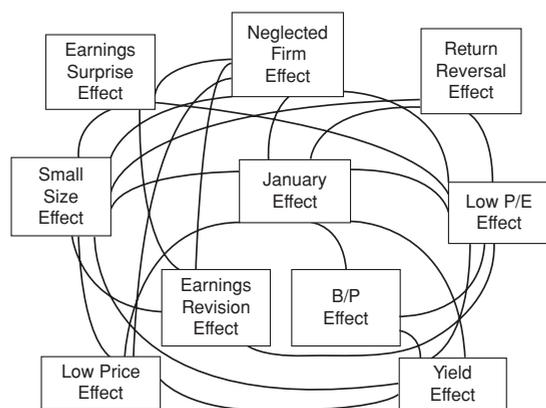


Figure 1 Return Effects Form a Tangled Web

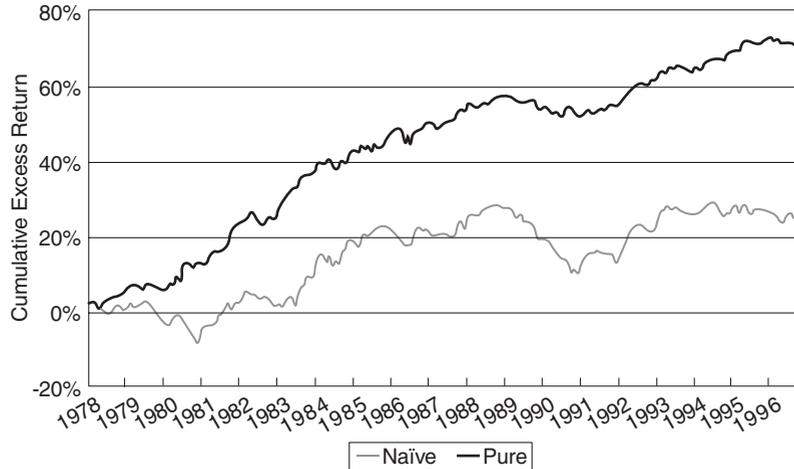


Figure 2 Naïve and Pure Returns to High Book-to-Price Ratio

But industry-related events such as Three Mile Island have no necessary bearing on the B/P variable. An investor could, for example, hold a high-B/P portfolio that does not overweight utilities, and such a portfolio would not have experienced the decline reflected in the naïve B/P measure in Figure 2. The naïve returns to B/P reflect noise from the inclusion of a utility industry effect. A pure B/P measure is not contaminated by such irrelevant variables.

Disentangling distinguishes real effects from mere proxies and thereby distinguishes between real and spurious investment opportunities. As it separates high B/P and industry affiliation, for example, it can also separate the effects of firm size from the effects of related variables. Disentangling shows that returns to small firms in January are not abnormal; the apparent January seasonal merely proxies for year-end tax-loss selling.⁶ Not all small firms will benefit from a January rebound; indiscriminately buying small firms at the turn of the year is not an optimal investment strategy. Ascertaining true causation leads to more profitable strategies.

Return Revelation

Disentangling can reveal hidden opportunities. Figure 3 plots the naïvely measured cumulative

monthly excess returns (relative to the 3,000-stock universe) to portfolios that rank lower than average in market capitalization and price per share and higher than average in terms of analyst neglect. These results derive from monthly univariate regressions. The small-cap line thus represents the cumulative excess returns to a portfolio of stocks naïvely chosen on the basis of their size, with no attempt made to control for other variables.

All three return series move together. The similarity between the small-cap and neglect series is particularly striking. This is confirmed by the correlation coefficients in the first column of Table 1. Furthermore, all series show a great deal of volatility within a broader up, down, up pattern.

Figure 4 shows the pure cumulative monthly excess returns to each size-related attribute over the period. These disentangled returns adjust for correlations not only between the three size

Table 1 Correlations Between Monthly Returns to Size-Related Variables*

Variable	Naïve	Pure
Small cap/low price	0.82	-0.12
Small cap/neglect	0.87	-0.22
Neglect/low price	0.66	-0.11

*A coefficient of 0.14 is significant at the 5% level.

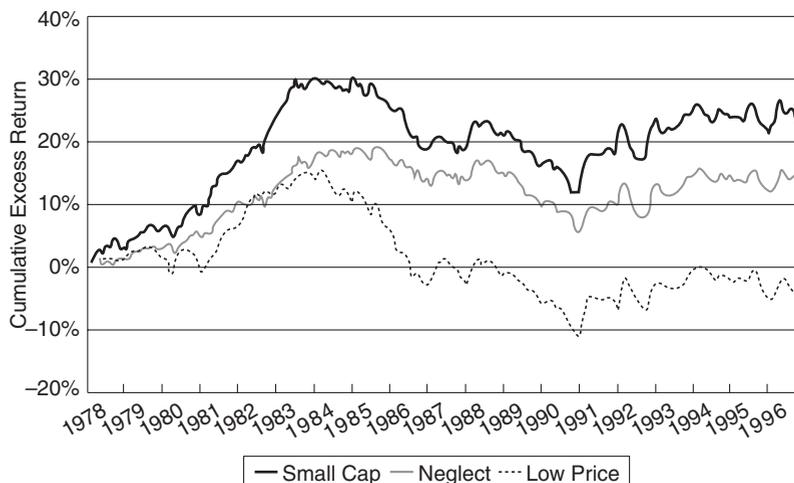


Figure 3 Naïve Returns Can Hide Opportunities (Three Size-Related Variables)

variables, but also between each size variable and industry affiliations and each variable and growth and value characteristics. Two findings are immediately apparent from Figure 4.

First, pure returns to the size variables do not appear to be nearly as closely correlated as the naïve returns displayed in Figure 3. In fact, over the second half of the period, the three return series diverge substantially. This is confirmed by the correlation coefficients in the second column of Table 1.

In particular, pure returns to small capitalization accumulate quite a gain over the pe-

riod; they are up 30%, versus an only 20% gain for the naïve returns to small cap. Purifying returns reveals a profit opportunity not apparent in the naïve returns. Furthermore, pure returns to analyst neglect amount to a substantial loss over the period. Because disentangling controls for proxy effects, and thereby avoids redundancies, these pure return effects are additive. A portfolio could have aimed for superior returns by selecting small-cap stocks with a higher-than-average analyst following (that is, a negative exposure to analyst neglect).

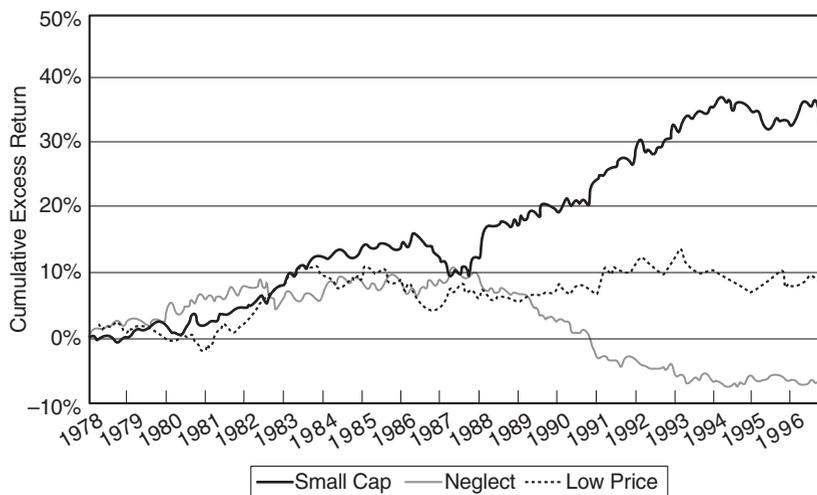


Figure 4 Pure Returns Can Reveal Opportunities (Three Size-Related Variables)

Table 2 Pure Returns Are Less Volatile, More Predictable: Standard Deviations of Monthly Returns to Size-Related Variables*

Variable	Naïve	Pure
Small cap	0.87	0.60
Neglect	0.87	0.67
Low price	1.03	0.58

*All differences between naïve and pure return standard deviations are significant at the 1% level.

Second, the pure returns appear to be much less volatile than the naïve returns. The naïve returns in Figure 3 display much month-to-month volatility within their more general trends. By contrast, the pure series in Figure 4 are much smoother and more consistent. This is confirmed by the standard deviations given in Table 2.

The pure returns in Figure 4 are smoother and more consistent than the naïve return responses in Figure 3 because the pure returns capture more signal and less noise. And because they are smoother and more consistent than naïve returns, pure returns are also more predictive.

Predictive Power

Disentangling improves the predictive power of estimated returns by providing a clearer picture of the relationships between investor behavior, fundamental variables, and macroeconomic conditions. For example, investors often prefer value stocks in bearish market environments, because growth stocks are priced more on the basis of high expectations, which get dashed in more pessimistic eras. But the success of such a strategy will depend on the variables one has chosen to define value.

Table 3 displays the results of regressing both naïve and pure monthly returns to various value-related variables on market (S&P 500) returns over the 1978–1996 period.⁷ The results indicate that DDM value is a poor indicator of a stock's ability to withstand a tide of receding market prices. The regression coeffi-

Table 3 Market Sensitivities of Monthly Returns to Value-Related Variables

Variable	Naïve	(<i>t</i> -stat.)	Pure	(<i>t</i> -stat.)
DDM	0.06	(5.4)	0.04	(5.6)
B/P	−0.10	(−6.2)	−0.01	(−0.8)
Yield	−0.08	(−7.4)	−0.03	(−3.5)

cient in the first column indicates that a portfolio with a one-standard-deviation exposure to DDM value will tend to outperform by 0.06% when the market rises by 1.00% and to underperform by a similar margin when the market falls by 1.00%. The coefficient for pure returns to DDM is similar. Whether their returns are measured in pure or naïve form, stocks with high DDM values tend to behave procyclically.

High B/P appears to be a better indicator of a defensive stock. It has a regression coefficient of −0.10 in naïve form. In pure form, however, B/P is virtually uncorrelated with market movements; pure B/P signals neither an aggressive nor a defensive stock. B/P as naïvely measured apparently picks up the effects of truly defensive variables, such as high yield.

The value investor in search of a defensive posture in uncertain market climates should consider moving toward high yield. The regression coefficients for both naïve and pure returns to high yield indicate significant negative market sensitivities. Stocks with high yields may be expected to lag in up markets but to hold up relatively well during general market declines.

These results make broad intuitive sense. DDM is forward-looking, relying on estimates of future earnings. In bull markets, investors take a long-term outlook, so DDM explains security pricing behavior. In bear markets, however, investors become myopic; they prefer today's tangible income to tomorrow's promise. Current yield is rewarded.

Pure returns respond in intuitively satisfying ways to macroeconomic events. Figure 5 illustrates, as an example, the estimated effects of changes in various macroeconomic variables on the pure returns to small size (as measured by

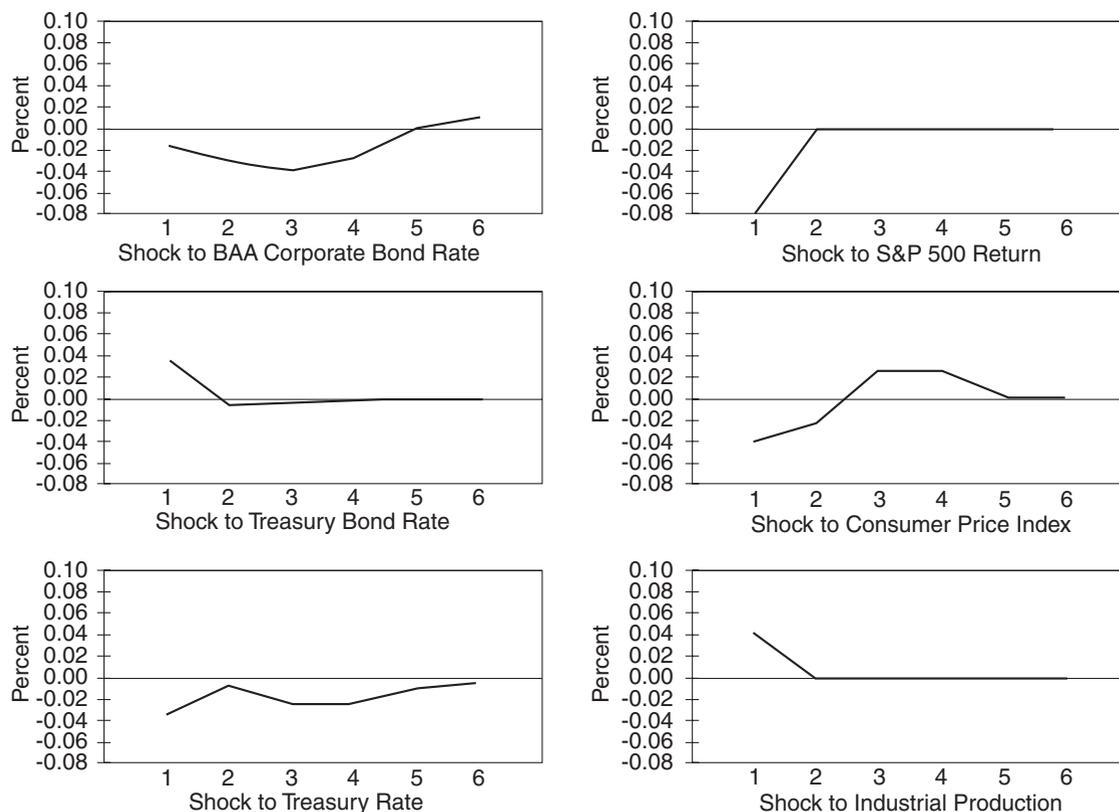


Figure 5 Forecast Response of Small Size to Macroeconomic Shocks

market capitalization). Consistent with the capital constraints on small firms and their relatively greater sensitivity to the economy, pure returns to small size may be expected to be negative in the first four months following an unexpected increase in the Baa corporate rate and positive in the first month following an unexpected increase in industrial production.⁸ Investors can exploit such predictable behavior by moving into and out of the small-cap market segment as economic conditions evolve.⁹

These examples serve to illustrate that the use of numerous, finely defined fundamental variables can provide a rich representation of the complexity of security pricing. The model can be even more finely tuned, however, by including variables that capture such subtleties as the effects of investor psychology, possible non-

linearities in variable-return relationships, and security transaction costs.

Additional Complexities

In considering possible variables for inclusion in a model of stock price behavior, the investor should recognize that pure stock returns are driven by a combination of economic fundamentals and investor psychology. That is, economic fundamentals such as interest rates, industrial production, and inflation can explain much, but by no means all, of the systematic variation in returns. Psychology, including investors' tendency to overreact, their desire to seek safety in numbers, and their selective memories, also plays a role in security pricing.

What's more, the modeler should realize that the effects of different variables, fundamental

and otherwise, can differ across different types of stocks. The value sector, for example, includes more financial stocks than the growth sector. Investors may thus expect value stocks in general to be more sensitive than growth stocks to changes in interest rate spreads.

Psychologically based variables such as short-term overreaction and price correction also seem to have a stronger effect on value than on growth stocks. Earnings surprises and earnings estimate revisions, by contrast, appear to be more important for growth than for value stocks. Thus, Google shares can take a nosedive when earnings come in a penny under expectations, whereas Duke Energy shares remain unmoved even by fairly substantial departures of actual earnings from expectations.

The relationship between stock returns and relevant variables may not be linear. The effects of positive earnings surprises, for instance, tend to be arbitrated away quickly; thus positive earnings surprises offer less opportunity for the investor. The effects of negative earnings surprises, however, appear to be more long-lasting. This nonlinearity may reflect the fact that sales of stock are limited to those investors who already own the stock (and to a relatively small number of short-sellers).¹⁰

Risk-variable relationships may also differ across different types of stock. In particular, small-cap stocks generally have more idiosyncratic risk than large-cap stocks. Diversification is thus more important for small-stock than for large-stock portfolios.

Return-variable relationships can also change over time. Recall the difference between DDM and yield value measures: high-DDM stocks tend to have high returns in bull markets and low returns in bear markets; high-yield stocks experience the reverse. For consistency of performance, return modeling must consider the effects of market dynamics, the changing nature of the overall market.

The investor may also want to decipher the informational signals generated by informed agents. Corporate decisions to issue or buy back

shares, split stock, or initiate or suspend dividends, for example, may contain valuable information about company prospects. So, too, may insiders' (legal) trading in their own firms' shares.

Finally, a complex model containing multiple variables is likely to turn up a number of promising return-variable relationships. But are these perceived profit opportunities translatable into real economic opportunities? Are some too ephemeral? Too small to survive frictions such as trading costs? Estimates of expected returns must be combined with estimates of the costs of trading to arrive at realistic returns net of trading costs.

CONSTRUCTING, TRADING, AND EVALUATING PORTFOLIOS

To maximize implementation of the model's insights, the portfolio construction process should consider exactly the same dimensions found relevant by the stock selection model. Failure to do so can lead to mismatches between model insights and portfolio exposures.¹¹

Consider a commercially available *portfolio optimizer* that recognizes only a subset of the variables in the valuation model. Risk reduction using such an optimizer will reduce the portfolio's exposures only along the dimensions the optimizer recognizes. As a result, the portfolio is likely to wind up more exposed to those variables recognized by the model, but not the optimizer, and less exposed to those variables common to both the model and the optimizer.

Imagine an investor who seeks low-P/E stocks that analysts are recommending for purchase, but who uses a commercial optimizer that incorporates a P/E factor but not analyst recommendations. The investor is likely to wind up with a portfolio that has a less-than-optimal level of exposure to low P/E and a greater-than-optimal level of exposure to

analyst purchase recommendations. Optimization using all relevant variables ensures a portfolio whose risk and return opportunities are balanced in accordance with the model's insights. Furthermore, the use of more numerous variables allows portfolio risk to be more finely tuned.

Insofar as the investment process, both stock selection and portfolio construction, is model-driven, it is more adaptable to electronic trading venues. This should benefit the investor in several ways. First, electronic trading is generally less costly, with lower commissions, market impact, and opportunity costs. Second, it allows real-time monitoring, which can further reduce trading costs. Third, an automated trading system can take account of more factors, including the urgency of a particular trade and market conditions, than individual traders can be expected to bear in mind.

Finally, the *performance attribution* process should be congruent with the dimensions of the selection model (and portfolio optimizer). Insofar as performance attribution identifies sources of return, a process that considers all the sources identified by the selection model will be more insightful than a commercial performance attribution system applied in a one-size-fits-all manner. Our investor who has sought exposure to low P/E and positive analyst recommendations, for example, will want to know how each of these factors has paid off and will be less interested in the returns to factors that are not a part of the stock selection process.

A performance evaluation process tailored to the model also functions as a monitor of the model's reliability. Has portfolio performance supported the model's insights? Should some be reexamined? Equally important, does the model's reliability hold up over time? A model that performs well in today's economic and market environments may not necessarily perform well in the future. A feedback loop between the evaluation and the research/modeling processes can help ensure that the model retains robustness over time.

PROFITING FROM COMPLEXITY

H. L. Mencken is supposed to have noted, "For every complex problem, there is a simple solution, and it is almost always wrong." Complex problems more often than not require complex solutions.

A complex approach to stock selection, portfolio construction, and performance evaluation is needed to capture the complexities of the stock market. Such an approach combines the breadth of coverage and the depth of analysis needed to maximize investment opportunity and potential reward.

Grinold presents a formula that identifies the relationships between the depth and breadth of investment insights and investment performance:¹²

$$IR = IC\sqrt{BR}$$

IR is the manager's information ratio, a measure of the success of the investment process. IR equals annualized excess return over annualized residual risk (e.g., 2% excess return with 4% tracking error provides 0.5 IR). IC, the information coefficient, or correlation between predicted and actual security returns, measures the goodness of the manager's insights, or the manager's skill. BR is the breadth of the strategy, measurable as the number of independent insights upon which investment decisions are made.

One can increase IR by increasing IC or BR. Increasing IC means coming up with some means of improving predictive accuracy. Increasing BR means coming up with more "investable" insights. A casino analogy may be apt (if anathema to prudent investors).

A gambler can seek to increase IC by card counting in blackjack or by building a computer model to predict probable roulette outcomes. Similarly, some investors seek to outperform by concentrating their research efforts on a few stocks: by learning all there is to know about Microsoft, for example, one may be able to

outperform all the other investors who follow this stock. But a strategy that makes a few concentrated stock bets is likely to produce consistent performance only if it is based on a very high level of skill, or if it benefits from extraordinary luck.

Alternatively, an investor can place a larger number of smaller stock bets and settle for more modest returns from a greater number of investment decisions. That is, rather than behaving like a gambler in a casino, the investor can behave like the casino. A casino has only a slight edge on any spin of the roulette wheel or roll of the dice, but many spins of many roulette wheels can result in a very consistent profit for the house. Over time, the odds will strongly favor the casino over the gambler.

A complex approach to the equity market, one that has both breadth of inquiry and depth of focus, can enhance the number and the goodness of investment insights. A complex approach to the equity market requires more time, effort, and ability, but it will be better positioned to capture the complexities of security pricing. The rewards are worth the effort.

KEY POINTS

- Ordered systems are definable and predictable by relatively simple rules; random systems cannot be modeled and are inherently unpredictable; complex systems can be at least partly comprehended and modeled, but only with difficulty.
- Stock price behavior is permeated by a complex web of interrelated return effects, and it requires a complex approach to stock selection, portfolio construction, and performance evaluation to capture this complexity.
- A complex approach combines the breadth of coverage and the depth of analysis needed to maximize investment opportunity and potential reward.
- Simple methods of measuring return effects (such as quintiling or univariate, single-variable regression) are naïve because they

assume that prices are responding only to the single variable under consideration.

- Simultaneous analysis of all relevant variables via multivariate regression takes into account and adjusts for interrelationships between effects, giving the return to each variable separately.
- Disentangling distinguishes real effects from mere proxies and thereby distinguishes between real and spurious investment opportunities.
- Because disentangling controls for proxy effects, pure return effects are additive, each having the potential to improve portfolio performance.
- In general, disentangling enhances the predictive power of estimated returns by providing a clearer picture of the relationships between investor behavior, fundamental variables, and macroeconomic conditions.
- To maximize implementation of insights gained from disentangling the market's complexity, the portfolio construction process should consider exactly the same dimensions found relevant by the stock selection process.
- Performance attribution should be congruent with the stock selection and portfolio construction processes so that it can be used to monitor the reliability of the stock selection process and provide input for research.

NOTES

1. See Pagels (1988) and Wolfram (2002).
2. Jacobs and Levy (1989a).
3. See Jacobs and Levy (1995b).
4. See Jacobs, Levy, and Krask (1997).
5. See Jacobs and Levy (1988b).
6. Jacobs and Levy (1988a).
7. Jacobs and Levy (1988c).
8. See Jacobs and Levy (1989b).
9. Jacobs and Levy (1996).
10. See Jacobs and Levy (1993).
11. See Jacobs and Levy (1995a).
12. Grinold (1989).

REFERENCES

- Grinold, R. C. (1989). The fundamental law of active management. *Journal of Portfolio Management* 15, 3: 30–37.
- Jacobs, B. I., and Levy, K. N. (1988a). Calendar anomalies: Abnormal returns at calendar turning points. *Financial Analysts Journal* 44, 6: 28–39.
- Jacobs, B. I., and Levy, K. N. (1988b). Disentangling equity return regularities: New insights and investment opportunities. *Financial Analysts Journal* 44, 3: 18–44.
- Jacobs, B. I., and Levy, K. N. (1988c). On the value of 'value.' *Financial Analysts Journal* 44, 4: 47–62.
- Jacobs, B. I., and Levy, K. N. (1989a). The complexity of the stock market. *Journal of Portfolio Management* 16, 1: 19–27.
- Jacobs, B. I., and Levy, K. N. (1989b) Forecasting the size effect. *Financial Analysts Journal* 45, 3: 38–54.
- Jacobs, B. I., and Levy, K. N. (1993). Long/short equity investing. *Journal of Portfolio Management* 20, 1: 52–63.
- Jacobs, B. I., and Levy, K. N. (1995a). Engineering portfolios: A unified approach. *Journal of Investing* 4, 4: 8–14.
- Jacobs, B. I., and Levy, K. N. (1995b). The law of one alpha. *Journal of Portfolio Management* 21, 4: 78–79.
- Jacobs, B. I., and Levy, K. N. (1996). High definition style rotation. *Journal of Investing* 6, 1: 14–23.
- Jacobs, B. I., Levy, K. N., and Krask, M. C. (1997). Earnings estimates, predictor specification, and measurement error. *Journal of Investing* 6, 4: 29–46.
- Pagels, H. (1988). *The Dreams of Reason: The Computer and the Rise of the Sciences of Complexity*. New York: Simon & Schuster.
- Wolfram, S. (2002). *A New Kind of Science*. Champaign, IL: Wolfram Media Inc.

Equity Portfolio Selection Models in Practice

DESSISLAVA A. PACHAMANOVA, PhD

Associate Professor of Operations Research, Babson College

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Quantitative equity portfolio selection often involves extending the classical mean-variance framework or more advanced tail-risk portfolio allocation frameworks to include different constraints that take specific investment guidelines and institutional features into account. Examples of such constraints are holding constraints that set limits on the total concentration of assets in an industry, sector, or country; turnover constraints that restrict the amount of trading; tracking error constraints that limit the difference between the performance of the portfolio and a benchmark; and risk factor constraints that limit the exposure of the portfolio to a risk factor such as the market. Portfolio allocation models can also account for transaction costs, taxes, and optimization of trades across multiple client accounts. An important practical issue in quantitative equity portfolio selection is how to mitigate the effect of model and estimation errors on the optimal allocation. Techniques that are used to address this issue include robust statistical techniques for parameter estimation, portfolio resampling, and robust optimization.

An integrated investment process generally involves the following activities:¹

1. An investor's objectives, preferences, and constraints are identified and specified to develop explicit investment policies.
2. Strategies are developed and implemented through the choice of optimal combinations of financial and real assets in the marketplace.
3. Market conditions, relative asset values, and the investor's circumstances are monitored.

4. Portfolio adjustments are made as appropriate to reflect significant changes in any or all of the relevant variables.

In this entry we focus on the second activity of the investment process, developing and implementing a portfolio strategy. The development of the portfolio strategy itself is typically done in two stages: First, funds are allocated among asset classes. Then, they are managed within the asset classes. The mean-variance framework is used at both stages,

but in this entry, we discuss the second stage. Specifically, we introduce quantitative formulations of portfolio allocation problems used in equity portfolio management. Quantitative equity portfolio selection often involves extending the classical mean-variance framework or more advanced tail-risk portfolio allocation frameworks to include different constraints that take specific investment guidelines and institutional features into account.

We begin by providing a classification of the most common *portfolio constraints* used in practice. We then discuss extensions such as index tracking formulations, the inclusion of transaction costs, optimization of trades across multiple client accounts, and tax-aware strategies. We conclude with a review of methods for incorporating robustness in quantitative portfolio allocation procedures by using robust statistics, simulation, and robust optimization techniques.

PORTFOLIO CONSTRAINTS COMMONLY USED IN PRACTICE

Institutional features and investment policy specifications often lead to more complicated requirements than simple minimization of risk (whatever the definition of risk may be) or maximization of expected portfolio return. For instance, there can be constraints that limit the number of trades, the exposure to a specific industry, or the number of stocks to be kept in the portfolio. Some of these constraints are imposed by the clients, while others are imposed by regulators. For example, in the case of regulated investment companies, restrictions on asset allocation are set forth in the prospectus and may be changed only with the approval of the fund's board of directors. Pension funds must comply with Employee Retirement Income Security Act (ERISA) requirements. The objective of the *portfolio optimization* problem can also be

modified to consider specifically the trade-off between risk and return, transactions costs, or taxes.

In this section, we will take a single-period view of investing, in the sense that the goal of the portfolio allocation procedure will be to invest optimally over a single predetermined period of interest, such as one month.² We will use \mathbf{w}_0 to denote the vector array of stock weights in the portfolio at the beginning of the period, and \mathbf{w} to denote the weights at the end of the period (to be determined).

Many investment companies, especially institutional investors, have a long investment horizon. However, in reality, they treat that horizon as a sequence of shorter period horizons. Risk budgets are often stated over a time period of a year, and return performance is monitored quarterly or monthly.

Long-Only (No-Short-Selling) Constraints

Many funds and institutional investors face restrictions or outright prohibitions on the amount of short selling they can do. When short selling is not allowed, the portfolio allocation optimization model contains the constraints $\mathbf{w} \geq \mathbf{0}$.

Holding Constraints

Diversification principles argue against investing a large proportion of the portfolio in a single asset, or having a large concentration of assets in a specific industry, sector, or country. Limits on the holdings of a specific stock can be imposed with the constraints

$$\mathbf{l} \leq \mathbf{w} \leq \mathbf{u}$$

where \mathbf{l} and \mathbf{u} are vectors of lower and upper bounds of the holdings of each stock in the portfolio.

Consider now a portfolio of 10 stocks. Suppose that the issuers of assets 1, 3, and 5 are in

the same industry, and that we would like to limit the portfolio exposure to that industry to be at least 20% but at most 40%. To limit exposure to that industry, we add the constraint

$$0.20 \leq w_1 + w_3 + w_5 \leq 0.40$$

to the portfolio allocation optimization problem.

More generally, if we have a specific set of stocks I_j out of the investment universe I consisting of stocks in the same category (such as industry or country), we can write the constraint

$$L_j \leq \sum_{j \in I_j} w_j \leq U_j$$

In words, this constraint requires that the sum of all stock weights in the particular category of investments with indexes I_j is greater than or equal to a lower bound L_j and less than or equal to a maximum exposure of U_j .

Turnover Constraints

High portfolio turnover can result in large transaction costs that make portfolio rebalancing inefficient and costly. Thus, some portfolio managers limit the amount of turnover allowed when trading their portfolio. (Another way to control for transaction costs is to minimize them explicitly; we will discuss the appropriate formulations later in this entry.)

Most commonly, turnover constraints are imposed for each stock:

$$|w_i - w_{0,i}| \leq u_i,$$

that is, the absolute magnitude of the difference between the final and the initial weight of stock i in the portfolio is restricted to be less than some upper bound u_i . Sometimes, a constraint is imposed to minimize the portfolio turnover as a whole:

$$\sum_{j \in I_j} |w_j - w_{0,j}| \leq U_j$$

that is, the total absolute difference between the initial and the final weights of the stocks in the portfolio is restricted to be less than or equal to an upper bound U_j . Under this constraint, some stock weights may deviate a lot more than others from their initial weights, but the total deviation is limited.

Turnover constraints are often imposed relative to the *average daily volume* (ADV) of a stock.³ For example, we may want to restrict turnover to be no more than 5% of the ADV. (In the latter case, the upper bound u_i is set to a value equal to 5% of the ADV.) Modifications of these constraints, such as limiting turnover in a specific industry or sector, are also frequently applied.

Risk Factor Constraints

In practice, it is very common for quantitatively oriented portfolio managers to use factor models to control for risk exposures to different risk factors. Such risk factors could include the market return, size, and style. Let us assume that the return on stock i has a factor structure with K risk factors, that is, it can be expressed through the equality

$$r_i = \alpha_i + \sum_{k=1}^K \beta_{ik} \cdot f_k + \varepsilon_i$$

The factors f_k are common to all securities. The coefficient β_{ik} in front of each factor f_k shows the sensitivity of the return on stock i to factor k . The value of α_i shows the expected excess return of the return on stock i , and ε_i is the idiosyncratic (called “nonsystematic”) part of the return of stock i . The coefficients α_i and β_{ik} are typically estimated by multiple regression analysis.

To limit the exposure of a portfolio of N stocks to the k th risk factor, we impose the constraint

$$\sum_{i=1}^N \beta_{ik} \cdot w_i \leq U_k$$

To understand this constraint, note that the total return on the portfolio can be written as

$$\begin{aligned} \sum_{i=1}^N w_i \cdot r_i &= \sum_{i=1}^N w_i \cdot (\alpha_i + \sum_{k=1}^K \beta_{ik} \cdot f_k + \varepsilon_i) \\ &= \sum_{i=1}^N w_i \cdot \alpha_i + \sum_{i=1}^N \left(w_i \cdot \left(\sum_{k=1}^K \beta_{ik} \cdot f_k \right) \right) \\ &\quad + \sum_{i=1}^N w_i \cdot \varepsilon_i \end{aligned}$$

The sensitivity of the portfolio to the different factors is represented by the second term, which can be also written as

$$\sum_{k=1}^K \left(\left(\sum_{i=1}^N w_i \cdot \beta_{ik} \right) \cdot f_k \right)$$

Therefore, the exposure to a particular factor k is the coefficient in front of f_k , that is,

$$\sum_{i=1}^N \beta_{ik} \cdot w_i$$

On an intuitive level, the sensitivity of the portfolio to a factor k will be larger the larger the presence of factor k in the portfolio through the exposure of the individual stocks. Thus, when we compute the total exposure of the portfolio to factor k , we need to take into consideration both how important this factor is for determining the return on each of the securities in the portfolio, and how much of each security we have in the portfolio.

A commonly used version of the maximum factor exposure constraint is

$$\sum_{i=1}^N \beta_{ik} \cdot w_i = 0$$

This constraint forces the portfolio optimization algorithm to find portfolio weights so that the overall risk exposure to factor k is 0, that is, so that the portfolio is neutral with respect to changes in factor k . Portfolio allocation strategies that claim to be “market-neutral” typically employ this constraint, and the factor is in fact the return on the market.

Cardinality Constraints

Depending on the portfolio allocation model used, sometimes the optimization subroutine recommends holding small amounts of a large number of stocks, which can be costly when one takes into consideration the transaction costs incurred when acquiring these positions. Alternatively, a portfolio manager may be interested in limiting the number of stocks used to track a particular index. (We will discuss index tracking later in this entry.) To formulate the constraint on the number of stocks to be held in the portfolio (called the cardinality constraint), we introduce binary variables, one for each of the N stocks in the portfolio. Let us call these binary variables $\delta_1, \dots, \delta_N$. Variable δ_i will take value 1 if stock i is included in the portfolio, and 0 otherwise.

Suppose that out of the N stocks in the investment universe, we would like to include a maximum of K stocks in the final portfolio. K here is a positive integer and is less than N . This constraint can be formulated as

$$\sum_{i=1}^N \delta_i \leq K$$

$$\delta_i \text{ binary, } i = 1, \dots, N$$

We need to make sure, however, that if a stock is not selected in the portfolio, then the binary variable that corresponds to that stock is set to 0, so that the stock is not counted as one of the K stocks left in the portfolio. When the portfolio weights are restricted to be nonnegative, this can be achieved by imposing the additional constraints

$$0 \leq w_i \leq \delta_i, \quad i = 1, \dots, N$$

If the optimal weight for stock i turns out to be different from 0, then the binary variable δ_i associated with stock i is forced to take value 1, and stock i will be counted as one of the K stocks to be kept in the portfolio. If the optimal weight for stock i is 0, then the binary variable δ_i associated with stock i can be either 0 or 1, but that will not matter for all practical

purposes, because the solver will set it to 0 if there are too many other attractive stocks that will be counted as the K stocks to be kept in the portfolio. At the same time, since the portfolio weights w_i are between 0 and 1, and δ_i is 0 or 1, the constraint $w_i \leq \delta_i$ does not restrict the values that the stock weight w_i can take.

The constraints are a little different if short sales are allowed, in which case the weights may be negative. We have

$$-M \cdot \delta_i \leq w_i \leq M \cdot \delta_i, \quad i = 1, \dots, N$$

where M is a “large” constant (large relative to the size of the inputs in the problem; so in this portfolio optimization application $M = 10$ can be considered “large”). You can observe that if the weight w_i is anything but 0, the value of the binary variable δ_i will be forced to be different from 0, that is, δ_i will need to be 1, since it can only take values 0 or 1.

Minimum Holding and Transaction Size Constraints

Cardinality constraints are often used in conjunction with minimum holding/trading constraints. The latter set a minimum limit on the amount of a stock that can be held in the portfolio, or the amount of a stock that can be traded, effectively eliminating small trades. Both cardinality and minimum holding/trading constraints aim to reduce the amount of transaction costs.

Threshold constraints on the amount of stock i to be held in the portfolio can be imposed with the constraint

$$|w_i| \geq L_i \cdot \delta_i$$

where L_i is the smallest holding size allowed for stock i , and δ_i is a binary variable, analogous to the binary variables δ_i defined in the previous section—it equals 1 if stock i is included in the portfolio, and 0 otherwise. (All additional constraints relating δ_i and w_i described in the previous section still apply.)

Similarly, constraints can be imposed on the minimum trading amount for stock i . As we ex-

plained earlier in this section, the size of the trade for stock i is determined by the absolute value of the difference between the current weight of the stock, $w_{0,i}$, and the new weight w_i that will be found by the solver: $|w_i - w_{0,i}|$. The minimum trading size constraint formulation is

$$|w_i - w_{0,i}| \geq L_i^{\text{trade}} \cdot \delta_i$$

where L_i^{trade} is the smallest trading size allowed for stock i .

Adding binary variables to an optimization problem makes the problem more difficult for the solver and can increase the computation time substantially. That is why in practice, portfolio managers often omit minimum holding and transaction size constraints from the optimization problem formulation, selecting instead to eliminate weights and/or trades that appear too small manually, after the optimal portfolio is determined by the optimization solver. It is important to realize, however, that modifying the optimal solution for the simpler portfolio allocation problem (the optimal solution in this case is the weights/trades for the different stocks) by eliminating small positions manually does not necessarily produce the optimal solution to an optimization problem that contained the minimum holding and transaction size constraints from the beginning. In fact, there can be pathological cases in which the solution is very different from the true optimal solution. However, for most cases in practice, the small manual adjustments to the optimal portfolio allocation do not cause tremendous discrepancies or inconsistencies.

Round Lot Constraints

So far, we have assumed that stocks are infinitely divisible, that is, that we can trade and invest in fractions of stocks, bonds, and so on. This is, of course, not true—in reality, securities are traded in multiples of minimum transaction lots, or rounds (e.g., 100 or 500 shares).

In order to represent the condition that securities should be traded in rounds, we need to introduce additional decision variables (let us

call them z_i , $i = 1, \dots, N$) that are integers and will correspond to the number of lots of a particular security that will be purchased. Each z_i will then be linked to the corresponding portfolio weight w_i through the equality

$$w_i = z_i \cdot f_i, \quad i = 1, \dots, N$$

where f_i is measured in dollars, and is a fraction of the total amount to be invested. For example, suppose there is a total of \$100 million to be invested, and stock i trades at \$50 in round lots of 100. Then

$$f_i = \frac{50 \cdot 100}{100,000,000} = 5 \cdot 10^{-7}$$

All remaining constraints in the portfolio allocation can be expressed through the weights w_i , as usual. However, we also need to specify for the solver that the decision variables z_i are integers.

An issue with imposing round lot constraints is that the budget constraint

$$\mathbf{w}'\mathbf{t} = 1$$

which is in fact

$$\sum_{i=1}^N z_i \cdot f_i = 1$$

may not be satisfied exactly. One possibility to handle this problem is to relax the budget constraint. For example, we can state the constraint as

$$\mathbf{w}'\mathbf{t} \leq 1$$

or, equivalently,

$$\sum_{i=1}^N z_i \cdot f_i \leq 1$$

This will ensure that we do not go over budget.

If our objective is stated as expected return maximization, the optimization solver will attempt to make this constraint as tight as possible, that is, we will end up using up as much of the budget as we can. Depending on the objective function and the other constraints in the formulation, however, this may not always happen. We can try to force the solver to minimize

the slack in the budget constraint by introducing a pair of nonnegative decision variables (let us call them ε^+ and ε^-) that account for the amount that is “overinvested” or “underinvested.” These variables will pick up the slack left over because of the inability to round the amounts for the different investments. Namely, we impose the constraints

$$\begin{aligned} \sum_{i=1}^N z_i \cdot f_i + \varepsilon^- - \varepsilon^+ &= 1 \\ \varepsilon^- \geq 0, \varepsilon^+ &\geq 0 \end{aligned}$$

and subtract the following term from the objective function:

$$\lambda_{\text{rl}} \cdot (\varepsilon^- + \varepsilon^+)$$

where λ_{rl} is a penalty term associated with the amount of over- or underinvestment the portfolio manager is willing to tolerate (selected by the portfolio manager). In the final solution, the violation of the budget constraint will be minimized. Note, however, that this formulation technically allows for the budget to be overinvested.

The optimal portfolio allocation we obtain after solving this optimization problem will not be the same as the allocation we would obtain if we solve an optimization problem without round lot constraints, and then round the amounts to fit the lots that can be traded in the market.

Cardinality constraints, minimum holding/trading constraints, and especially round lot constraints require more sophisticated binary and integer programming solvers, and are difficult problems to solve in the case of large portfolios.

BENCHMARK EXPOSURE AND TRACKING ERROR MINIMIZATION

Expected portfolio return maximization under the mean-variance framework or other risk measure minimization are examples of *active*

investment strategies, that is, strategies that identify a universe of attractive investments and ignore inferior investments opportunities. A different approach, referred to as a *passive investment strategy*, argues that in the absence of any superior forecasting ability, investors might as well resign themselves to the fact that they cannot beat the market. From a theoretical perspective, the analytics of portfolio theory tell them to hold a broadly diversified portfolio anyway. Many mutual funds are managed relative to a particular benchmark or stock universe, such as the S&P 500 or the Russell 1000. The portfolio allocation models are then formulated in such a way that the tracking error relative to the benchmark is kept small.

Standard Definition of Tracking Error

To incorporate a passive investment strategy, we can change the objective function of the portfolio allocation problem so that instead of minimizing a portfolio risk measure, we minimize the tracking error with respect to a benchmark that represents the market, such as the Russell 3000, or the S&P 500. Such strategies are often referred to as *indexing*. The *tracking error* can be defined in different ways. However, practitioners typically mean a specific definition: the variance (or standard deviation) of the difference between the portfolio return, $\mathbf{w}'\tilde{\mathbf{r}}$, and the return on the benchmark, $\mathbf{w}_b'\tilde{\mathbf{r}}$. Mathematically, the tracking error (TE) can be expressed as

$$\begin{aligned} \text{TE} &= \text{Var}(\mathbf{w}'\tilde{\mathbf{r}} - \mathbf{w}_b'\tilde{\mathbf{r}}) \\ &= \text{Var}((\mathbf{w} - \mathbf{w}_b)'\tilde{\mathbf{r}}) \\ &= (\mathbf{w} - \mathbf{w}_b)'\text{Var}(\tilde{\mathbf{r}})(\mathbf{w} - \mathbf{w}_b) \\ &= (\mathbf{w} - \mathbf{w}_b)'\boldsymbol{\Sigma}(\mathbf{w} - \mathbf{w}_b) \end{aligned}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the stock returns. One can observe that the formula is very similar to the formula for the portfolio variance; however, the portfolio weights in the formula for the variance are replaced by differences between the weights of the stocks in the portfolio and the weights of the stocks in the index.

Why do we need to optimize portfolio weights in order to track a benchmark, when technically the most effective way to track a benchmark is by investing the portfolio in the stocks in the benchmark portfolio in the same proportions as the proportions of these securities in the benchmark? The problem with this approach is that, especially with large benchmarks like the Russell 3000, the transaction costs of a proportional investment and the subsequent rebalancing of the portfolio can be prohibitive (that is, dramatically adversely impact the performance of the portfolio relative to the benchmark). Furthermore, in practice securities are not infinitely divisible, so investing a portfolio of a limited size in the same proportions as the composition of the benchmark will still not achieve zero tracking error. Thus, the optimal formulation is to require that the portfolio follows the benchmark as closely as possible.

While indexing has become an essential part of many portfolio strategies, most portfolio managers cannot resist the temptation to identify at least some securities that will outperform others. Hence, restrictions on the tracking error are often imposed as a constraint, while the objective function is something different from minimizing the tracking error. The tracking error constraint takes the form

$$(\mathbf{w} - \mathbf{w}_b)'\boldsymbol{\Sigma}(\mathbf{w} - \mathbf{w}_b) \leq \sigma_{\text{TE}}^2$$

where σ_{TE}^2 is a limit (imposed by the investor) on the amount of tracking error the investor is willing to tolerate. This is a quadratic constraint, which is convex and computationally tractable, but requires specialized optimization software.

Alternative Ways of Defining Tracking Error

There are alternative ways in which tracking-error type constraints can be imposed.

For example, we may require that the absolute deviations of the portfolio weights (\mathbf{w}) from the index weights (\mathbf{w}_b) are less than or equal to a

given vector array of upper bounds \mathbf{u} :

$$|\mathbf{w} - \mathbf{w}_b| \leq \mathbf{u}$$

where the absolute values $|\cdot|$ for the vector differences are taken componentwise, that is, for pairs of corresponding elements from the two vector arrays. These constraints can be stated as linear constraints by rewriting them as

$$\begin{aligned} \mathbf{w} - \mathbf{w}_b &\leq \mathbf{u} \\ -(\mathbf{w} - \mathbf{w}_b) &\leq \mathbf{u} \end{aligned}$$

Similarly, we can require that for stocks within a specific industry (whose indexes in the portfolio belong to a subset I_j of the investment universe I), the total tracking error is less than a given upper bound U_j :

$$\sum_{j \in I_j} (w_j - w_{b,j}) \leq U_j$$

Finally, tracking error can be expressed through risk measures other than the absolute deviations or the variance of the deviations from the benchmark. Rockafellar and Uryasev (2000) suggest using conditional value-at-risk (CVaR) to manage the tracking error. Conditional value-at-risk measures the average loss that can happen with probability less than some small probability, that is, the average loss in the tail of the distribution of portfolio losses. (Using CVaR as a risk measure results in computationally tractable optimization formulations for portfolio allocation, as long as the data are presented in the form of scenarios.⁴) We provide below a formulation that is somewhat different from Rockafellar and Uryasev, but preserves the main idea.

Suppose that we are given S scenarios for the return of a benchmark portfolio (or an instrument we are trying to replicate), $b_s, s = 1, \dots, S$. These scenarios can be generated by simulation or taken from historical data. We also have N stocks with returns $r_i^{(s)} (i = 1, \dots, N, s = 1, \dots, S)$ in each scenario. The value of the portfolio in scenario s is

$$\sum_{i=1}^N r_i^{(s)} \cdot w_i$$

or, equivalently, $(\mathbf{r}^{(s)})' \mathbf{w}$, where $\mathbf{r}^{(s)}$ is the vector of returns for the N stocks in scenario s . Consider the differences between the return on the benchmark and the return on the portfolio,

$$b_s - (\mathbf{r}^{(s)})' \mathbf{w} = -((\mathbf{r}^{(s)})' \mathbf{w} - b_s)$$

If this difference is positive, we have a loss; if the difference is negative, we have a gain; both gains and losses are computed relative to the benchmark. Rationally, the portfolio manager should not worry about differences that are negative; the only cause for concern would be if the portfolio underperforms the benchmark, which would result in a positive difference. Thus, it is not necessary to limit the variance of the deviations of the portfolio returns from the benchmark, which penalizes for positive and negative deviations equally. Instead, we can impose a limit on the amount of loss we are willing to tolerate in terms of the CVaR of the distribution of losses relative to the benchmark.

The tracking error constraint in terms of the CVaR can be stated as the following set of constraints:⁵

$$\begin{aligned} \xi + \frac{1}{[\varepsilon \cdot S]} \cdot \sum_{s=1}^S y_s &\leq U_{TE} \\ y_s &\geq -\left((\mathbf{r}^{(s)})' \mathbf{w} - b_s\right) - \xi, \quad s = 1, \dots, S \\ y_s &\geq 0, \quad s = 1, \dots, S \end{aligned}$$

where U_{TE} is the upper bound on the negative deviations.

This formulation of tracking error is appealing in two ways. First, it treats positive and negative deviations relative to the benchmark differently, which agrees with the strategy of an investor seeking to maximize returns overall. Second, it results in a linear set of constraints, which are easy to handle computationally, in contrast to the first formulation of the tracking error constraint in this section, which results in a quadratic constraint.

Actual Versus Predicted Tracking Error

The tracking error calculation in practice is often backward-looking. For example, in computing the covariance matrix Σ in the standard tracking error definition as the variance of the deviations of the portfolio returns from the index, or in selecting the scenarios used in the CVaR-type tracking error constraint in the previous section, we may use historical data. The tracking error calculated in this manner is called the *ex post* tracking error, backward-looking error, or actual tracking error.

The problem with using the actual tracking error for assessing future performance relative to a benchmark is that the actual tracking error does not reflect the effect of the portfolio manager's current decisions on the future active returns and hence the tracking error that may be realized in the future. The actual tracking error has little predictive value and can be misleading regarding portfolio risk.

Portfolio managers need forward-looking estimates of tracking error to reflect future portfolio performance more accurately. In practice, this is accomplished by using the services of a commercial vendor that has a multifactor risk model that has identified and defined the risks associated with the benchmark, or by building such a model in-house. Statistical analysis of historical return data for the stocks in the benchmark is used to obtain the risk factors and to quantify the risks. Using the manager's current portfolio holdings, the portfolio's current exposure to the various risk factors can be calculated and compared to the benchmark's exposures to the risk factors. From the differential factor exposures and the risks of the factors, a forward-looking tracking error for the portfolio can be computed. This tracking error is also referred to as an *ex ante* tracking error or predicted tracking error.

There is no guarantee that the predicted tracking error will match exactly the tracking error realized over the future time period of interest. However, this calculation of the tracking error

has its use in risk control and portfolio construction. By performing a simulation analysis on the factors that enter the calculation, the manager can evaluate the potential performance of portfolio strategies relative to the benchmark, and eliminate those that result in tracking errors beyond the client-imposed tolerance for risk. The actual tracking error, on the other hand, is useful for assessing actual performance relative to a benchmark.

INCORPORATING TRANSACTION COSTS

Transaction costs can be generally divided into two categories: (1) explicit such as bid-ask spreads, commissions, and fees, and (2) implicit such as price movement risk costs and market impact costs. Price movement risk costs are the costs resulting from the potential for a change in market price between the time the decision to trade is made and the time the trade is actually executed. Market impact is the effect a trader has on the market price of an asset when it sells or buys the asset. It is the extent to which the price moves up or down in response to the trader's actions. For example, a trader who tries to sell a large number of shares of a particular stock may drive down the stock's market price.

The typical portfolio allocation models are built on top of one or several forecasting models for expected returns and risk. Small changes in these forecasts can result in reallocations that would not occur if transaction costs are taken into account. In practice, the effect of transaction costs on portfolio performance is far from insignificant. If transaction costs are not taken into consideration in allocation and rebalancing decisions, they can lead to poor portfolio performance.

This section describes some common transaction cost models for portfolio rebalancing. We use the mean-variance framework as the basis for describing the different approaches. However, it is straightforward to extend the

transaction cost models into other portfolio allocation frameworks.

The earliest, and most widely used, model for transaction costs is the mean-variance risk-aversion formulation with transaction costs.⁶ The optimization problem has the following objective function:

$$\max_{\mathbf{w}} \mathbf{w}'\boldsymbol{\mu} - \lambda \cdot \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} - \lambda_{\text{TC}} \cdot \text{TC}$$

where TC is a transaction cost penalty function and λ_{TC} is the transaction cost aversion parameter. In other words, the objective is to maximize the expected portfolio return less the cost of risk and transaction costs. We can imagine that as the transaction costs increase, at some point it becomes optimal to keep the current portfolio rather than to rebalance. Variations of this formulation exist. For example, it is common to maximize expected portfolio return minus transaction costs, and to impose limits on the risk as a constraint (i.e., to move the second term in the objective function to the constraints).

Transaction costs models can involve complicated nonlinear functions. Although software exists for general nonlinear optimization problems, the computational time required for solving such problems is often too long for realistic investment applications, and the quality of the solution is not guaranteed. In practice, an observed complicated nonlinear transaction costs function is often approximated with a computationally tractable function that is assumed to be separable in the portfolio weights, that is, it is often assumed that the transaction costs for each individual stock are independent of the transaction costs for another stock. For the rest of this section, we will denote the individual cost function for stock i by TC_i .

Next, we explain several widely used models for the transaction cost function.

Linear Transaction Costs

Let us start simple. Suppose that the transaction costs are proportional, that is, they are a percentage c_i of the transaction size $|t| =$

$|w_i - w_{0,i}|$.⁷ Then, the portfolio allocation problem with transaction costs can be written simply as

$$\max_{\mathbf{w}} \mathbf{w}'\boldsymbol{\mu} - \lambda \cdot \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} - \lambda_{\text{TC}} \cdot \sum_{i=1}^N c_i \cdot |w_i - w_{0,i}|$$

The problem can be made solver-friendly by replacing the absolute value terms with new decision variables y_i , and adding two sets of constraints. Hence, we rewrite the objective function as

$$\max_{\mathbf{w}, \mathbf{y}} \mathbf{w}'\boldsymbol{\mu} - \lambda \cdot \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} - \lambda_{\text{TC}} \cdot \sum_{i=1}^N c_i \cdot y_i$$

and add the constraints

$$\begin{aligned} y_i &\geq w_i - w_{0,i} \\ y_i &\geq -(w_i - w_{0,i}) \end{aligned}$$

This preserves the quadratic optimization problem formulation, a formulation that can be passed to quadratic optimization solvers such as Excel Solver and MATLAB's `quadprog` function, because the constraints are linear expressions, and the objective function contains only linear and quadratic terms.

In the optimal solution, the optimization solver will in fact set the value for y_i to $|w_i - w_{0,i}|$. This is because this is a maximization problem and y_i occurs with a negative sign in the objective function, so the solver will try to set y_i to the minimum value possible. That minimum value will be the maximum of $(w_i - w_{0,i})$ or $-(w_i - w_{0,i})$, which is in fact the absolute value $|w_i - w_{0,i}|$.

Piecewise-Linear Transaction Costs

Taking the model in the previous section a step further, we can introduce piecewise-linear approximations to transaction cost function models. This kind of function is more realistic than the linear cost function, especially for large trades. As the trading size increases, it becomes increasingly more costly to trade because of the market impact of the trade.

An example of a piecewise-linear function of transaction costs for a trade of size t of a

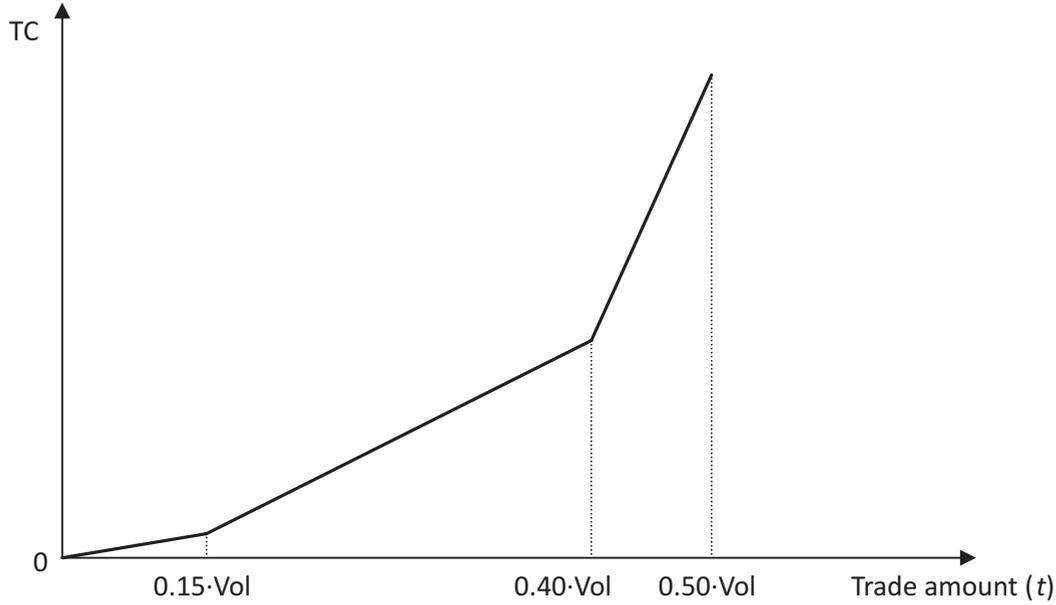


Figure 1 Example of Modeling Transaction Costs (TC) as a Piecewise-Linear Function of Trade Size t

particular security is illustrated in Figure 1. The transaction cost function in the graph assumes that the rate of increase of transaction costs (reflected in the slope of the function) changes at certain threshold points. For example, it is smaller in the range 0 to 15% of daily volume than in the range 15% to 40% of daily volume (or some other trading volume index). Mathematically, the transaction cost function in Figure 1 can be expressed as

$$TC(t) = \begin{cases} s_1 t, & 0 \leq t \leq 0.15 \cdot \text{Vol} \\ s_1(0.15 \cdot \text{Vol}) + s_2(t - 0.15 \cdot \text{Vol}), & 0.15 \cdot \text{Vol} \leq t \leq 0.40 \cdot \text{Vol} \\ s_1(0.15 \cdot \text{Vol}) + s_2(0.25 \cdot \text{Vol}) + s_3(t - 0.40 \cdot \text{Vol}), & 0.40 \cdot \text{Vol} \leq t \leq 0.50 \cdot \text{Vol} \end{cases}$$

where s_1, s_2, s_3 are the slopes of the three linear segments on the graph. (They are given data.)

To include piecewise-linear functions for transaction costs in the objective function of a mean-variance (or any general mean-risk) portfolio optimization problem, we need to introduce new decision variables that correspond to the number of pieces in the piecewise-linear approximation of the transaction cost function (in this case, there are three linear segments, so

we introduce variables z_1, z_2, z_3). We write the penalty term in the objective function for an individual stock as⁸

$$\lambda_{TC} \cdot (s_1 \cdot z_1 + s_2 \cdot z_2 + s_3 \cdot z_3)$$

If there are N stocks in the portfolio, the total transaction cost will be the sum of the transaction costs for each individual stock, that is, the penalty term that involves transaction costs in the objective function becomes

$$-\lambda_{TC} \sum_{i=1}^N (s_{1,i} \cdot z_{1,i} + s_{2,i} \cdot z_{2,i} + s_{3,i} \cdot z_{3,i})$$

In addition, we specify the following constraints on the new decision variables:

$$\begin{aligned} 0 &\leq z_{1,i} \leq 0.15 \cdot \text{Vol}_i \\ 0 &\leq z_{2,i} \leq 0.25 \cdot \text{Vol}_i \\ 0 &\leq z_{3,i} \leq 0.10 \cdot \text{Vol}_i \end{aligned}$$

Note that because of the increasing slopes of the linear segments and the goal of making that term as small as possible in the objective function, the optimizer will never set the decision variable corresponding to the second segment, $z_{2,i}$, to a number greater than 0 unless the decision variable corresponding to the first segment, $z_{1,i}$, is at its upper bound. Similarly,

the optimizer would never set $z_{3,i}$ to a number greater than 0 unless both $z_{1,i}$ and $z_{2,i}$ are at their upper bounds. So, this set of constraints allows us to compute the amount of transaction costs incurred in the trading of stock i as $z_{1,i} + z_{2,i} + z_{3,i}$.

Of course, we also need to link the amount of transaction costs incurred in the trading of stock i to the optimal portfolio allocation. This can be done by adding a few more variables and constraints. We introduce variables y_i , one for each stock in the portfolio, that would represent the amount traded (but not the direction of the trade) and would be nonnegative. Then, we require that

$$y_i = z_{1,i} + z_{2,i} + z_{3,i} \quad \text{for each stock } i,$$

and also that y_i equals the change in the portfolio position of stock i . The latter condition can be imposed by writing the constraint

$$y_i = |w_i - w_{0,i}|$$

where $w_{0,i}$ and w_i are the initial and the final amount of stock i in the portfolio, respectively.⁹

Despite their apparent complexity, piecewise-linear approximations for transaction costs are very solver-friendly, and save time (relative to nonlinear models) in the actual portfolio optimization. Although modeling transaction costs this way requires introducing new decision variables and constraints, the increase in the dimension of the portfolio optimization problem does not affect significantly the running time or the performance of the optimization solver, because the problem formulation is easy from a computational perspective.

Quadratic Transaction Costs

The transaction cost function is often parameterized as a quadratic function of the form

$$\text{TC}_i(t) = c_i \cdot |t| + d_i \cdot |t|^2$$

The coefficients c_i and d_i are calibrated from data—for example, by fitting a quadratic function to an observed pattern of transaction costs

realized for trading a particular stock under normal conditions.

Including this function in the objective function of the portfolio optimization problem results in a quadratic program that can be solved with widely available quadratic optimization software.

Fixed Transaction Costs

In some cases, we need to model fixed transaction costs. Those are costs that are incurred independently of the amount traded. To include such costs in the portfolio optimization problem, we need to introduce binary variables $\delta_1, \dots, \delta_N$ corresponding to each stock, where δ_i equals 0 if the amount traded of stock i is 0, and 1 otherwise. The idea is similar to the idea we used to model the requirement that only a given number of stocks can be included in the portfolio.

Suppose the fixed transaction cost is a_i for stock i . Then, the transaction cost function is

$$\text{TC}_i = a_i \cdot \delta_i$$

The objective function formulation is then

$$\max_{\mathbf{w}, \delta} \mathbf{w}'\boldsymbol{\mu} - \lambda \cdot \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} - \lambda_{\text{TC}} \cdot \sum_{i=1}^N a_i \cdot \delta_i$$

and we need to add the following constraints to make sure that the binary variables are linked to the trades $|w_i - w_{0,i}|$:

$$|w_i - w_{0,i}| \leq M \cdot \delta_i, \quad i = 1, \dots, N,$$

δ_i binary

where M is a “large” constant. When the trading size $|w_i - w_{0,i}|$ is nonzero, δ_i will be forced to be 1. When the trading size is 0, then δ_i can be either 0 or 1, but the optimizer will set it to 0, since it will try to make its value the minimum possible in the objective function.

Of course, combinations of different trading cost models can be used in practice. For example, if the trade involves both a fixed and a variable quadratic transaction cost, then we could

use a transaction cost function of the kind

$$TC_i(t) = a_i \cdot \delta_i + c_i \cdot |t| + d_i \cdot |t|^2$$

The important takeaway from this section is that when transaction costs are included in the portfolio rebalancing problem, the result is a reduced amount of trading and rebalancing, and a different portfolio allocation than the one that would be obtained if transaction costs are not taken into consideration.

INCORPORATING TAXES

When stocks in a portfolio appreciate or depreciate in value, capital gains (respectively, losses) accumulate. When stocks are sold, investors pay *taxes* on the realized net capital gains. The taxes are computed as a percentage of the difference between the current market value of the stocks and their tax basis, where the tax basis is the price at which the stocks were bought originally.¹⁰ The percentage is less for long-term capital gains (when stocks have been held for more than a year) than it is for short-term capital gains (when stocks have been held for less than a year).¹¹ Since shares of the same stock could have been bought at different points in time (in different lots), selling one lot of the stock as opposed to another could incur a different amount of tax. In addition to capital gains taxes, investors who are not exempt from taxes owe taxes on the dividends paid on stocks in their portfolios. Those dividends are historically taxed at a higher rate than capital gains, and may eventually be taxed as income, that is, at the investor's personal tax rate. The tax liability of a particular portfolio therefore depends on the timing of the execution of trades, on the tax basis of the portfolio, on the accumulated short-term and long-term capital gains, and on the tax bracket of the investor.

Over two-thirds of marketable portfolio assets in the United States are held by individuals, insurance, and holding companies who pay taxes on their returns. (Exceptions are, for ex-

ample, pension funds, which do not pay taxes year-to-year.) Studies have indicated that taxes are the greatest expense investors face—greater than commissions and investment management fees. To gain some intuition about the effect of taxes on the income of an investor over the investor's lifetime, consider a portfolio that has a capital appreciation of 6.00% per year. After 30 years, \$1,000 invested in that portfolio will turn into $\$1,000 \cdot (1 + 0.06)^{30} = \$5,743.49$. Now suppose that the capital gains are realized each year, and a tax of 35% is paid on the gains (the remainder is reinvested). After 30 years, \$1,000 invested in the portfolio will turn into $\$1,000 \cdot (1 + (1 - 0.35) \cdot 0.06)^{30} = \$3,151.13$, about half of the amount without taxes even when the tax is about one third of the capital gains. In fact, in order to provide the same return as the portfolio with no taxes, the portfolio with annual realized capital gains would need to generate a capital appreciation of 9.23% per year! One can imagine that the same logic would make benchmark tracking and performance measurement very difficult on an after-tax basis.

As investors have become more aware of the dramatic impact of taxes on their returns, there is increasing pressure on portfolio managers to include tax considerations in their portfolio rebalancing decisions and to report after-tax performance. Consequently, the demand for computationally efficient and quantitatively rigorous methods for taking taxes into consideration in portfolio allocation decisions has grown in recent years. The complexity of the problem of incorporating taxes, however, is considerable, both from a theoretical and practical perspective:

1. The presence of tax liabilities changes the interpretation of even fundamental portfolio performance summary measures such as market value and risk. Thus, well-established methods for evaluating portfolio performance on a pretax basis do not work well in the case of tax-aware portfolio optimization. For example, in traditional

portfolio management a loss is associated with risk and is therefore minimized whenever possible. However, in the presence of taxes, losses may be less damaging, because they can be used to offset capital gains and reduce the tax burden of portfolio rebalancing strategies. Benchmarking is also not obvious in the presence of taxes: Two portfolios that have exactly the same current holdings are not equivalent if the holdings have a different tax basis.¹²

2. Tax considerations are too complex to implement in a nonautomated fashion; at the same time, their automatic inclusion in portfolio rebalancing algorithms requires the ability to solve very difficult, large-scale optimization problems.
3. The best approach for portfolio management with tax considerations is optimization problem formulations that look at return forecasts over several time periods (e.g., until the end of the year) before recommending new portfolio weights. However, the latter multiperiod view of the portfolio optimization problem is very difficult to handle computationally—the dimension of the optimization problem, that is, the number of variables and constraints, increases exponentially with the number of time periods under consideration.

We need to emphasize that while many of the techniques described in the previous sections of this entry are widely known, there are no standard practices for tax-aware portfolio management that appear to be established. Different asset management firms interpret tax-aware portfolio allocation and approach the problem differently. To some firms, minimizing turnover,¹³ for example, by investing in index funds, or selecting strategies that minimize the portfolio dividend yield,¹⁴ qualify as tax-aware portfolio strategies. Other asset management firms employ complex optimization algorithms that incorporate tax considerations directly in portfolio rebalancing decisions, so

that they can keep up with the considerable burden of keeping track of thousands of managed accounts and their tax preferences. The fact is, even using simple rules of thumb, such as always selling stocks from the oldest lots after rebalancing the portfolio with classical portfolio optimization routines, can have a positive effect on after-tax portfolio returns. The latter strategy minimizes the likelihood that short-term gains will be incurred, which in turn reduces taxes, because short-term capital gains are taxed at a higher rate than long-term capital gains.

Apelfeld, Fowler, and Gordon (1996) suggest a tax-aware portfolio rebalancing framework that incorporates taxes directly into the portfolio optimization process. The main idea of the approach is to treat different lots of the same stock as different securities, and then penalize for taxes as if they were different transaction costs associated with the sale of each lot. (This means, for example, that Microsoft stock bought on Date 1 is treated as a different security from Microsoft stock bought on Date 2.) Many tax-aware quantitative investment strategies employ versions of this approach, but there are a few issues to beware of when using it in practice:

- The first one is a general problem for all tax-aware approaches when they are used in the context of active portfolio management. For a portfolio manager who handles thousands of different accounts with different tax exposures, it is virtually impossible to pay attention to the tax cost incurred by each individual investor. While the tax-aware method described above minimizes the overall tax burden by reducing the amount of realized short-term sales, it has no provisions for differentiating between investors in different tax brackets because it is difficult to think of each trade as divided between all investors, and adjusted for each individual investor's tax circumstances. This issue is so intractable that in practice it is not really brought under consideration.

- The dimension of the problem can become unmanageable very quickly. For example, a portfolio of 1,000 securities, each of which has 10 different lots, is equivalent to a portfolio of 10,000 securities when each lot is treated as a different security. Every time a new purchase is realized, a new security is added to the portfolio, since a new lot is created. One needs to exercise care and “clean up” lots that have been sold and therefore have holdings of zero each time the portfolio is rebalanced.
- Practitioners typically use factor models for forecasting returns and estimating risk. One of the assumptions when measuring portfolio risk through factor models is that the specific risk of a particular security is uncorrelated with the specific risk of other securities. (The only risk they share is the risk expressed through the factors in the factor model.) This assumption clearly does not hold when different “securities” are in fact different lots of the same stock.

DiBartolomeo (2000) describes a modification to the model used by Northfield Information Service’s portfolio management software that eliminates the last two problems. Instead of treating each lot as a separate security, the software imposes piecewise-linear transaction costs (see Figure 1) where the break points on the horizontal axis correspond to the current size of different lots of the same security. The portfolio rebalancing algorithm goes through several iterations for the portfolio weights, and at each iteration, only the shares in the highest cost basis tax lot can be traded. Other shares of the same stock can be traded in subsequent iterations of the algorithm, with their appropriate tax costs attached.

The approaches we described so far take into consideration the short-term or long-term nature of capital gains, but do not incorporate the ability to offset capital gains and losses accumulated over the year. This is an inherent limitation of single-period portfolio rebalancing approaches and is a strong argument in fa-

vor of adopting more realistic multiperiod portfolio optimization approaches. The rebalancing of the portfolio at each point in time should be made not only by considering the immediate consequences for the market value of the portfolio, but also the opportunity to correct for tax liabilities by realizing other capital gains or losses by the end of the taxable year. The scarce theoretical literature on multiperiod tax-aware portfolio optimization contains some characterizations of optimal portfolio strategies under numerous simplifying assumptions.¹⁵ However, even under such simplifying assumptions, the dimension of the problem grows exponentially with the number of stocks in a portfolio, and it is difficult to come up with computationally viable algorithms for portfolios of realistic size.

MULTIACCOUNT OPTIMIZATION

Portfolio managers who handle multiple accounts face an important practical issue. When individual clients’ portfolios are managed, portfolio managers incorporate their clients’ preferences and constraints. However, on any given trading day, the necessary trades for multiple diverse accounts are pooled and executed simultaneously. Moreover, typically trades may not be crossed, that is, it is not simply permissible to transfer an asset that should be sold on behalf of one client into the account of another client for whom the asset should be bought.¹⁶ The trades should be executed in the market. Thus, each client’s trades implicitly impact the results for the other clients: The market impact of the combined trades may be such that the benefits sought for individual accounts through trading are lost due to increased overall transaction costs. A robust multiaccount management process should ensure accurate accounting and fair distribution of transaction costs among the individual accounts.

One possibility to handle the effect of trading in multiple accounts is to use an iterative

process, in which at each iteration the market impact of the trades in previous iterations is taken into account.¹⁷ More precisely, single clients' accounts are optimized as usual, and once the optimal allocations are obtained, the portfolio manager aggregates the trades and computes the actual marginal transaction costs based on the aggregate level of trading. The portfolio manager then reoptimizes individual accounts using these marginal transaction costs, and aggregates the resulting trades again to compute new marginal transaction costs, and so on. The advantage of this approach is that little needs to be changed in the way individual accounts are typically handled, so the existing single-account optimization and management infrastructure can be reused. The disadvantage is that most generally, this iterative approach does not guarantee a convergence (or its convergence may be slow) to a "fair equilibrium," in which clients' portfolios receive an unbiased treatment with respect to the size and the constraint structure of their accounts.¹⁸ The latter equilibrium is the one that would be attained if all clients traded independently and competitively in the market for liquidity, and it is thus the correct and fair solution to the aggregate trading problem.

An alternative, more comprehensive approach is to optimize trades across all accounts simultaneously. O'Connell, Scherer, and Xu (2006) describe such a model and show that it attains the fair equilibrium we mentioned above.¹⁹ Assume that client k 's utility function is given by u_k and is in the form of a dollar return penalized for risk. Assume also that a transaction cost model τ gives the cost of trading in dollars, and that τ is a convex increasing function.²⁰ Its exact form will depend on the details of how trading is implemented. Let \mathbf{t} be the vector of trades. It will typically have the form $(t_1^+, \dots, t_N^+, t_1^-, \dots, t_N^-)$, that is, it will specify the aggregate buys t_i^+ and the aggregate sells t_i^- for each asset $i=1, \dots, N$, but it may also incorporate information about how the trade could be carried out.²¹

The multiaccount optimization problem can be formulated as

$$\begin{aligned} \max_{\mathbf{w}_1, \dots, \mathbf{w}_K, \mathbf{t}} \quad & E[u_1(\mathbf{w}_1)] + \dots + E[u_K(\mathbf{w}_K)] - \tau(\mathbf{t}) \\ \text{s.t.} \quad & \mathbf{w}_k \in C_k, k = 1, \dots, K \end{aligned}$$

where \mathbf{w}_k is the N -dimensional vector of asset holdings (or weights) of client k , and C_k is the collection of constraints on the portfolio structure of client k . The objective can be interpreted as maximization of net expected utility, that is, as maximization of the expected dollar return penalized for risk and net of transaction costs.

The problem can be simplified by making some reasonable assumptions. For example, it can be assumed that the transaction cost function τ is additive across different assets, that is, that trades in one asset do not influence trading costs in another. In such a case, the trading cost function can be split into more manageable terms, that is,

$$\tau(\mathbf{t}) = \sum_{i=1}^N \tau_i(t_i^+, t_i^-)$$

where $\tau_i(t_i^+, t_i^-)$ is the cost of trading asset i as a function of the aggregate buys and sells of that asset. Splitting the terms $\tau_i(t_i^+, t_i^-)$ further into separate costs of buying and selling, however, is not a reasonable assumption, because simultaneous buying and selling of an asset tends to have an offsetting effect on its price.

To formulate the problem completely, let \mathbf{w}_k^0 be the vector of original holdings (or weights) of client k 's portfolio, \mathbf{w}_k be the vector of decision variables for the optimal holdings (or weights) of client k 's portfolio, and $\eta_{k,i}$ be constants that convert the holdings (or weight) of each asset i in client k 's portfolio $w_{k,i}$ to dollars, that is, $\eta_{k,i}w_{k,i}$ is client k 's dollar holdings of asset i .²² We also introduce new variables \mathbf{w}_k^+ to represent the an upper bound on the weight of each asset client k will buy:

$$w_{k,i} - w_{k,i}^0 \leq w_{k,i}^+, \quad i = 1, \dots, N, k = 1, \dots, K$$

The aggregate amount of asset i bought for all clients can then be computed as

$$t_i^+ = \sum_{k=1}^K \eta_{k,i} \cdot w_{k,i}^+$$

The aggregate amount of asset i sold for all clients can be easily expressed by noticing that the difference between the amounts bought and sold of each asset are exactly equal to the total amount of trades needed to get from the original position $w_{k,i}^0$ to the final position $w_{k,i}$ of that asset.²³

$$t_i^+ - t_i^- = \sum_{k=1}^K \eta_{k,i} \cdot (w_{k,i} - w_{k,i}^0)$$

Here t_i^+ and t_i^- are nonnegative variables.

The multiaccount optimization problem then takes the form

$$\begin{aligned} \max_{\mathbf{w}_1, \dots, \mathbf{w}_K, t^+, t^-} & E[u_1(\mathbf{w}_1)] + \dots + E[u_K(\mathbf{w}_K)] - \sum_{i=1}^N \tau_i(t_i^+, t_i^-) \\ \text{s.t. } & \mathbf{w}_k \in C_k, k = 1, \dots, K \\ & w_{k,i} - w_{k,i}^0 \leq w_{k,i}^+, i = 1, \dots, N, k = 1, \dots, K \\ & t_i^+ = \sum_{k=1}^K \eta_{k,i} w_{k,i}^+, i = 1, \dots, N \\ & t_i^+ - t_i^- = \sum_{k=1}^K \eta_{k,i} \cdot (w_{k,i} - w_{k,i}^0), i = 1, \dots, N \\ & t_i^+ \geq 0, t_i^- \geq 0, w_{k,i}^+ \geq 0, i = 1, \dots, N, k = 1, \dots, K \end{aligned}$$

O'Kinneide, Scherer, and Xu (2006) studied the behavior of the model in simulated experiments with a simple model for the transaction cost function, namely, one in which

$$\tau(t) = \theta \cdot t^\gamma$$

where t is the trade size, and θ and γ are constants satisfying $\theta \geq 0$ and $\gamma \geq 1$.²⁴ θ and γ are specified in advance and calibrated to fit observed trading costs in the market. The transaction costs for each client k can therefore be expressed as

$$\tau_k = \theta \sum_{i=1}^N |w_{k,i} - w_{k,i}^0|^\gamma$$

O'Kinneide, Scherer, and Xu (2006) observed that key portfolio performance measures, such

as the information ratio (IR),²⁵ turnover, and total transaction costs, change under this model relative to the traditional approach. Not surprisingly, the turnover and the net information ratios of the portfolios obtained with *multiaccount optimization* are lower than those obtained with single-account optimization under the assumption that accounts are traded separately, while transaction costs are higher. These results are in fact more realistic, and they are a better representation of the postoptimization performance of multiple client accounts in practice.

ROBUST PARAMETER ESTIMATION

The most commonly used approach for estimating security expected returns, covariances, and other parameters that are inputs to portfolio optimization models is to calculate the sample analogues from historical data. These are sample estimates for the parameters we need. It is important to remember that when we rely on historical data for estimation purposes, we in fact assume that the past provides a good representation of the future.

It is well known, however, that expected returns exhibit significant time variation (referred to as nonstationarity). They are impacted by changes in markets and economic conditions, such as interest rates, the political environment, consumer confidence, and the business cycles of different industry sectors and geographical regions. Consequently, extrapolated historical returns are often poor forecasts of future returns.

Similarly, the covariance matrix is unstable over time. Moreover, sample estimates of covariances for portfolios with thousands of stocks are notoriously unreliable, because we need large data sets to estimate them, and such large data sets of relevant data are difficult to procure. Estimates of the covariance matrix based on factor models are often used to reduce the number of statistical estimates needed from a limited set of data.

In practice, portfolio managers often alter historical estimates of different parameters subjectively or objectively, based on their expectations and forecasting models for future trends. They also use statistical methods for finding estimators that are less sensitive to outliers and other sampling errors, such as Bayesian and shrinkage estimators. A complete review of advanced statistical estimation topics is beyond the scope of this entry. We provide a brief overview of the most widely used concepts.²⁶

Shrinkage is a form of averaging different estimators. The shrinkage estimator typically consists of three components: (1) an estimator with little or no structure (like the sample mean); (2) an estimator with a lot of structure (the shrinkage target); and (3) a coefficient that reflects the shrinkage intensity. Probably the most well-known estimator for expected returns in the financial literature was proposed by Jorion (1986). The shrinkage target in Jorion's model is a vector array with the return on the minimum variance portfolio, and the shrinkage intensity is determined from a specific formula.²⁷ Shrinkage estimators are used for estimates of the covariance matrix of returns as well,²⁸ although equally weighted portfolios of covariance matrix estimators have been shown to be equally effective as shrinkage estimators.²⁹

Bayesian estimation approaches, named after the English mathematician Thomas Bayes, are based on subjective interpretations of the probability that a particular event will occur. A probability distribution, called the prior distribution, is used to represent the investor's knowledge about the probability before any data are observed. After more information is gathered (e.g., data are observed), a formula (known as Bayes' rule) is used to compute the new probability distribution, called the posterior distribution.

In the portfolio parameter estimation context, a posterior distribution of expected returns is derived by combining the forecast from the empirical data with a prior distribution. One of the most well-known examples of the applica-

tion of the Bayesian framework in this context is the Black-Litterman model,³⁰ which produces an estimate of future expected returns by combining the market equilibrium returns (i.e., returns that are derived from pricing models and observable data) with the investor's subjective views. The investor's views are expressed as absolute or relative deviations from the equilibrium together with confidence levels of the views (as measured by the standard deviation of the views).

The ability to incorporate exogenous insight, such as a portfolio manager's opinion, into quantitative forecasting models is important; this insight may be the most valuable input to the model. The Bayesian framework provides a mechanism for forecasting systems to use both important traditional information sources such as proprietary market data and subjective external information sources such as analyst's forecasts.

It is important to realize that regardless of how sophisticated the estimation and forecasting methods are, they are always subject to estimation error. What makes matters worse, however, is that different estimation errors can accumulate over the different activities of the portfolio management process, resulting in large aggregate errors at the final stage. It is therefore critical that the inputs evaluated at each stage are reliable and robust, so that the aggregate impact of estimation errors is minimized.

PORTFOLIO RESAMPLING

Robust parameter estimation is only one part of ensuring that the quantitative portfolio management process as a whole is reliable. It has been observed that portfolio allocation schemes are very sensitive to small changes in the inputs that go into the optimizer. In particular, a well-known study by Black and Litterman³¹ demonstrated that in the case of mean-variance optimization, small changes in the inputs for expected returns had a substantial impact on

the portfolio composition. “Optimal” portfolios constructed under conditions of uncertainty can have extreme or nonintuitive weights for some stocks.

With advances in computational capabilities and new research in the area of optimization under uncertainty, practitioners in recent years have been able to incorporate considerations for uncertainty not only at the estimation, but also at the portfolio optimization stage. Methods for taking into consideration inaccuracies in the inputs to the portfolio optimization problem include simulation (resampling) and robust optimization. We explain *portfolio resampling* in this section, and robust portfolio optimization in the following section.

A logical approach to making portfolio allocation more robust with respect to changes in the input parameters is to generate different scenarios for the values these parameters can take, and to find weights that remain stable for small changes in the input parameters. This method is referred to as *portfolio resampling*.³² To illustrate the resampling technique, we explain how it is applied to portfolio mean-variance optimization.

Suppose that we have initial estimates for the expected stock returns, $\hat{\boldsymbol{\mu}}$, and covariance matrix, $\hat{\boldsymbol{\Sigma}}$, for the N stocks in the portfolio. (We use “hat” to denote a statistical estimate.)

1. We simulate S samples of N returns from a multivariate normal distribution with mean $\hat{\boldsymbol{\mu}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$.
2. We use the S samples generated in (1) to compute S new estimates of vectors of expected returns $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_S$ and covariance matrices $\hat{\boldsymbol{\Sigma}}_1, \dots, \hat{\boldsymbol{\Sigma}}_S$.
3. We solve S portfolio optimization problems, one for each estimated pair of expected returns and covariances $(\hat{\boldsymbol{\mu}}_s, \hat{\boldsymbol{\Sigma}}_s)$, and save the weights for the N stocks in a vector array $\mathbf{w}^{(s)}$, where $s = 1, \dots, S$. (The optimization problem itself could be any of the standard mean-variance formulations: maximize expected return subject to constraints on

risk, minimize risk subject to constraints on the expected return, or maximize the utility function.)

4. To find the final portfolio weights, we average out the weight for each stock over the S weights found for that stock in each of the S optimization problems. In other words,

$$\mathbf{w} = \frac{1}{S} \sum_{s=1}^S \mathbf{w}^{(s)}$$

For example, stock i in the portfolio has final weight

$$w_i = \frac{w_i^{(1)} + \dots + w_i^{(S)}}{S}$$

5. Perhaps even more valuable than the average estimate of the weights obtained from the simulation and optimization iterations is the probability distribution we obtain for the portfolio weights. If we plot the weights for each stock obtained over the S iterations, $w_i^{(1)}, \dots, w_i^{(S)}$, we can get a sense for how variable this stock weight is in the portfolio. A large standard deviation computed from the distribution of portfolio weight i will be an indication that the original portfolio weight was not very precise due to estimation error.

An important question, of course, is how large is “large enough.” Do we have evidence that the portfolios we obtained through resampling are statistically different from one another? We can evaluate that by using a test statistic. For example, it can be shown that the test statistic

$$d(\mathbf{w}^*, \mathbf{w}) = (\mathbf{w}^* - \mathbf{w})' \boldsymbol{\Sigma} (\mathbf{w}^* - \mathbf{w})$$

follows a chi-square (χ^2) distribution with degrees of freedom equal to the number of securities in the portfolio. If the value of this statistic is statistically “large,” then there will be evidence that the portfolio weights \mathbf{w}^* and \mathbf{w} are statistically different. This is an important insight for the portfolio manager, and its applications extend beyond just resampling. Let us provide some intuition as to why.

Suppose that we are considering rebalancing our current portfolio. Given our forecasts of expected returns and risk, we could calculate a set of new portfolios through the resampling procedure. Using the test statistic above, we determine whether the new set of portfolio weights is statistically different from our current weights and, therefore, whether it would be worthwhile to rebalance or not. If we decide that it is worthwhile to rebalance, we could choose any of the resampled portfolios that are statistically different from our current portfolio. Which one should we choose? A natural choice would be to select the portfolio that would lead to the lowest transaction costs. The idea of determining statistically equivalent portfolios, therefore, has much wider implications than the ones illustrated in the context of resampling.

Resampling has its drawbacks:

- Since the resampled portfolio is calculated through a simulation procedure in which a portfolio optimization problem needs to be solved at each step, the approach is computationally cumbersome, especially for large portfolios. There is a trade-off between the number of resampling steps and the accuracy of estimation of the effect of errors on the portfolio composition.
- Due to the averaging in the calculation of the final portfolio weights, it is highly likely that all stocks will end up with nonzero weights. This has implications for the amount of transaction costs that will be incurred if the final portfolio is to be attained. One possibility is to include constraints that limit both the turnover and the number of stocks with nonzero weights. As we saw earlier, however, the formulation of such constraints adds another level of complexity to the optimization problem and will slow down the resampling procedure.
- Since the averaging process happens *after* the optimization problems are solved, the final weights may not actually satisfy some of the constraints in the optimization formulation.

In general, only convex (such as linear) constraints are guaranteed to be satisfied by the averaged final weights. Turnover constraints, for example, may not be satisfied. This is a serious limitation of the resampling approach for practical applications.

Despite these limitations, resampling has advantages and presents a good alternative to using only point estimates of inputs to the optimization problem.

ROBUST PORTFOLIO OPTIMIZATION

Another way in which uncertainty about the inputs can be modeled is by incorporating it directly into the optimization process. *Robust optimization* is an intuitive and efficient way to deal with uncertainty. Robust portfolio optimization does not use the traditional forecasts, such as expected returns and covariances, but rather uncertainty sets containing these point estimates. An example of such an uncertainty set is a confidence interval around the forecast for each expected return (“alpha”). This uncertainty shape looks like a “box” in the space of the input parameters. (See Figure 2(A).) We can also formulate advanced uncertainty sets that incorporate more knowledge about the estimation error. For instance, a widely used uncertainty set is the ellipsoidal uncertainty set, which takes into consideration the covariance structure of the estimation errors. (See Figure 2(B).) We will see examples of both uncertainty sets in this section.

The robust optimization procedure for portfolio allocation is as follows. First, we specify the uncertainty sets around the input parameters in the problem. Then, we ask what the optimal portfolio allocation is when the input parameters take the worst possible value inside these uncertainty sets. In effect, we solve an inner problem that determines the worst possible realization of the uncertain parameters over

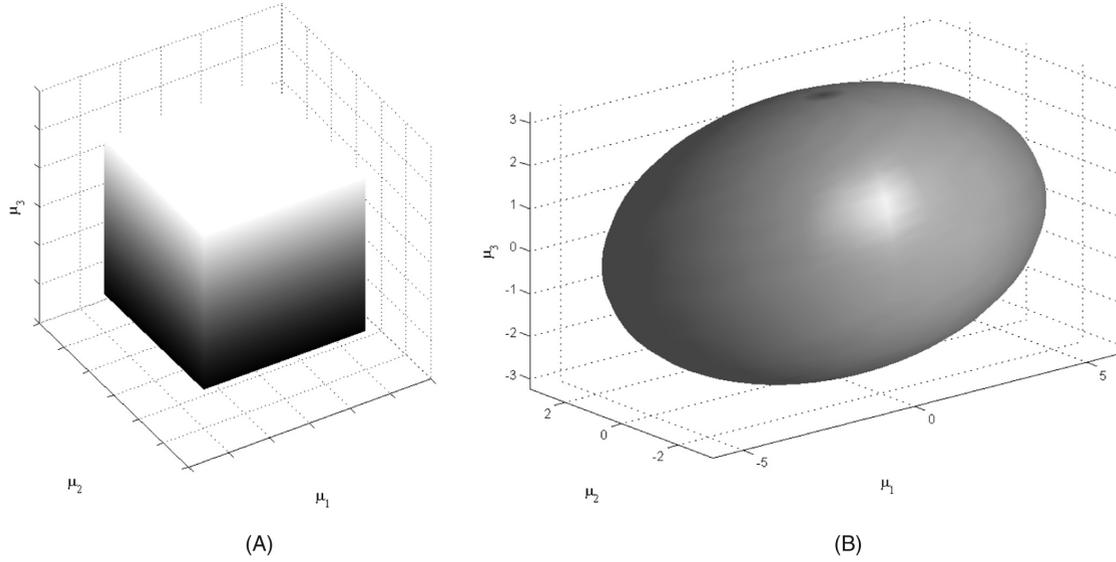


Figure 2 (A) Box Uncertainty Set in Three Dimensions; (B) Ellipsoidal Uncertainty Set in Three Dimensions

the uncertainty set before we solve the original problem of optimal portfolio allocation.

Let us give a specific example of how the robust optimization framework can be applied in the portfolio optimization context. Consider the utility function formulation of the mean-variance portfolio allocation problem:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}'\boldsymbol{\mu} - \lambda \cdot \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{1} = 1 \end{aligned}$$

Suppose that we have estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ of the vector of expected returns and the covariance matrix. Instead of the estimate $\hat{\boldsymbol{\mu}}$, however, we will consider a set of vectors $\boldsymbol{\mu}$ that are “close” to $\hat{\boldsymbol{\mu}}$. We define the box uncertainty set

$$U_{\delta}(\hat{\boldsymbol{\mu}}) = \{\boldsymbol{\mu} \mid |\mu_i - \hat{\mu}_i| \leq \delta_i, i = 1, \dots, N\}$$

In words, the set $U_{\delta}(\hat{\boldsymbol{\mu}})$ contains all vectors $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$ such that each component μ_i is in the interval $[\hat{\mu}_i - \delta_i, \hat{\mu}_i + \delta_i]$. We then solve the following problem:

$$\max_{\mathbf{w}} \left\{ \min_{\boldsymbol{\mu} \in U_{\delta}(\hat{\boldsymbol{\mu}})} \{\boldsymbol{\mu}'\mathbf{w}\} - \lambda \cdot \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} \right\} \quad \text{s.t.} \quad \mathbf{w}'\mathbf{1} = 1$$

This is called the robust counterpart of the original problem. It is a max-min problem that searches for the optimal portfolio weights when the estimates of the uncertain returns take their worst-case values within the prespecified uncertainty set in the sense that the value of the objective function is the worst it can be over all possible values for the expected returns in the uncertainty set.

It can be shown³³ that the max-min problem above is equivalent to the following problem

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}'\boldsymbol{\mu} - \delta'|\mathbf{w}| - \lambda \cdot \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{1} = 1 \end{aligned}$$

where $|\mathbf{w}|$ denotes the absolute value of the entries of the vector of weights \mathbf{w} . To gain some intuition, notice that if the weight of stock i in the portfolio is negative, the worst-case expected return for stock i is $\mu_i + \delta_i$ (we lose the largest amount possible). If the weight of stock i in the portfolio is positive, then the worst-case expected return for stock i is $\mu_i - \delta_i$ (we gain the smallest amount possible). Observe that $\mu_i w_i - \delta_i |w_i|$ equals $(\mu_i - \delta_i) w_i$ if the weight w_i is positive and $(\mu_i + \delta_i) w_i$ if the weight w_i is

negative. Hence, the mathematical expression in the objective agrees with our intuition: It minimizes the worst-case expected portfolio return. In this robust version of the mean-variance formulation, stocks whose mean return estimates are less accurate (i.e., have a larger estimation error δ_i) are therefore penalized in the objective function and will tend to have a smaller weight in the optimal portfolio allocation.

This optimization problem has the same computational complexity as the nonrobust mean-variance formulation—namely, it can be stated as a quadratic optimization problem. The latter can be achieved by using a standard trick that allows us to get rid of the absolute values for the weights. The idea is to introduce an N -dimensional vector of additional variables ψ to replace the absolute values $|w|$, and to write an equivalent version of the optimization problem,

$$\begin{aligned} \max_{\mathbf{w}, \psi} \quad & \mathbf{w}'\hat{\boldsymbol{\mu}} - \delta'\boldsymbol{\psi} - \lambda \cdot \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{1} = 1 \\ & \psi_i \geq w_i; \psi_i \geq -w_i, i = 1, \dots, N \end{aligned}$$

Therefore, incorporating considerations about the uncertainty in the estimates of the expected returns in this example has virtually no computational cost.

We can view the effect of this particular “robustification” of the mean-variance portfolio optimization formulation in two different ways. On the one hand, we can see that the values of the expected returns for the different stocks have been adjusted downward in the objective function of the optimization problem. The robust optimization model “shrinks” the expected return of stocks with large estimation error, that is, in this case the robust formulation is related to statistical shrinkage methods, which we introduced earlier in this entry. On the other hand, we can interpret the additional term in the objective function as a “risk-like” term that represents penalty for estimation error. The size of the penalty is determined by the investor’s aversion to estimation risk and is reflected in the magnitude of the deltas.

More complicated specifications for uncertainty sets have more involved mathematical representations, but can still be selected so that they preserve an easy computational structure for the robust optimization problem. For example, we can use the ellipsoidal uncertainty set from Figure 2(B), which can be expressed mathematically as

$$U_\delta(\hat{\boldsymbol{\mu}}) = \{ \boldsymbol{\mu} \mid (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})' \boldsymbol{\Sigma}_\mu^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}) \leq \delta^2 \}.$$

Here $\boldsymbol{\Sigma}_\mu$ is the covariance matrix of estimation errors for the vector of expected returns $\boldsymbol{\mu}$. This uncertainty set represents the requirement that the sum of squares (scaled by the inverse of the covariance matrix of estimation errors) between all elements in the set and the point estimates $\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_N$ can be no larger than δ^2 . We note that this uncertainty set cannot be interpreted as individual confidence intervals around each point estimate. Instead, it captures the idea of a joint confidence region. In practical applications, the covariance matrix of estimation errors is often assumed to be diagonal. In the latter case, the set contains all vectors of expected returns that are within a certain number of standard deviations from the point estimate of the vector of expected returns, and the resulting robust portfolio optimization problem would protect the investor if the vector of expected returns is within that range.

It can be shown that the robust counterpart of the mean-variance portfolio optimization problem with an ellipsoidal uncertainty set for the expected return estimates is the following optimization problem formulation:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}'\boldsymbol{\mu} - \lambda \cdot \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} - \delta \cdot \sqrt{\mathbf{w}'\boldsymbol{\Sigma}_\mu\mathbf{w}} \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{1} = 1 \end{aligned}$$

This is a second-order cone optimization problem and requires specialized software to solve, but the methods for solving it are very efficient.

Similarly to the case of the robust counterpart with a box uncertainty set, we can interpret the extra term in the objective function ($\delta \cdot \sqrt{\mathbf{w}'\boldsymbol{\Sigma}_\mu\mathbf{w}}$) as the penalty for estimation risk,

where δ incorporates the degree of the investor's aversion to estimation risk. Note, by the way, that the covariance matrix in the estimation error penalty term, Σ_{μ} , is not necessarily the same as the covariance matrix of returns Σ . In fact, it is not immediately obvious how Σ_{μ} can be estimated from data. Σ_{μ} is the covariance matrix of the errors in the estimation of the expected (average) returns. Thus, if a portfolio manager forecasts 5% active return over the next time period, but gets 1%, the manager cannot argue that there was a 4% error in the expected return—the actual error would consist of both an estimation error in the expected return and the inherent volatility in actual realized returns. In fact, critics of the approach such as Lee, Stefek, and Zhelenyak (2006) have argued that the realized returns typically have large stochastic components that dwarf the expected returns, and hence estimating Σ_{μ} from data is very hard, if not impossible.

Several approximate methods for estimating Σ_{μ} have been found to work well in practice. For example, Stubbs and Vance (2005) observe that simpler estimation approaches, such as using just the diagonal matrix containing the variances of the estimates (as opposed to the complete error covariance matrix), often provide most of the benefit in robust portfolio optimization. In addition, standard approaches for estimating expected returns, such as Bayesian statistics and regression-based methods, can produce estimates for the estimation error covariance matrix in the process of generating the estimates themselves.³⁴

Among practitioners, the notion of robust portfolio optimization is often equated with the robust mean-variance model we discussed in this section, with the box or the ellipsoidal uncertainty sets for the expected stock returns. While robust optimization applications often involve one form or another of this model, the actual scope of robust optimization can be much broader. We note that the term *robust optimization* refers to the technique of incorporating information about uncertainty sets for the pa-

rameters in the optimization model, and not to the specific definitions of uncertainty sets or the choice of parameters to model as uncertain. For example, we can use the robust optimization methodology to incorporate considerations for uncertainty in the estimate of the covariance matrix in addition to the uncertainty in expected returns, and obtain a different robust portfolio allocation formulation. Robust optimization can be applied also to portfolio allocation models that are different from the mean-variance framework, such as Sharpe ratio optimization and value-at-risk optimization.³⁵ Finally, robust optimization has the potential to provide a computationally efficient way to handle portfolio optimization over multiple stages—a problem for which so far there have been few satisfactory solutions.³⁶ There are numerous useful robust formulations, but a complete review is beyond the scope of this entry.³⁷

Is implementing robust optimization formulations worthwhile? Some tests with simulated and real market data indicate that robust optimization, when inaccuracy is assumed in the expected return estimates, outperforms classical mean-variance optimization in terms of total excess return a large percentage (70–80%) of the time (see, for example, Ceria and Stubbs, 2006). Other tests have not been as conclusive (see, for example, Lee, Stefek, and Zhelenyak, 2006). The factor that accounts for much of the difference is how the uncertainty in parameters is modeled. Therefore, finding a suitable degree of robustness and appropriate definitions of uncertainty sets can have a significant impact on portfolio performance.

Independent tests by practitioners and academics using both simulated and market data appear to confirm that robust optimization generally results in more stable portfolio weights, that is, that it eliminates the extreme corner solutions resulting from traditional mean-variance optimization. This fact has implications for portfolio rebalancing in the presence of transaction costs and taxes, as transaction costs

and taxes can add substantial expenses when the portfolio is rebalanced. Depending on the particular robust formulations employed, robust mean-variance optimization also appears to improve worst-case portfolio performance and results in smoother and more consistent portfolio returns. Finally, by preventing large swings in positions, robust optimization typically makes better use of the turnover budget and risk constraints.

Robust optimization, however, is not a panacea. By using robust portfolio optimization formulations, investors are likely to trade off the optimality of their portfolio allocation in cases in which nature behaves as they predicted for protection against the risk of inaccurate estimation. Therefore, investors using the technique should not expect to do better than classical portfolio optimization when estimation errors have little impact, or when typical scenarios occur. They should, however, expect insurance in scenarios in which their estimates deviate from the actual realized values by up to the amount they have prespecified in the modeling process.

KEY POINTS

- Commonly used constraints in practice include long-only (no short-selling) constraints, turnover constraints, holding constraints, risk factor constraints, and tracking error constraints. These constraints can be handled in a straightforward way by the same type of optimization algorithms used for solving the classical mean-variance portfolio allocation problem.
- Minimum holding constraints, transaction size constraints, cardinality constraints, and round-lot constraints are also widely used in practice, but their nature is such that they require binary and integer modeling, which necessitates the use of mixed-integer and other specialized optimization solvers.
- Transaction costs can easily be incorporated in standard portfolio allocation models. Typical functions for representing transaction

costs include linear, piecewise linear, and quadratic.

- Taxes can have a dramatic effect on portfolio returns; however, it is difficult to incorporate them into the classical portfolio optimization framework. Their importance to the individual investor is a strong argument for taking a multiperiod view of investments, but the computational burden of multiperiod portfolio optimization formulations with taxes is extremely high.
- For investment managers who handle multiple accounts, increased transaction costs because of the market impact of simultaneous trades can be an important practical issue and should be taken into consideration when individual clients' portfolio allocation decisions are made to ensure fairness across accounts.
- As the use of quantitative techniques has become widespread in the investment industry, the consideration of estimation risk and model risk has grown in importance. Methods for robust statistical estimation of parameters include shrinkage and Bayesian techniques.
- Portfolio resampling is a technique that uses simulation to generate multiple scenarios for possible values of the input parameters in the portfolio optimization problem and aims to determine portfolio weights that remain stable with respect to small changes in model parameters.
- Robust portfolio optimization incorporates uncertainty directly into the optimization process. The uncertain parameters in the optimization problem are assumed to vary in prespecified uncertainty sets that are selected subjectively or based on data.

NOTES

1. See Chapter 1 in Maginn and Tuttle (1990).
2. Multiperiod portfolio optimization models are still rarely used in practice, not because the value of multiperiod modeling is

questioned, but because such models are often too intractable from a computational perspective.

3. As the term intuitively implies, the ADV measures the total amount of a given asset traded in a day on average, where the average is taken over a prespecified time period.
4. Another computationally tractable situation for minimizing CVaR is when the data are normally distributed. In that case, minimizing CVaR is equivalent to minimizing the standard deviation of the portfolio.
5. See Chapters 8 and 9 in Pachamanova and Fabozzi (2010) for a more detailed explanation of CVaR and a derivation of the optimization formulation.
6. Versions of this model have been suggested in Pogue (1970), Schreiner (1980), Adcock and Meade (1994), Lobo, Fazel, and Boyd (2000), and Mitchell and Braun (2004).
7. Here we are thinking of w_i as the portfolio weights, but in fact it may be more intuitive to think of the transaction costs as a percentage of amount traded. It is easy to go back and forth between portfolio weights and portfolio amounts by simply multiplying w_i by the total amount in the portfolio. In fact, we can switch the whole portfolio optimization formulation around and write it in terms of allocation of dollars, instead of weights. We just need to replace the vector of weights \mathbf{w} by a vector \mathbf{x} of dollar holdings.
8. See, for example, Bertsimas, Darnell, and Soucy (1999).
9. As we explained earlier, this constraint can be written in an equivalent, more optimization solver-friendly form, namely,

$$y_i \geq w_i - w_{0,i}$$

$$y_i \geq -(w_i - w_{0,i})$$
10. The computation of the tax basis is different for stocks and bonds. For bonds, there are special tax rules, and the original price is not the tax basis.
11. The exact rates vary depending on the current version of the tax code, but the main idea behind the preferential treatment of long-term gains to short-term gains is to encourage long-term capital investments and fund entrepreneurial activity.
12. See Stein (1998).
13. See Apelfeld, Fowler, and Gordon (1996) who show that a manager can outperform on an after-tax basis with high turnover as well, as long as the turnover does not result in net capital gains taxes. (There are other issues with high turnover, however, such as higher transaction costs that may result in a lower overall portfolio return.)
14. Dividends are taxed as regular income, i.e., at a higher rate than capital gains, so minimizing the portfolio dividend yield should theoretically result in a lower tax burden for the investor.
15. See Constantinides (1983), Dammon and Spatt (1996), and Dammon, Spatt, and Zhang (2001 and 2004).
16. The Securities and Exchange Commission (SEC) in general prohibits cross-trading but does provide exemptions if prior to the execution of the cross trade the asset manager can demonstrate to the SEC that a particular cross trade benefits both parties. Similarly, Section 406(b)(3) of the Employee Retirement Income Security Act of 1974 (ERISA) forbids cross-trading, but there is new cross-trading exemption in Section 408(b)(19) adopted in the Pension Protection Act of 2006.
17. Khodadadi, Tutuncu, and Zangari (2006).
18. The iterative procedure is known to converge to the equilibrium, however, under special conditions. See O’Cinneide, Scherer, and Xu (2006).
19. The issue of considering transaction costs in multiaccount optimization has been discussed by others as well. See, for example, Bertsimas, Darnell, and Soucy (1999).
20. As we mentioned earlier in this entry, realistic transaction costs are in fact described by

- nonlinear functions, because costs per share traded typically increase with the size of the trade due to market impact.
21. For example, if asset i is a euro-pound forward, then a trade in that asset can also be implemented as a euro-dollar forward plus a dollar-forward, so there will be two additional assets in the aggregate trade vector \mathbf{t} .
 22. Note that $\eta_{k,i}$ equals 1 if $w_{k,i}$ is the actual dollar holdings.
 23. Note that, similarly to \mathbf{w}_k^+ , we could introduce additional sell variables \mathbf{w}_k^- , but this is not necessary. By expressing aggregate sales through aggregate buys and total trades, we reduce the dimension of the optimization problem, because there are fewer decision variables. This would make a difference for the speed of obtaining a solution, especially in the case of large portfolios and complicated representation of transaction costs.
 24. Note that $\gamma = 1$ defines linear transaction costs. For linear transaction costs, multi-account optimization produces the same allocation as single-account optimization, because linear transaction costs assume that an increased aggregate amount of trading does not have an impact on prices.
 25. The information ratio is the ratio of (annualized) portfolio residual return (alpha) to (annualized) portfolio residual risk, where risk is defined as standard deviation.
 26. For further details, see Chapters 6, 7, and 8 in Fabozzi, Kolm, Pachamanova, and Focardi (2007).
 27. See Chapter 8 in Fabozzi, Kolm, Pachamanova, and Focardi (2007).
 28. See, for example, Ledoit and Wolf (2003).
 29. For an overview of such models, see Disatnik and Benninga (2007).
 30. For a step-by-step description of the Black-Litterman model, see Chapter 8 in Fabozzi, Kolm, Pachamanova, and Focardi (2007).
 31. See Black and Litterman (1992).
 32. See Michaud (1998), Jorion (1992), and Scherer (2002).
 33. For derivation, see, for example, Chapter 12 in Fabozzi, Kolm, Pachamanova, and Focardi (2007) or Chapter 9 in Pachamanova and Fabozzi (2010).
 34. For a more in-depth coverage of the topic of estimating input parameters for robust optimization formulations, see Chapter 12 in Fabozzi, Kolm, Pachamanova, and Focardi (2007).
 35. See, for example, Goldfarb and Iyengar (2003) and Natarajan, Pachamanova, and Sim (2008).
 36. See Ben-Tal, Margalit, and Nemirovski (2000) and Bertsimas and Pachamanova (2008).
 37. For further details, see Fabozzi, Kolm, Pachamanova, and Focardi (2007).

REFERENCES

- Adcock, C., and Meade, N. (1994). A simple algorithm to incorporate transaction costs in quadratic optimization. *European Journal of Operational Research* 79, 1: 85–94.
- Apelfeld, R., Fowler, G. B., and Gordon, J. P. (1996). Tax-aware equity investing. *Journal of Portfolio Management* 22, 2: 18–28.
- Ben-Tal, A., Margalit, T., and Nemirovski, A. (2000). Robust modeling of multi-stage portfolio problems. In H. Frenk, K. Roos, T. Terlaky, and S. Zhang (eds.), *High-Performance Optimization* (Dordrecht: Kluwer Academic Publishers, pp. 303–328).
- Bertsimas, D., Darnell, C., and Soucy, R. (1999). Portfolio construction through mixed-integer programming at Grantham, Mayo, Van Otterloo and Company. *Interfaces* 29, 1: 49–66.
- Bertsimas, D., and Pachamanova, D. (2008). Robust multiperiod portfolio management with transaction costs. *Computers and Operations Research*, special issue on *Applications of Operations Research in Finance* 35, 1: 3–17.
- Black, F., and Litterman, R. (1992). Global portfolio optimization. *Financial Analysts Journal* 48, 5: 28–43.
- Ceria, S., and Stubbs, R. (2006). Incorporating estimation errors into portfolio selection: Robust portfolio construction. *Journal of Asset Management* 7, 2: 109–127.

- Constantinides, G. (1983). Capital market equilibrium with personal taxes. *Econometrica* 5: 611–636.
- Dammon, R. M., and Spatt, C. S. (1996). The optimal trading and pricing of securities with asymmetric capital gains taxes and transaction costs. *Review of Financial Studies* 9, 3: 921–952.
- Dammon, R. M., Spatt, C. S., and Zhang, H. H. (2001). Optimal consumption and investment with capital gains taxes. *Review of Financial Studies* 14, 3: 583–617.
- Dammon, R. M., Spatt, C. S., and Zhang, H. H. (2004). Optimal asset location and allocation with taxable and tax-deferred investing. *Journal of Finance* 59, 3: 999–1037.
- David, D., and Benninga, S. (2007). Shrinking the covariance matrix—simpler is better. *Journal of Portfolio Management* 33, 4: 56–63.
- DiBartolomeo, D. (2000). Recent advances in management of taxable portfolios. Manuscript, Northfield Information Services.
- Fabozzi, F. J., Kolm, P., Pachamanova, D., and Focardi, S. (2007). *Robust Portfolio Optimization and Management*. Hoboken, NJ: John Wiley & Sons.
- Goldfarb, D., and Iyengar, G. (2003). Robust portfolio selection problems. *Mathematics of Operations Research* 28, 1: 1–38.
- Jorion, P. (1986). Bayes-Stein estimator for portfolio analysis. *Journal of Financial and Quantitative Analysis* 21, 3: 279–292.
- Jorion, P. (1992). Portfolio optimization in practice. *Financial Analysts Journal* 48, 1: 68–74.
- Khodadadi, A., Tutuncu, R., and Zangari, P. (2006). Optimization and quantitative investment management. *Journal of Asset Management* 7, 2: 83–92.
- Ledoit, O., and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance* 10, 5: 603–621.
- Lee, J-H., Stefek, D., and Zhelenyak, A. (2006). Robust portfolio optimization—a closer look. MSCI Barra Research Insights Report, June.
- Lobo, M. S., Fazel, M., and Boyd, S. (2000). Portfolio optimization with linear and fixed transaction costs and bounds on risk. *Annals of Operations Research*.
- Maginn, J. L., and Tuttle, D. L. (eds.). (1990). *Managing Investment Portfolios: A Dynamic Process*. New York: Warren, Gorham & Lamont, sponsored by the Institute of Chartered Financial Analysts, Second Edition.
- Michaud, R. O. (1998). *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation*. Oxford: Oxford University Press.
- Mitchell, J. E., and Braun, S. (2004). Rebalancing an investment portfolio in the presence of convex transaction costs. *Technical Report*, Department of Mathematical Sciences, Rensselaer Polytechnic Institute.
- Natarajan, K., Pachamanova, D., and Sim, M. (2008). Incorporating asymmetric distributional information in robust value-at-risk optimization. *Management Science* 54, 3: 573–585.
- O’Cinneide, C., Scherer, B., and Xu, X. (2006). Pooling trades in a quantitative investment process. *Journal of Portfolio Management* 32, 4: 33–43.
- Pachamanova, D. A., and Fabozzi, F. J. (2010). *Simulation and Optimization in Finance: Modeling with MATLAB, @RISK, and VBA*. Hoboken, NJ: John Wiley & Sons.
- Pogue, G. (1970). An extension of the Markowitz portfolio selection model to include variable transactions costs, short sales, leverage policies, and taxes. *Journal of Finance* 25, 5: 1005–1027.
- Rockafellar, R. T., and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance* 26: 1443–1471.
- Rockefeller, R. T., and Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk* 3: 21–41.
- Schreiner, J. (1980). Portfolio revision: A turnover-constrained approach. *Financial Management* 9, 1: 67–75.
- Scherer, B. (2002). Portfolio resampling: Review and critique, *Financial Analysts Journal* 58, 6: 98–109.
- Stein, D. M. (1998). Measuring and evaluating portfolio performance after taxes. *Journal of Portfolio Management* 24, 2: 117–124.
- Stubbs, R., and Vance, P. (2005). Computing return estimation error matrices for robust optimization. *Report*, Axioma.

Basics of Quantitative Equity Investing

PAUL BUKOWSKI, CFA, FCAS

Executive President, Head of Equities, Hartford Investment Management

Abstract: Quantitative equity investing is one method used by investors to identify attractive stocks and gain a competitive advantage. In contrast to fundamental investors who focus on a single company at a time, quantitative investors focus on stock characteristics. Quantitative investors look for sources of information or company characteristics that help to explain why one stock outperforms another stock. They assemble a group of characteristics into a unique stock selection model, which is the core of the quantitative investment process. The quantitative investment process can be divided into three main phases: research, portfolio construction, and monitoring. During the research phase, the stock selection model is created. During the portfolio construction phase, the quantitative investor uses the stock selection model to create a live portfolio. Finally, during the monitoring phase, the quantitative investor makes sure the portfolio is performing as expected and modifies it as needed. While quantitative investing can be very different from fundamental investing, they are complementary and combined can lead to a more well-rounded overall investment approach.

The goal of this entry is to provide the basics of *quantitative equity investing* and an explanation of the quantitative investing process. More specifically, I focus on the following three questions. First, how do quantitative and *fundamental equity* investors differ? Second, what are the core steps in a quantitative equity investment process? Finally, what are the basic building blocks used by quantitative equity investors?

In answering these questions, I will pull back the curtain on the quantitative equity investment process, showing how it is similar to many other approaches, all searching for the best stocks. Where it differs is in the creation of a repeatable process that uses several key cri-

teria to find the most attractive companies—its *stock selection model*. Finally, some of the most common techniques used by quantitative equity investors are covered.

It is important to understand that this entry is dedicated to a traditional quantitative equity investing approach. There are many other types of investing that are quantitative in nature (e.g., high-frequency trading, statistical arbitrage, etc.), which will not be covered.

EQUITY INVESTING

Investing can take many forms, but it starts with an investor assigning a value to a security.



Figure 1 The Value of a Stock Comes from Multiple Information Sources

Whether this value exceeds or is less than the current market price usually determines whether the investor will buy or sell the security. In the case of equities, the investor often seeks to understand the specific company under consideration, the broader economic environment, and the interplay between the two. This encompasses a wide range of information for the investor to consider as displayed in Figure 1. How this information is used differentiates the quantitative from the fundamental investor.

FUNDAMENTAL VS. QUANTITATIVE INVESTOR

Let's start with a basic question. How do portfolio managers select stocks from a broad universe of 1,000 or more companies?

Fundamental managers start with a basic company screen. For instance, they may first look for companies that satisfy conditions such as a price-earnings (P/E) ratio that is less than 15, earnings growth greater than 10%, and profit margins in excess of 20%. Filtering by those *characteristics* may result in, say, 200 potential candidates. Next, portfolio managers in consultation with their group of stock analysts will

spend the majority of their time thoroughly reviewing each of the potential candidates to arrive at the best 50 to 100 stocks for their portfolio. Quantitative managers, in contrast, spend the bulk of their time determining the characteristics for the initial stock screen, their stock selection model. They will look for five or more unique characteristics that are good at identifying the most attractive 200 stocks of the universe. Quantitative managers will then purchase all 200 stocks for their portfolio.

So let's expand on how these two investors—fundamental and quantitative—differ. Figure 2 details the main attributes of the two approaches discussed further below.

Focus: Company versus Characteristic: The fundamental investor's primary analysis is on a single company at a time, while the quantitative investor's primary analysis is on a single characteristic at a time. For example, a fundamental investor may analyze a health care company to assess whether a company's sales prospects look strong and whether this stronger sales growth is reflected in the company's current stock price. A quantitative investor may also invest in a company based on its sales growth, but

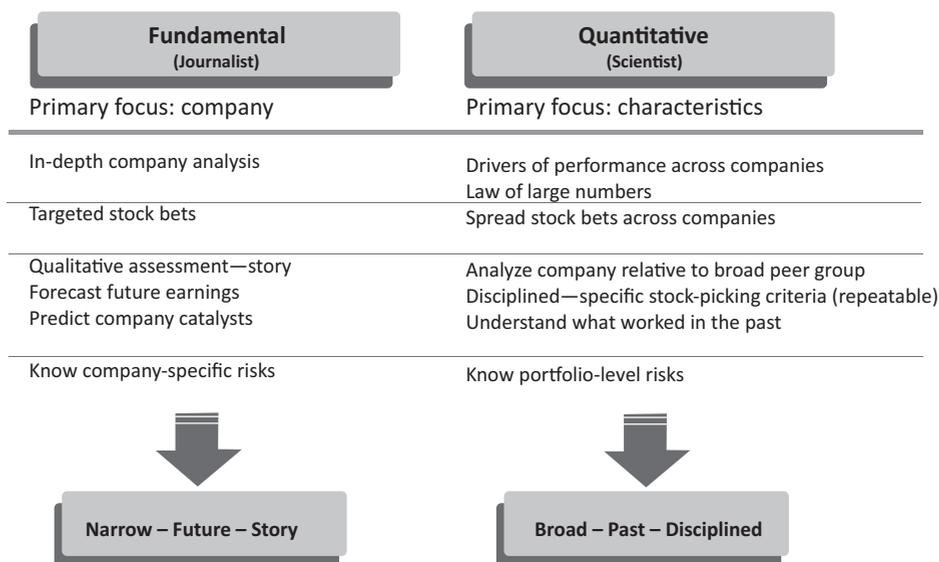


Figure 2 Fundamental vs. Quantitative Investor: Viewing Information

will start by assessing the sales growth characteristic. The quantitative investor will determine whether stocks within the group, health care companies, with higher sales growth also have higher stock returns. If they do, then the quantitative investor will buy health care stocks with higher sales growth. In the end, both types of investors may buy a stock due to its good sales prospects, but both come at the decision from a different point of view.

Narrow vs. Broad: Fundamental investors focus their attention narrowly on a small group of stocks. They cover fewer companies since they make more in-depth reviews of each company. Fundamental investors immerse themselves in the company, studying everything from financial information, to new products, to meeting management. Ideally, they are searching for exploitable differences between their detailed assessment of the company's value and the market's perception of that value. In contrast, quantitative investors focus more broadly. Rather than reviewing one company at a time, they look across a large group of companies. Quan-

titative investors focus on what separates companies from one another; they search for pieces of information (characteristics) that they can use to exploit differences between securities. Since they are dealing with a great deal of data from a large number of companies, they employ quantitative techniques to quickly sift through the information.

Position Concentration/Size of Bets: Another difference in the two approaches is the size of the positions within a portfolio; they tend to be larger for a fundamental investor and smaller for a quantitative investor. Fundamental investors perform in-depth company analysis so they will have greater conviction in taking larger positions in their selected stocks. Quantitative investors perform in-depth analysis across a group of companies, so they will tend to spread their bets across this larger group of companies.

Risk Perspective: The fundamental investor sees risk at the company level while the quantitative investor is more focused at the portfolio level. Fundamental investors will review the risk to both their forecasts and catalysts for the company. They understand

Fundamental	Quantitative
Small portfolio	Large portfolio
Larger positions	Smaller positions
Performance at sector/company level	Performance at characteristic level
Emphasize stock-specific risk	Diversify stock-specific risk

Figure 3 Fundamental vs. Quantitative Investor: Process Differences

how a changing macro picture can impact their valuation of the company. In contrast, the quantitative investor's broader view relates to understanding the risks across the portfolio. They understand if there are risk characteristics in their portfolio that are different from their chosen stock selection model. For example, a quantitative investor who does not believe growth prospects matter to a company's stock performance would want to investigate if the model had the investor buying many very high- or low-growth companies.

Past vs. Future: Finally, the fundamental investor often places greater emphasis on the future prospects of the company while the quantitative investor studies the company's past. Fundamental investors tend to paint a picture of the company's future; they will craft a story around the company and its prospects; and they will look for catalysts generating future growth for a company. They rely on their ability to predict changes in a company. In contrast, the quantitative investor places more emphasis on the past, using what is known or has been reported by a company. Quantitative investors rely on historical accounting data as well as historical strategy simulations, or backtests, to search for the best company characteristics to select stocks. For instance, they will look at whether technology companies with stronger profitability have performed better than those without, or whether retail companies with stronger inventory controls have performed better than those without.

Quantitative investors are looking for stock picking criteria that can be tested and incorporated into a stock selection model.

In the end, we have two types of investors viewing information, often the same information, quite differently. The fundamental investor is a journalist focused on crafting a unique story of a company's future prospects and predicting the potential for gain in the company's stock. The quantitative investor is a scientist, broadly focused, relying on historical information to differentiate across all companies, using statistical techniques to create a stock selection model.

These two investors can and often do create different portfolios based on their different approaches as shown in Figure 3. Fundamental investors are more focused, with higher conviction in their stocks resulting in fewer, larger positions in their portfolios. Quantitative investors, reviewing a large group of companies, generally take a large number of smaller positions in their portfolio. Fundamental investors are investing in a stock (or sector) and therefore are most concerned with how much each of their stocks (or sectors) is contributing to performance. Quantitative investors are investing in a characteristic and how well it differentiates stocks. They want to know how each of their characteristics is contributing to performance. Finally, fundamental investors' detailed view into the company allows them to understand the intrinsic risk of each investment they make—the potential stumbling blocks for each company. Quantitative investors' goal is to

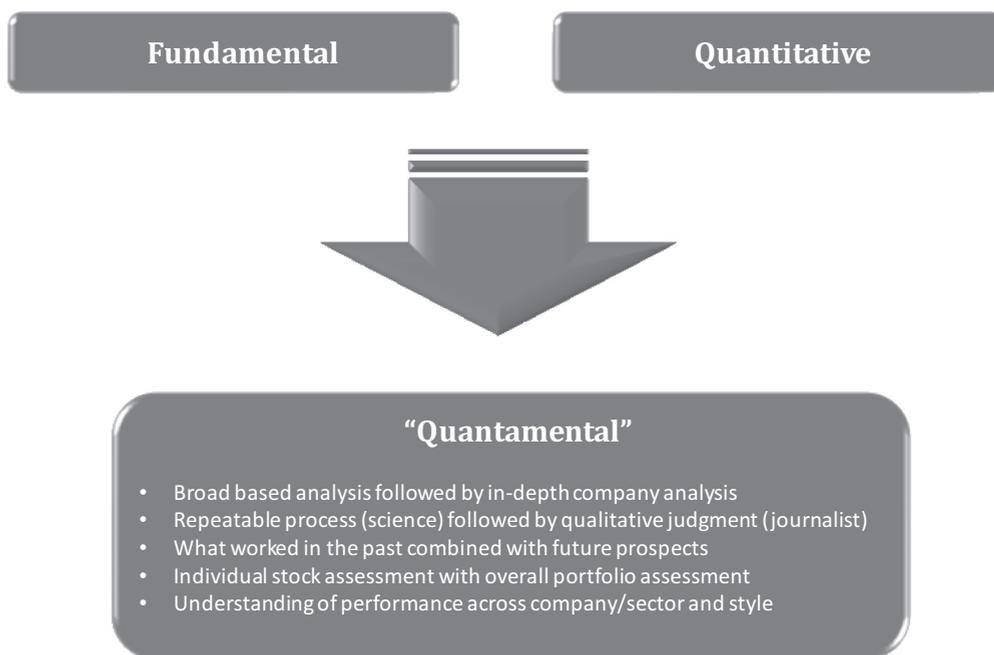


Figure 4 Benefits of a Combined Fundamental and Quantitative Approach

understand specific characteristics across a broad universe of stocks. They look at risks across their entire portfolio, attempting to diversify away any firm-specific risks ancillary to their strategy.

Now that you understand the basic differences between the two approaches, it might also be clear how using both investment styles can be very appealing. As Figure 4 shows, the two styles are quite complementary in nature and can provide a robust, well-rounded view of a company or portfolio. Combining the two approaches provides the following benefits:

- **Breadth and depth.** In-depth analysis across a large group of stocks selecting the best subset of companies, which is followed by in-depth review of the small subset of attractive companies.
- **Facts balanced with human insight.** The scientific approach reviewing large amounts of data across many companies complemented by personal judgment at the company level.
- **Past and future perspective.** A detailed historical review of companies combined with a

review of future potential prospects of a company.

- **Full risk analysis.** A broad look at risk both within each company owned and across the entire portfolio.
- **Clear portfolio performance.** A thorough understanding of which companies, sectors, and characteristics are driving a portfolio's performance.

In fact, over the years, the defining line between the two approaches has been blurring. Some have coined a term for this joint process: "quantamental." Many investment managers are combining both approaches in one investment process, which is why whether you are a fundamental or quantitative investor, it is important to understand both perspectives.

Given our preceding discussion, the distinction between the quantitative and fundamental approaches should now be better appreciated. In the remainder of this entry we restrict our focus to the quantitative equity investment process, addressing the last two topics listed at the beginning of this entry: the core steps in a

quantitative equity investment process and some of the basic building blocks used by quantitative investors.

THE QUANTITATIVE STOCK SELECTION MODEL

Before diving into the details of the quantitative investment process, let's look at what is at its core—the stock selection model. As explained in the previous section, the quantitative investment approach is rooted in understanding what separates strong-performing stocks from weak-performing stocks.¹ The quantitative investor looks for sources of information or company characteristics (often referred to as factors or signals)² that help to explain why one stock outperforms another stock. They assemble these characteristics into a stock selection model, which can be run daily to provide an updated view on every stock in their investment universe.

The stock selection model is at the heart of the quantitative process. To build the model, the quantitative investor will look throughout history and see what characteristics drive performance differences between stocks in a group such as a universe (i.e., small cap, small-cap value, and large-cap growth) or a sector (i.e., technology, financials, materials).

The quantitative investor's typical stock selection methodology is buying stocks with the most attractive attributes and not investing in (or shorting, if permitted by investment guidelines) stocks with the least attractive attributes. For instance, let's suppose retail stocks that have the highest profitability tend to have higher stock returns than those with the lowest profitability. In this case, if a retail stock had strong profitability, there is a greater chance a portfolio manager would purchase it. Profitability is just one characteristic of a company. The quantitative investor will look at a large number of characteristics, from 25 to over 100, to include in the stock selection model. In the



Figure 5 Sample Stock Selection Model for the Retail Sector

end, they will narrow their final model to a few characteristics that are best at locating performance differences among stocks in a particular universe or sector.

Figure 5 is an example of a stock selection model for the retail sector. If a stock has good margins and positive earnings growth, sell-side analysts like it, solid inventory management and is attractively valued, especially as pertains to earnings, then the quantitative investor would buy it. And if it did not have these characteristics, a quantitative investor would not own it, sell it, or short it. This example is for a retail sector; a quantitative investor could also have different models to select stocks in the bank sector or utilities sector or among small-cap value stocks.

So how does a quantitative investor create and use the stock selection model? A good analogy is a professional golfer. Like a quantitative investor, golfers create a model of their game. First, golfers analyze all elements of their basic swing from backswing to follow through. They then alter their swing to different conditions (high winds, rain, cold), and different

course types (links, woodlands, fast greens). Next, golfers put their model into action. While they are golfing, they make mental notes about what is and isn't working to help enhance their game. Could they tweak their swing? What has been effective under the current weather conditions? How are they playing this type of course?

Overall, the golfers' process is much like quantitative investors' process. They create a model, implement it, and then monitor it, assessing their ability to shoot below par. Like professional golfers who go to the driving range for countless hours to perfect their swing, quantitative investors will spend countless hours perfecting their model, understanding how it works under many different market (weather/course) conditions.

With that analogy in mind, we now turn to the entire quantitative investment process.

THE OVERALL QUANTITATIVE INVESTMENT PROCESS

The quantitative process can be divided into the following three main phases (shown in Figure 6):

- Research
- Portfolio construction
- Monitoring

During the *research* phase, the stock selection model is created. During the *portfolio construction* phase, the quantitative investor

“productionalizes” the stock selection model or gets it ready to invest in a live portfolio. Finally, during the *monitoring* phase, the quantitative investor makes sure the portfolio is performing as expected.

RESEARCH

Let's start with the research phase since it is the basic building block of the quantitative process. It is where the fact-finding mission begins. This is similar to when the golfer spends countless hours at the driving range perfecting his (or her) swing. In this phase, the quantitative investor determines what aspects of a company make its stock attractive or unattractive. The research phase begins by the quantitative investors testing all the characteristics they have at their disposal, and it finishes with assembling the chosen characteristics into a stock selection model (see Figure 7).

1. **Characteristic Testing.** First, quantitative investors determine which characteristics are good at differentiating strong-performing from weak-performing stocks. Initially, the quantitative investor segments the stocks. This could be by sector, such as consumer discretionary; industry, such as consumer electronics; or a universe, such as small-cap value stocks. Once the stocks have been grouped, each of the characteristics is tested to see if it can delineate the strong-performing stocks from the weak-performing stocks.

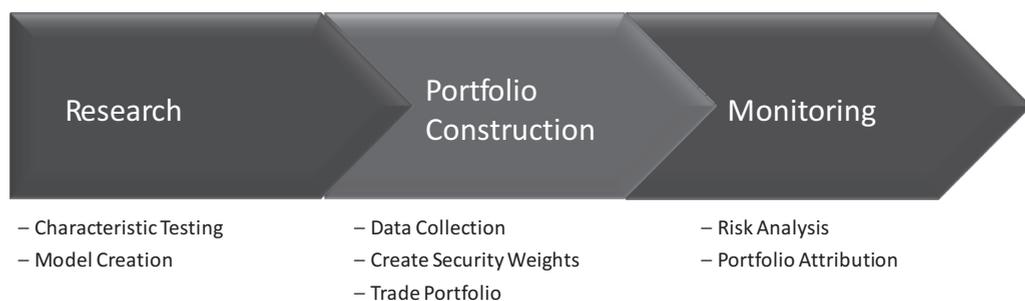


Figure 6 Three Core Phases of the Quantitative Equity Investment Process

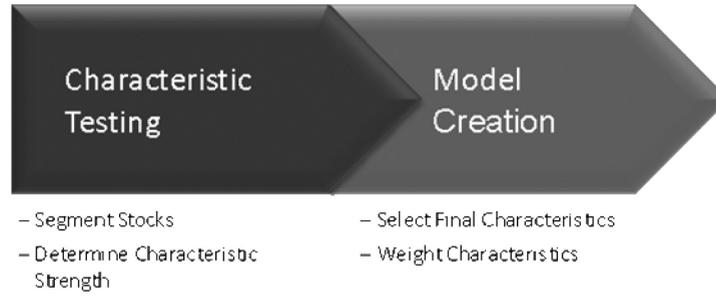


Figure 7 Two Core Steps in the Research Phase of the Quantitative Equity Investment Process

2. *Model Creation.* Second, quantitative investors select the final characteristics that are best at picking the most attractive stocks. Then they weight each characteristic in the stock selection model—determining which characteristics should be more relied upon when picking stocks, or if they all should be treated equally.

looks at historical information over 20 years or more in order to cover multiple market cycles. While testing, many performance metrics are reviewed to get an expansive view of a characteristic’s ability to differentiate stocks. These metrics span the return category, risk category, and other metrics as outlined in Figure 8. Using an array of metrics, quantitative investors are better able to confirm a characteristic’s consistency. They make sure that the selected characteristics score well on more than a single metric. Before continuing with the research process, let’s review a few of the more commonly used metrics.

During the research phase, the quantitative investor tries to get a broad picture of a characteristic, making sure it performs well under a diverse set of conditions and performance measures. For testing, the quantitative investor

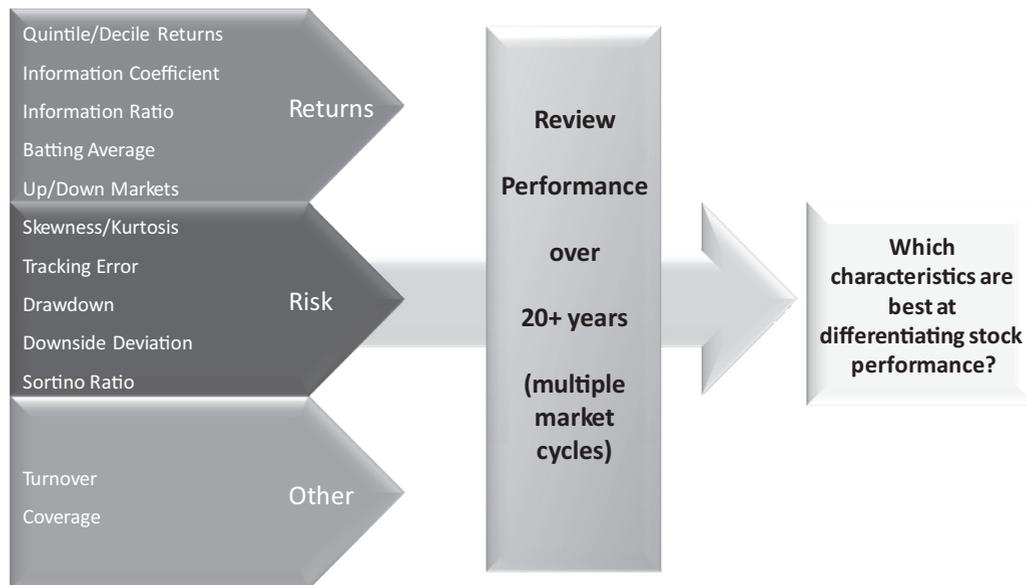


Figure 8 Characteristic Testing in the Research Phase

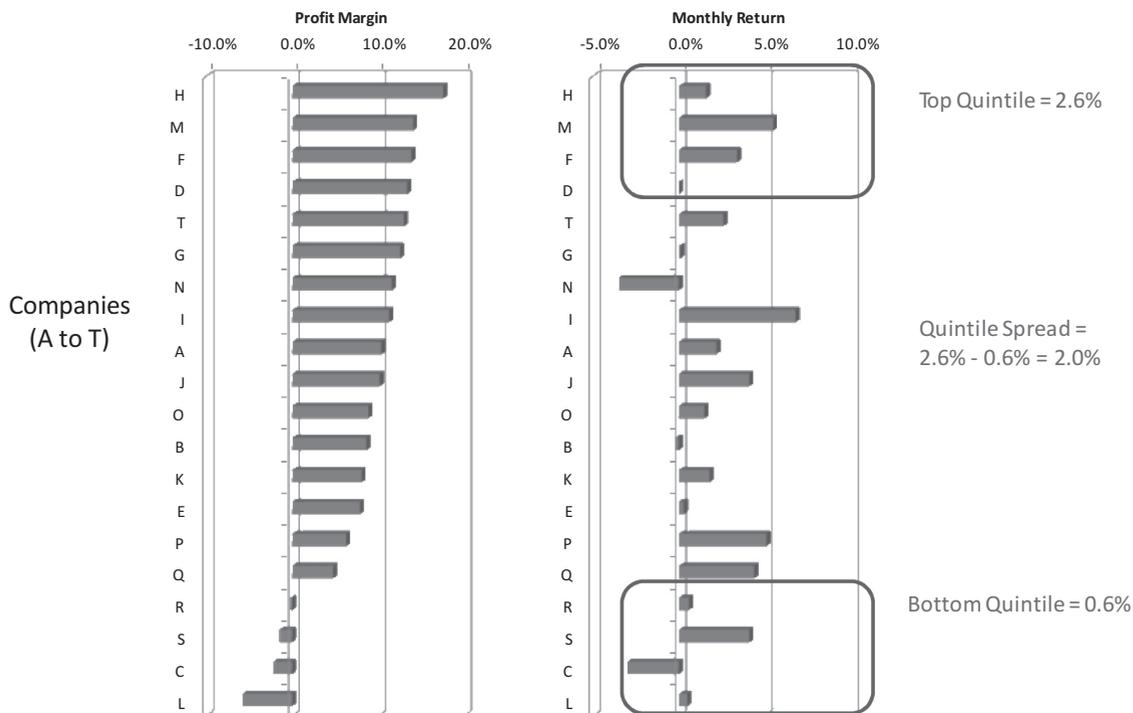


Figure 9 Determining the Characteristic's Quintile Spread

Characteristic Testing: Key Quantitative Research Metrics

In this section we will review quintile returns and information coefficients, which measure whether a characteristic can differentiate between winning and losing stocks. Although profitability was chosen for the examples, other characteristics such as sales growth, P/E ratio, or asset turnover also could have been chosen.

Quintile Returns

The *quintile return* is already prevalent across most research publications, but is gaining popularity in more and more mainstream publications such as the *Wall Street Journal*, *Barron's*, and the like. Quintile returns measure how well a characteristic differentiates stocks. In essence, the stocks that are being reviewed are segmented into five groups (quintiles) and then are tested to determine if the companies in the group with the best attributes (top quintile) out-

perform the group with the least desirable attributes (bottom quintile).

Figure 9 provides an example. In this example, we start with 20 companies that we refer to as A through T. The first step—the left-hand side of the exhibit—is to order the 20 companies by profitability from highest to lowest. In the second step, this ordered list is divided into five groups, creating a most profitable group (top quintile) down to the least profitable group (bottom quintile). The top and bottom quintile groups are boxed on the right-hand chart of the figure. Finally, the performance of the top quintile is compared to the bottom quintile.

As Figure 9 shows, the stocks with highest profitability (top quintile) returned 2.6% while the stocks with the lowest profitability (bottom quintile) returned only 0.6%. So the top quintile stocks outperformed the bottom quintile stocks by 2.0%, meaning for this month, the most profitable companies outperformed the least profitable companies by 2.0%. This is commonly

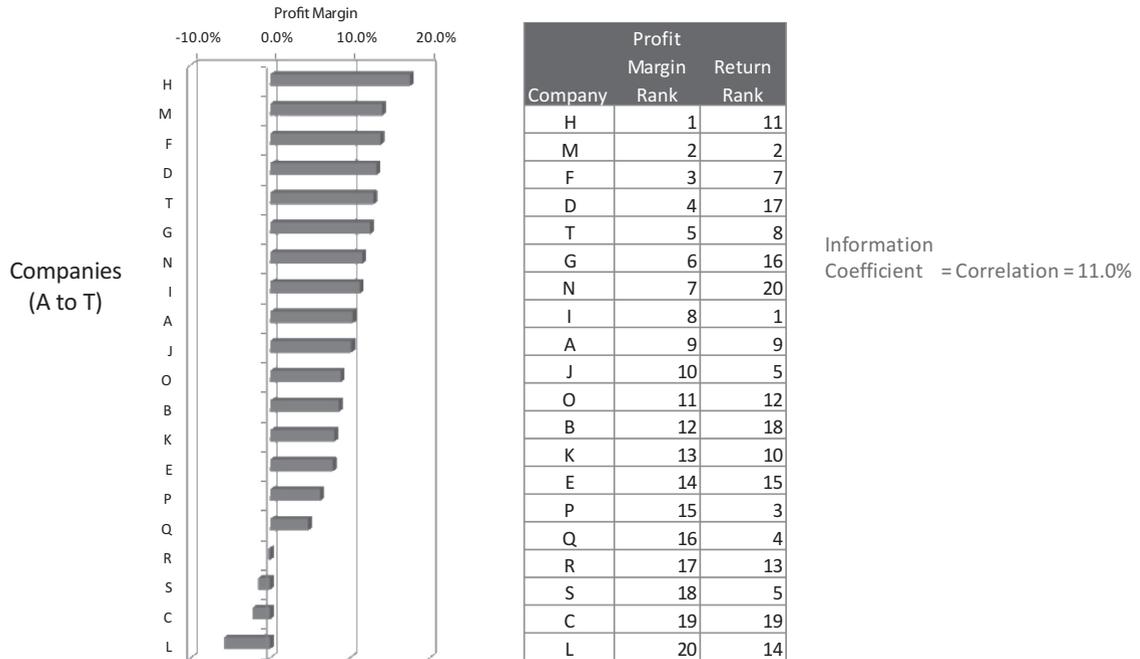


Figure 10 Determining the Characteristic's Information Coefficient

referred to as the characteristic's *quintile return* or quintile spread. The higher the quintile spread, the more attractive the characteristic is.

Information Coefficient

Another common metric used for determining if a characteristic is good at separating the strong- from the weak-performing stocks is the *information coefficient* (IC). It does so by measuring the correlation between a stock's characteristic (i.e., profitability) and its return. The major difference between the IC and quintile return is that the IC looks across all of the stocks, while the quintile return only focuses on the best and worst stocks, ignoring those stocks in the middle. The IC is more concerned with differentiating performance across all stocks rather than the extremes.

The calculation of the IC is detailed in Figure 10. Similar to assessing the quintile return, the sort ordering of the companies based on profitability is done first. However, the next

step is different. In the second step, each stock is ranked on both profitability and return. The most profitable company is assigned a rank of 1 all the way down to the least profitable company, which is assigned a rank of 20. Likewise for stock returns: The highest returning stock is assigned a rank of 1 down to the lowest returning stock receiving a rank of 20. In the third step, the rank of the company's profitability is correlated with the rank of the company's return. The correlation of the two ranks is the IC, which is 11% as shown in Figure 10. The higher the correlation (i.e., IC), the more likely companies with higher profitability also have higher returns and the more effective the characteristic.

When is it better to employ an IC over a quintile spread? IC is a better metric when a quantitative investor is considering owning a greater number of stocks in the portfolio. The reason is that the IC looks at the relationships across all of the stocks in the group. The quintile return is better suited for more concentrated bets in

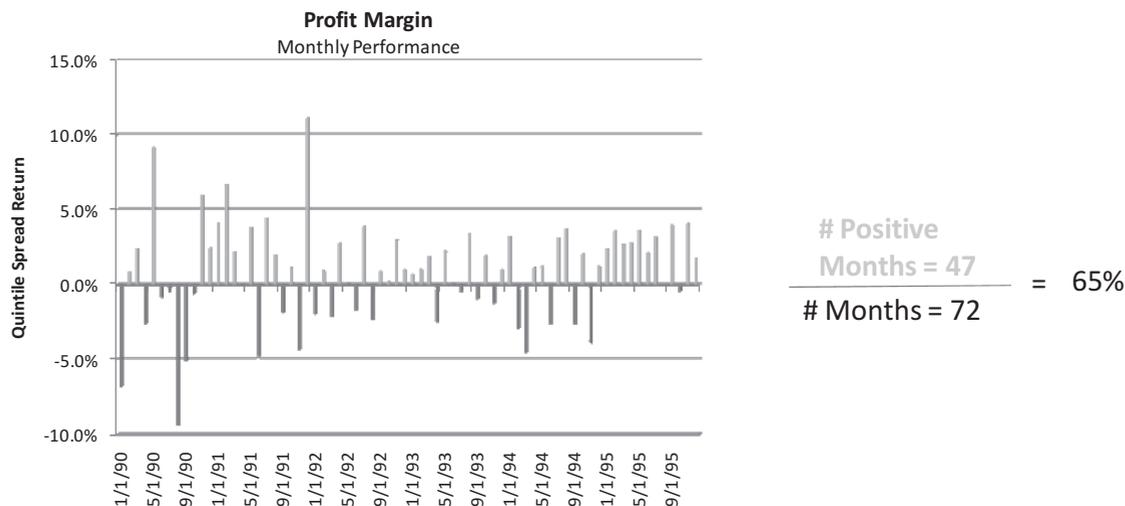


Figure 11 Determining the Characteristic's Batting Average

fewer stocks as it places a greater emphasis on measuring the few stocks at the extremes.

The last two examples reviewed how a characteristic (profitability) was able to explain the next month's return for a group of stocks. In both cases it looked effective—a quintile return of 2.0% and an IC of 11%. However, in practice, it is also necessary to assess whether the characteristic was effective for not only one month, but over decades of investing encompassing multiple market cycles. To that end, during the research process a quantitative investor will look at the average quintile returns or ICs over an extended period of up to 20 years or more. When looking at these longer time series, quantitative investors use additional metrics to understand the characteristic's effectiveness.

Characteristic Testing: Key Measures of Consistency

Two commonly used measures of consistency are batting average and information ratio.

Batting Average

Batting average is a straightforward metric. In baseball a player's batting average is the number of hits divided by the number of times at bat. A similar metric is used in investing.

Batting average is the number of positive performance months (hits) divided by the number of total months (at bats). The higher the batting average, the more consistently the characteristic generates positive performance.

As Figure 11 displays, to arrive at the batting average we take the number of months the quintile return was positive divided by the number of months tested. In our example, in 47 of the 72 months profitability was effective, resulting in a positive return. This translates to a batting average of 65%, which is quite high. Imagine walking into a casino in Las Vegas where you have a 65% chance of winning every bet. That casino would not be in business very long with you at the table.

Information Ratio

Information ratio is also used to measure consistency. This measure is defined as the average return of a characteristic divided by its volatility—basically a measure of return per unit of risk or risk reward ratio. For volatility, quantitative investors use *tracking error*, which is the standard deviation of excess returns.

Figure 12 demonstrates the calculation of the information ratio. In this example, there are two characteristics. Which one should be selected?



Figure 12 Determining the Characteristic's Information Ratio

Based only on returns, we would choose characteristic 2 since it has a higher excess return (3.0%) than characteristic 1 (2.0%). However, as we can see in the figure, characteristic 2 also has much larger swings in performance than characteristic 1 and therefore more risk. The higher risk of characteristic 2 is confirmed by its high tracking error of 12.0%, three times greater than characteristic 1's tracking error of 4.0%. Characteristic 1 looks much better on a risk-adjusted basis with an information ratio of 0.50 (2.0%/4.0%) or twice characteristic 2's information ratio of 0.25 (3.0%/12.0%). So even though characteristic 1 has a lower return than characteristic 2, it also has much less risk, making it preferred since investors are rewarded more for the risk they are taking.

Model Creation

After reviewing and selecting the best characteristics, the quantitative investor then needs to assemble them into a stock selection model. This step of the research process is called model creation. It usually involves two main components:

1. Ascertaining whether the characteristics selected are not measuring the same effect (i.e., are not highly correlated).
2. Assigning weights to the selected characteristics, potentially placing greater emphasis on those in which the quantitative investor has stronger convictions.

Let us begin by discussing the first component in model creation: measuring *correlation*. When including characteristics in a stock selection model, the quantitative investor does not want to include two characteristics that have very similar performance since they may be measuring similar aspects of the company. In these cases, quantitative managers could be potentially doubling their position in a stock for the same reason. For instance, stocks with a historically high sales growth may perform similarly to stocks with high expected growth in the future, or stocks with strong gross margins may perform similarly to stocks with strong profit margins. In either case, we would not include both similar characteristics.

An example is provided in Figure 13, which shows the cumulative quintile spread return over 10 years for three characteristics (which we have labeled A, B, and C). Characteristic A did the best at differentiating the winners from losers—the stocks it liked outperformed the stocks it did not like by almost 10% over the 10-year period. Characteristic B was next with a return slightly greater than 8%, and characteristic C was the lowest with an almost 4% cumulative 10-year return. Given that all three characteristics have good performance, which two should the quantitative investor retain in the model?

Although characteristics A and B are better at differentiating winners from losers than characteristic C, A's return pattern looks very

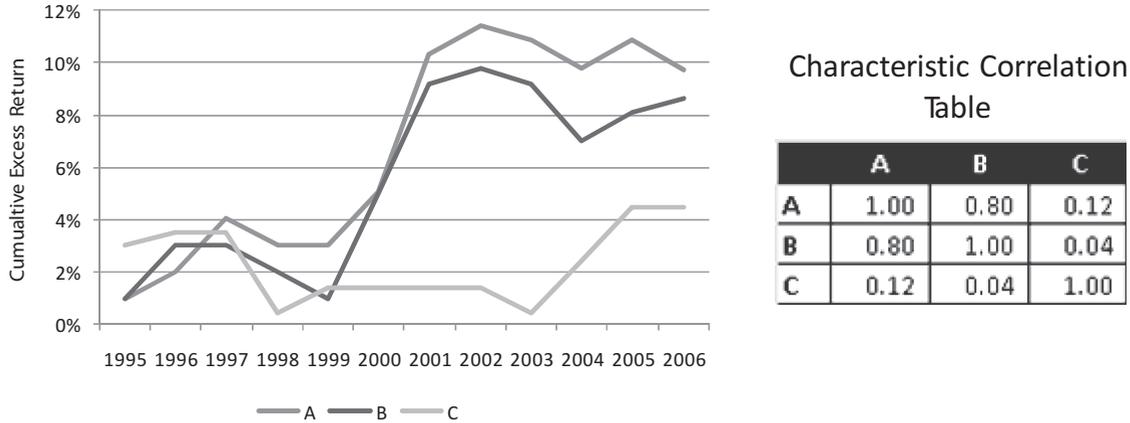


Figure 13 Model Creation: Correlation Review / Table 1 Characteristic Correlations

similar to B's. This is confirmed by Table 1 where characteristics A and B have a correlation of 0.80. Since a correlation of 1.00 means their returns move in lockstep, a correlation of 0.80 indicates they are very similar. Rather than keeping both A and B and potentially doubling our positions from similar characteristics, it would be best to keep either A or B and combine the characteristic retained with C. Even though characteristic C is the worst performing of the three, for the stock selection model C provides a good uncorrelated source of performance.

Once the characteristics to select stocks are identified, quantitative investors are ready to determine the importance or weight of each

characteristic. They must decide whether all characteristics should have the same weight or whether better characteristics should have greater weight in the stock selection model.

There are many ways to determine the weights of the characteristics. We can simply equal weight them or use some other process such as creating an algorithm, performing regressions, or optimizing. Figure 14 shows how a typical stock selection model is created. In this step, the selected characteristics are combined to determine a target for each stock whether it be a return forecast, rank, or a position size.

Once the combination of characteristics for the model is selected, the quantitative investor determines their weights and then reviews the

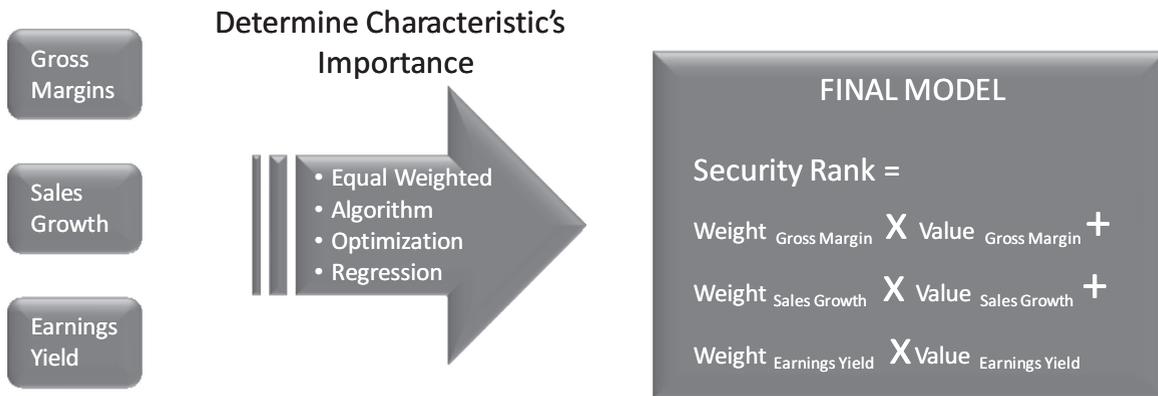


Figure 14 Stock Selection Model: Characteristic Weightings

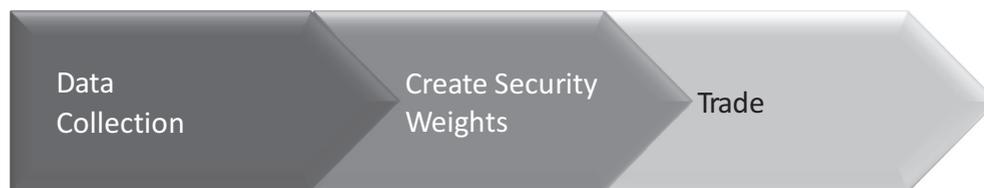


Figure 15 Three Main Steps of the Portfolio Construction Phase

model. Model review is similar to reviewing a single characteristic. The model is looked at from many perspectives, calculating all of the metrics described in Figure 8. The quantitative investor would look at how the top-quintile stocks of the model perform versus the bottom and look at information coefficients of the stock selection model over time. In addition, how much trading or turnover the stock selection model creates is reviewed or if there are any biases in the stock selections (e.g., too many small-cap stocks, or a reliance on high- or low-beta stocks). In practice, the review is much more extensive, covering many more metrics. If the stock selection model does not hold up under this final review, then the quantitative investor will need to change the stock selection model to eliminate the undesirable effects.

PORTFOLIO CONSTRUCTION

In the second phase of the investment process, the quantitative investor uses the stock selection model to buy stocks. It is in this phase that the quantitative investor puts the model into production. Returning to our golfer analogy, this is when they travel to the course to play a round of golf.

During the portfolio construction phase, the model is ready to create a daily portfolio. This phase consists of three main steps as shown in Figure 15 and described below.

Step 1: Collect data. Data are collected on a nightly basis, making sure the data are correct and do not contain any errors.

Step 2: Create security weights. New data are used to both select the stocks that should be

purchased for the portfolio as well as to specify how large its position should be.

Step 3: Trade. The stock selection model that has incorporated the most current information is used for trading.

Data Collection

As Figure 16 shows, data come from many different sources, such as a company's fundamental, pricing, economic, and other data (specialized data sources). All of these data are updated nightly, so it is important to have robust systems and processes established to handle large amounts of data, clean the data (check for errors), and process it in a timely fashion. The quantitative investor seeks to have everything ready to trade at the market opening.

Creating Security Weights

After the data are collected and verified, the next step is running all of the updated company information through the stock selection model. This will create final positions for every stock in the screened universe. In this step, each stock is ranked using the stock selection model, with the better scoring companies making it into the portfolio.

Figure 17 provides a simplified example of this, showing a stock selection model with three characteristics: gross margins, sales growth, and earnings yield (i.e., earnings-to-price ratio; the higher the ratio, the more attractively priced the stock is). From the example, Company ABC is in the top 10% of companies based on gross margin, in the top 30% in sales growth, and average on earnings yield. Company ABC

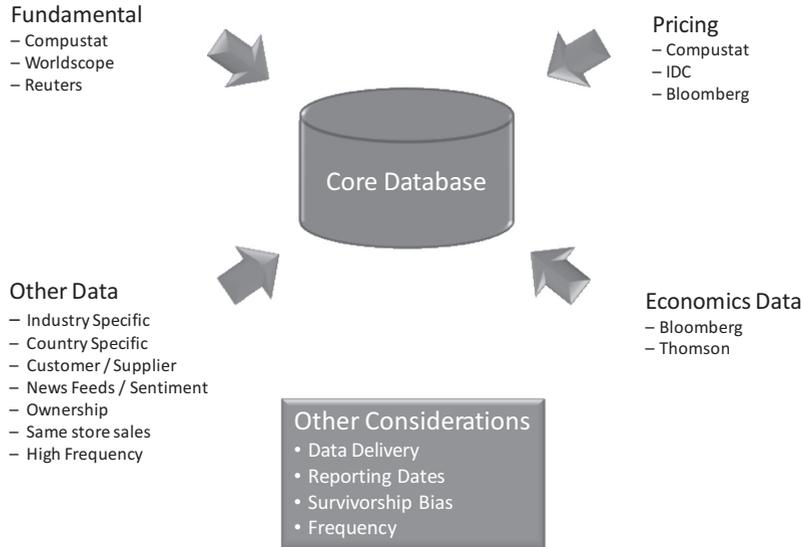


Figure 16 Data Collection Step of the Portfolio Construction Phase

may represent a company finding a profitable market and growing into it, and the rest of the market has not caught on to its prospects, so it is still valued like an average stock. In this case, the stock rates favorably by the stock selection model and would be purchased. The other stock, the stock of Company XYZ, is not as fa-

vorable and either would not be held in the portfolio or, if permitted, could be shorted. Although Company XYZ also has good margins, its growth is slowing and it is relatively expensive compared to its earnings. The company could be one that had a profitable niche, but its niche may be shrinking as sales are dwindling.



Figure 17 Creating Security Weights Step of the Portfolio Construction Phase

Furthermore, the investment community has not discounted the slowing growth and hence the stock is still expensive.

Trade

The final step in the portfolio construction process is to trade into the new positions chosen by the stock selection model. While many investment approaches trade regularly, even daily, quantitative investors tend not to. Quantitative investors tend to trade monthly or longer. They may wait for the views from their stock selection model to change significantly from their current portfolio before trading into the new views.

MONITORING

The third and final phase in the quantitative equity investment process is monitoring performance and risk. This step is important to check if any hidden biases are embedded in the portfolio and that the portfolio is performing in line with expectations. Returning one last time to our golfer analogy, this is when the golfer is making mental notes as to what is and isn't working during the round to improve his or her game in the future. This step can be broken into two activities: risk management and performance attribution.

Risk Management

In *risk management*, the main emphasis is on making sure that the quantitative investor is buying companies consistent with the stock selection model. Returning to the retail model discussed earlier in this entry, the model liked companies with good profit margins but had no view on the company's beta. So the quantitative investor would want to make sure that the companies included in the portfolio have high profit margins but average beta. If the portfolio started to include high-beta stocks, the quantitative investor would want to make adjustments to the process to eliminate this high-beta

bias. There are many types of risk management software and techniques that can be used to detect any hidden risks embedded in the portfolio and provide ways to remedy those identified.

Another aspect of risk management is to make sure that the portfolio's risk level is consistent with the modeling phase. The quantitative investor wants to ensure that the tracking error is not too high or low relative to expectations. Again, risk management techniques and software can be used to monitor tracking error and sources of tracking error, and to remedy any deviations from expectations.

Performance Attribution

Performance attribution is critical in ensuring that the actual live portfolio's performance is coming from the characteristics in the stock selection model and is in line with performance expected during the modeling stage. Performance attribution is like monitoring a car's gas mileage: If the gas mileage begins to dip below what the driver expects, or what it is known to be, then the driver would want to look under the car's hood. Similarly, if the stock selection model is not producing the desired results, or the results have changed, then the quantitative investor would need to look under the hood of the stock selection model. If performance is not being generated from the selected characteristics, then the quantitative manager would want to check out the model in more detail. One possibility is that another characteristic is canceling the desired characteristics, or the model should be providing more weight to the desired characteristic.

The monitoring phase is critical in making sure that the stock selection model is being implemented as expected.

CURRENT TRENDS

Let's look at some recent trends in the quantitative investment industry.

Many quantitative equity investors are looking for additional sources of alpha by using

alternative data sources to help select stocks. One notable source is industry-specific data (e.g., banking, airlines, and retail). Additionally, quantitative investors are turning to the Internet to better understand news flows for companies through Web-based search engines. Furthermore, quantitative investors are using more conditioning models. Conditioning occurs when two characteristics are combined rather than choosing them side by side in a stock selection model. Traditional models would look for companies that have either attractive margins or growth. With conditioning models, companies that have both attractive margins and growth are sought.

Dynamic modeling is gaining renewed popularity. It consists of timing characteristics, determining when they should enter or leave a stock selection model based on business cycle analysis, technical market indicators, or other information. For instance, during recessionary periods, a quantitative investor may want companies with strong profitability, while in expansionary periods companies with good growth prospects are sought. A stock selection model would contain profitability when the economy is entering a recession, and then include the growth characteristic once it felt that the economy is moving into an expansionary period. This is an example of how quantitative investors are bringing more personal judgment to the process, similar to fundamental investors.

Finally, with the advent of high-frequency trading and more advanced trading analytics, many quantitative investors are reviewing how best to implement their stock selection models. Some characteristics such as earnings surprise may have short-lived alpha prospects, so quantitative investors would want to trade into these stocks more quickly. Other characteristics are longer term in nature, such as valuation metrics, so investors would not have to trade into companies with attractive valuations as quickly. Furthermore, trading costs are being measured with greater granularity, allowing quantitative investors to measure transaction cost and incor-

porate these better estimates into their research modeling phase.

KEY POINTS

- Investing begins with processing many different types of information to find the most attractively priced assets. Fundamental and quantitative investors differ in their approach to the available information. The fundamental investor's primary focus is on a single company at a time, while the quantitative investor's primary focus is on a single characteristic at a time.
- Quantitative and fundamental approaches are complementary. By combining the two approaches you can obtain a more well-rounded investment process including breadth and depth in analysis, facts based on human judgment, a past and future perspective of a company, and a more well-rounded view of risk and performance of the portfolio.
- The quantitative equity investment process is made up of three phases: research, portfolio construction, and monitoring. During the research phase, the stock selection model is created. During the portfolio construction phase, the quantitative investor "productionalizes" the stock selection model or gets it ready to invest in a live portfolio. Finally, during the monitoring phase, the quantitative investor makes sure the portfolio is performing as expected.
- At the heart of the quantitative equity investment process is the stock selection model. The model includes those characteristics that are best at delineating the highest from lowest returning stocks. Models can be created for industries, sectors, or styles.
- Two common metrics used to judge a characteristic's effectiveness are quintile returns and information coefficients. Two more metrics used to understand the consistency of a characteristic's performance over time are batting average and information ratio.

- During the portfolio construction phase, data are collected from multiple sources and run through the investor's stock selection model to arrive at a list of buy and sell candidates. The buy candidates will have the strongest characteristic values in the investor's stock selection model, and the sell candidates the weakest characteristic values.
- The monitoring phase is when the investor assures that the performance in the portfolio is consistent with expectations. During this phase, the investor will make sure there are no hidden bets in the portfolio and that the characteristics in the stock selection model are performing as expected.

NOTES

1. Throughout the entry we discuss whether characteristics can separate a stock with strong future returns from one with weak future returns. Many times reference will be made to a "strong" characteristic that can differentiate the strong- from weak-performing stocks.
2. In this entry, the term "characteristic" means the attributes that differentiate companies. Quantitative investors often refer to these same characteristics as factors or signals which they typically use in their stock selection model.

Quantitative Equity Portfolio Management

ANDREW ALFORD, PhD

Managing Director, Quantitative Investment Strategies, Goldman Sachs Asset Management

ROBERT JONES, CFA

Chairman, Arwen Advisors and Chairman and CIO, System Two Advisors

TERENCE LIM, PhD, CFA

CEO, Arwen Advisors

Abstract: Equity portfolio management has evolved considerably since the 1950s. Portfolio theories and asset pricing models, in conjunction with new data sources and powerful computers, have revolutionized the way investors select stocks and create portfolios. Consequently, what was once mostly an art is increasingly becoming a science: Loose rules of thumb are being replaced by rigorous research and complex implementation. While greatly expanding the frontiers of finance, these advances have not necessarily made it any easier for portfolio managers to outperform the market. The two approaches to equity portfolio management are the traditional approach and the quantitative approach. Despite the contrasting of these two approaches by their advocates, they actually share many traits such as applying economic reasoning to identify a small set of key drivers of equity values, using observable data to quantify these key drivers, using expert judgment to develop ways to map these key drivers into the final stock-selection decision, and evaluating their performance over time. The difference in the two approaches is how they perform these tasks.

Equity portfolio management has evolved considerably since Benjamin Graham and David Dodd published their classic text on security analysis in 1934 (Graham and Dodd, 1934). For one, the types of stocks available for investment have shifted dramatically, from companies with mostly physical assets (such as railroads and utilities) to companies with mostly intangible assets (such as technology stocks and pharma-

ceuticals). Moreover, theories such as the modern portfolio theory and the capital asset pricing model, in conjunction with new data sources and powerful computers, have revolutionized the way investors select stocks and create portfolios. Consequently, what was once mostly an art is increasingly becoming a science: Loose rules of thumb are being replaced by rigorous research and complex implementation.

Of course, these new advances, while greatly expanding the frontiers of finance, have not necessarily made it any easier for portfolio managers to beat the market. In fact, the increasing sophistication of the average investor has probably made it more difficult to find—and exploit—pricing errors. Several studies show that a majority of professional money managers have been unable to beat the market (see, for example, Malkiel, 1995). There are no sure bets, and mispricings, when they occur, are rarely both large and long lasting. Successful managers must therefore constantly work to improve their existing strategies and to develop new ones. Understanding fully the equity management process is essential to accomplishing this challenging task.

These new advances, unfortunately, have also allowed some market participants to stray from a sound investment approach. It is now easier than ever for portfolio managers to use biased, unfamiliar, or incorrect data in a flawed strategy, one developed from untested conjecture or haphazard trial and error. Investors, too, must be careful not to let the abundance of data and high-tech techniques distract them when allocating assets and selecting managers. In particular, investors should not allow popular but narrow rankings of short-term performance obscure important differences in portfolio managers' style exposure or investment process. To avoid these pitfalls, it helps to have a solid grasp of the constantly advancing science of equity investing.

This entry provides an overview of equity portfolio management aimed at current and potential investors, analysts, investment consultants, and portfolio managers. We begin with a discussion of the two major approaches to equity portfolio management: the *traditional approach* and the *quantitative approach*. The remaining sections of the entry are organized around four major steps in the investment process: (1) forecasting the unknown quantities needed to manage equity portfolios—returns, risks, and transaction costs; (2) constructing portfo-

lios that maximize expected risk-adjusted return net of transaction costs; (3) trading stocks efficiently; and (4) evaluating results and updating the process.

These four steps should be closely integrated: The return, risk, and transaction cost forecasts, the approach used to construct portfolios, the way stocks are traded, and performance evaluation should all be consistent with one another. A process that produces highly variable, fast-moving return forecasts, for example, should be matched with short-term risk forecasts, relatively high transaction costs, frequent rebalancing, aggressive trading, and short-horizon performance evaluation. In contrast, stable, slower-moving return forecasts can be combined with longer term risk forecasts, lower expected transaction costs, less frequent rebalancing, more patient trading, and longer-term evaluation. Mixing and matching incompatible approaches to each part of the investment process can greatly reduce a manager's ability to reap the full rewards of an investment strategy.

A well-structured investment process should also be supported by sound economic logic, diverse information sources, and careful empirical analysis that together produce reliable forecasts and effective implementation. And, of course, a successful investment process should be easy to explain; marketing professionals, consultants, and investors all need to understand a manager's process before they will invest in it.

TRADITIONAL AND QUANTITATIVE APPROACHES TO EQUITY PORTFOLIO MANAGEMENT

At one level, there are as many ways to manage portfolios as there are portfolio managers. After all, developing a unique and innovative investment process is one of the ways managers distinguish themselves from their peers. Nonetheless, at a more general level, there are

two basic approaches used by most managers: The traditional approach and the quantitative approach. Although these two approaches are often sharply contrasted by their proponents, they actually share many traits. Both apply economic reasoning to identify a small set of key drivers of equity values; both use observable data to help measure these key drivers; both use expert judgment to develop ways to map these key drivers into the final stock-selection decision; and both evaluate their performance over time. What differs most between traditional and quantitative managers is how they perform these steps.

Traditional managers conduct stock-specific analysis to develop a subjective assessment of each stock's unique attractiveness. Traditional managers talk with senior management, closely study financial statements and other corporate disclosures, conduct detailed, stock-specific competitive analysis, and usually build spreadsheet models of a company's financial statements that provide an explicit link between various forecasts of financial metrics and stock prices. The traditional approach involves detailed analysis of a company and is often well equipped to cope with data errors or structural changes at a company (e.g., restructurings or acquisitions). However, because the traditional approach relies heavily on the judgment of analysts, it is subject to potentially severe subjective biases such as selective perception, hindsight bias, stereotyping, and overconfidence that can reduce forecast quality. (For a discussion of the systematic errors in judgment and probability assessment that people frequently make, see Kahneman, Slovic, and Tversky, 1982.) Moreover, the traditional approach is costly to apply, which makes it impracticable for a large investment universe comprising many small stocks. The high cost and subjective nature also make it difficult to evaluate, because it is hard to create the history necessary for testing. Testing an investment process is important because it helps to distinguish factors that are reflected in stock prices from those that are not. Only

factors that are not yet impounded in stock prices can be used to identify profitable trading opportunities. Failure to distinguish between these two types of factors can lead to the familiar "good company, bad stock" problem in which even a great company can be a bad investment if the price paid for the stock is too high.

Quantitative managers use statistical models to map a parsimonious set of measurable factors into objective forecasts of each stock's return, risk, and cost of trading. The quantitative approach formalizes the relation between the key factors and forecasts, which makes the approach transparent and largely free of subjective biases. Quantitative analysis can also be highly cost effective. Although the fixed costs of building a robust quantitative model are high, the marginal costs of applying the model, or extending it to a broader investment universe, are low. Consequently, quantitative portfolio managers can choose from a large universe of stocks, including many small and otherwise neglected stocks that have attractive fundamentals. Finally, because the quantitative approach is model-based, it can be tested historically on a wide cross-section of stocks over diverse economic environments. While quantitative analysis can suffer from specification errors and overfitting, analysts can mitigate these errors by following a well-structured and disciplined research process.

On the negative side, quantitative models can be misleading when there are bad data or significant structural changes at a company (that is, "garbage in, garbage out"). For this reason, most quantitative managers like to spread their bets across many names so that the success of any one position will not make or break the strategy. Traditional managers, conversely, prefer to take fewer, larger bets given their detailed hands-on knowledge of the company and the high cost of analysis.

A summary of the major advantages of each approach to equity portfolio management is presented in Table 1. (Dawes, Faust, and Meehl

Table 1 Major Advantages of the Traditional and Quantitative Approaches to Equity Portfolio Management

Traditional approach	
Depth	Although they have views on fewer companies, traditional managers tend to have more in-depth knowledge of the companies they cover. Unlike a computerized model, they should know when data are misleading or unrepresentative.
Regime shifts	Traditional managers may be better equipped to handle regime shifts and recognize situations where past relationships might not be expected to continue (e.g., where back-tests may be unreliable).
Signal identification	Based on their greater in-depth knowledge, traditional managers can better understand the unique data sources and factors that are important for stocks in different countries or industries.
Qualitative factors	Many important factors that may affect an investment decision are not available in any database and are hard to evaluate quantitatively. Examples might include management and their vision for the company; the value of patents, brands, and other intangible assets; product quality; or the impact of new technology.
Quantitative approach	
Universe	Because a computerized model can quickly evaluate thousands of securities and can update those evaluations daily, it can uncover more opportunities. Further, by spreading their risk across many small bets, quantitative managers can add value with only slightly favorable odds.
Discipline	While individuals often base decisions on only the most salient or distinctive factors, a computerized model will simultaneously evaluate all specified factors before reaching a conclusion.
Verification	Before using any signal to evaluate stocks, quantitative managers will normally backtest its historical efficacy and robustness. This provides a framework for weighting the various signals.
Risk management	By its nature, the quantitative approach builds in the notion of statistical risk and can do a better job of controlling unintended risks in the portfolio.
Lower fees	The economies of scale inherent in a quantitative process usually allow quantitative managers to charge lower fees.

[1989] provide an excellent comparison of clinical (traditional) and actuarial (quantitative) decision analysis.) Our focus in the rest of this entry is the process of quantitative equity portfolio management.

FORECASTING STOCK RETURNS, RISKS, AND TRANSACTION COSTS

Developing good forecasts is the first and perhaps most critical step in the investment process. Without good forecasts, the difficult task of forming superior portfolios becomes nearly

impossible. In this section we discuss how to use a quantitative approach to generate *forecasts of stock returns, risks, and transaction costs*. These forecasts are then used in the portfolio construction step described in the next section.

It should be noted that some portfolio managers do not develop explicit forecasts of returns, risks, and transaction costs. Instead, they map a variety of individual stock characteristics directly into portfolio holdings. However, there are limitations with this abbreviated approach. Because the returns and risks corresponding to the various characteristics are not clearly identified, it is difficult to ensure the weights placed on the characteristics

are appropriate. Further, measuring risk at the portfolio level is awkward without reliable estimates of the risks of each stock, especially the correlations between stocks. Similarly, controlling turnover is hard when returns and transaction costs are not expressed in consistent units. And, of course, it is difficult to explain a process that occurs in one magical step.

Forecasting Returns

The process of building a quantitative return-forecasting model can be divided into four closely linked steps: (1) identifying a set of potential return forecasting variables, or signals; (2) testing the effectiveness of each signal, by itself and together with other signals; (3) determining the appropriate weight for each signal in the model; and (4) blending the model's views with market equilibrium to arrive at reasonable forecasts for expected returns.

Identifying a list of potential *signals* might seem like an overwhelming task; the candidate pool can seem almost endless. To narrow the list, it is important to start with fundamental relationships and sound economics. Reports published by Wall Street analysts and books about financial statement analysis are both good sources for ideas. Another valuable resource is academic research in finance and accounting. Academics have the incentive and expertise to identify and carefully analyze new and innovative information sources. Academics have studied a large number of stock price anomalies, and Table 2 lists several that have been adopted by investment managers. (For evidence on the performance of several well-known anomalies, see Fama and French [2008].)

For portfolio managers intent on building a successful investment strategy, it is not enough to simply take the best ideas identified by others and add them to the return-forecasting model. Instead, each potential signal must be thoroughly tested to ensure it works in the context of the manager's strategy across many stocks and during a variety of economic envi-

Table 2 Selected Stock Price Anomalies Used in Quantitative Models

<i>Growth/Value:</i> Value stocks (high B/P, E/P, CF/P) outperform growth stocks (low B/P, E/P, CF/P).
<i>Post-earnings-announcement drift:</i> Stocks that announce earnings that beat expectations outperform stocks that miss expectations.
<i>Short-term price reversal:</i> One-month losers outperform one-month winners.
<i>Intermediate-term price momentum:</i> Six-months to one-year winners outperform losers.
<i>Earnings quality:</i> Stocks with cash earnings outperform stocks with non-cash earnings.
<i>Stock repurchases:</i> Companies that repurchase shares outperform companies that issue shares.
<i>Analyst earnings estimates and stock recommendations:</i> Changes in analyst stock recommendations and earnings estimates predict subsequent stock returns.

ronments. The real challenge is winnowing the list of potential signals to a parsimonious set of reliable forecasting variables. When selecting a set of signals, it is a good idea to include a variety of variables to capture distinct investment themes, including valuation, momentum, and earnings quality. By diversifying over information sources and variables, there is a good chance that if one signal fails to add value another will be there to carry the load.

When evaluating a signal, it is important to make sure the underlying data used to compute the signal are available and largely error free. Checking selected observations by hand and screening for outliers or other influential observations is a useful way to identify data problems. It is also sometimes necessary to transform a signal—for instance, by subtracting the industry mean or taking the natural logarithm—to improve the “shape” of the distribution. To evaluate a signal properly, both univariate and multivariate analysis is important. Univariate analysis provides evidence on the signal's predictive ability when the signal is used alone, whereas multivariate analysis provides evidence on the signal's incremental predictive ability above and beyond other variables considered. For both univariate and

multivariate analysis, it is wise to examine the returns to a variety of portfolios formed on the basis of the signal. Sorting stocks into quintiles or deciles is popular, as is regression analysis, where the coefficients represent the return to a portfolio with unit exposure to the signal. These portfolios can be equal weighted, cap weighted, or even risk weighted depending on the model's ultimate purpose. Finally, the return forecasting model should be tested using a realistic simulation that controls the target level of risk, takes account of transaction costs, and imposes appropriate constraints (e.g., the nonnegativity constraint for long-only portfolios). In our experience, many promising return-forecasting signals fail to add value in realistic back-tests—either because they involve excessive trading; work only for small, illiquid stocks; or contain information that is already captured by other components of the model.

The third step in building a return forecasting model is determining each signal's weight. When computing expected returns, more weight should be put on signals that, over time, have been more stable; generated higher and more consistent returns; and provided superior diversification benefits. Maintaining exposures to signals that change slowly requires less trading, and hence lower costs, than is the case for signals that change rapidly. Other things being equal, a stable signal (such as the ratio of book-to-market equity) should get more weight than a less stable signal (such as one-month price reversal). High, consistent returns are essential to a profitable, low-risk investment strategy; hence, signals that generate high returns with little risk should get more weight than signals that produce lower returns with higher risk. Finally, signals with more diversified payoffs should get more weight because they can hedge overall performance when other signals in the model perform poorly.

The last step in forecasting returns is to make sure the forecasts are reasonable and internally consistent by comparing them with equilibrium views. Return forecasts that ignore equilibrium

expectations can create problems in the portfolio construction step. Seemingly reasonable return forecasts can cause an optimizer to maximize errors rather than expected returns, producing extreme, unbalanced portfolios. The problem is caused by return forecasts that are inconsistent with the assumed correlations across stocks. If two stocks (or subportfolios) are highly correlated, then the equilibrium expectation is that their returns should be similar; otherwise, the optimizer will treat the pair of stocks as a (near) arbitrage opportunity by going extremely long the high-return stock and extremely short the low-return stock. However, with hundreds of stocks, it is not always obvious whether certain stocks, or combinations of stocks, are highly correlated and therefore ought to have similar return forecasts. The Black-Litterman model was specifically designed to alleviate this problem. It blends a model's raw return forecasts with *equilibrium expected returns*—which are the returns that would make the benchmark optimal for a given risk model—to produce internally consistent return forecasts that reflect the manager's (or model's) views yet are consistent with the risk model. (For a discussion of how to use the Black-Litterman model to incorporate equilibrium views into a return-forecasting model, see Litterman [2003].)

Forecasting Risks

In a portfolio context, the risk of a single stock is a function of the variance of its returns, as well as the covariances between its returns and the returns of other stocks in the portfolio. The variance-covariance matrix of stock returns, or risk model, is used to measure the risk of a portfolio. For equity portfolio management, investors rarely estimate the full variance-covariance matrix directly because the number of individual elements is too large, and for a well-behaved (that is, non-singular) matrix, the number of observations used to estimate the matrix must significantly

exceed the number of stocks in the matrix. To see this, suppose that there are N stocks. Then the variance-covariance matrix has $N(N + 1)/2$ elements, consisting of N variances and $N(N - 1)/2$ covariances. For an S&P 500 portfolio, for instance, there are $500 \times (500 + 1)/2 = 125,250$ unknown parameters to estimate, 500 variances and 124,750 covariances. For this reason, most equity portfolio managers use a *factor risk model* in which individual variances and covariances are expressed as a function of a small set of stock characteristics—such as industry membership, size, and leverage. This greatly reduces the number of unknown risk parameters that the manager needs to estimate.

When developing an equity factor risk model, it is a good idea to include all of the variables used to forecast returns among the (potentially larger) set of variables used to forecast risks. This way, the risk model “sees” all of the potential risks in an investment strategy, both those managers are willing to accept and those they would like to avoid. Further, a mismatch between the variables in the return and risk models can produce less efficient portfolios in the optimizer. For instance, suppose a return model comprises two factors, each with 50% weight: the book-to-price ratio (B/P) and return on equity (ROE). Suppose the risk model, on the other hand, has only one factor: B/P. When forming a portfolio, the optimizer will manage risk only for the factors in the risk model—that is, B/P but not ROE. This inconsistency between the return and risk models can lead to portfolios with extreme positions and higher-than-expected risk. The portfolio will not reflect the original 50-50 weights on the two return factors because the optimizer will dampen the exposure to B/P, but not to ROE. In addition, the risk model’s estimate of tracking error will be too low because it will not capture any risk from the portfolio’s exposure to ROE. The most effective way to avoid these two problems is to make sure all of the factors in the return model are also included in the risk model (although the converse does not need to be true—that

is, there can be risk factors without expected returns).

A final issue to consider when developing or selecting a risk model is the frequency of data used in the estimation process. Many popular risk models use monthly returns, whereas some portfolio managers have developed proprietary risk models that use daily returns. Clearly, when estimating variances and covariances, the more observations, the better. High-frequency data produce more observations and hence more precise and reliable estimates. Further, by giving more weight to recent observations, estimates can be more responsive to changing economic conditions. As a result, risk models that use high-frequency returns should provide more accurate risk estimates. (For a detailed discussion of factor risk models, see Chapter 20 of Litterman [2003]).

Forecasting Transaction Costs

Although often overlooked, accurate trade-cost estimates are critical to the equity portfolio management process. After all, what really matters is not the gross return a portfolio might receive, but rather the actual return a portfolio does receive after deducting all relevant costs, including transaction costs. Ignoring transaction costs when forming portfolios can lead to poor performance because implementation costs can reduce, or even eliminate, the advantages achieved through superior stock selection. Conversely, taking account of transaction costs can help produce portfolios with gross returns that exceed the costs of trading.

Accurate trading-cost forecasts are also important after portfolio formation, when monitoring the realized costs of trading. A good transaction-cost model can provide a benchmark for what realized costs “should be,” and hence whether actual execution costs are reasonable. Detailed trade-cost monitoring can help traders and brokers achieve best execution by driving improvements in trading methods—such as more patient trading,

or the selective use of alternative trading mechanisms.

Transaction costs have two components: (1) explicit costs, such as commissions and fees; and (2) implicit costs, or market impact. Commissions and fees tend to be relatively small, and the cost per share does not depend on the number of shares traded. In contrast, market impact costs can be substantial. They reflect the costs of consuming liquidity from the market, costs that increase on a per-share basis with the total number of shares traded.

Market impact costs arise because suppliers of liquidity incur risk. One component of these costs is inventory risk. The liquidity supplier has a risk/return trade-off, and will demand a price concession to compensate for this inventory risk. The larger the trade size and the more illiquid or volatile the stock, the larger are inventory risk and market impact costs. Another consideration is adverse selection risk. Liquidity suppliers are willing to provide a better price to uninformed than informed traders, but since there is no reliable way to distinguish between these two types of traders, the market maker sets an average price, with expected gains from trading with uninformed traders compensating for losses incurred from trading with informed traders. Market impact costs tend to be higher for low-price and small-cap stocks for which greater adverse selection risk and informational asymmetry tend to be more severe.

Forecasting price impact is difficult. Because researchers only observe prices for completed trades, they cannot determine what a stock's price would have been without these trades. It is therefore impossible to know for sure how much prices moved as a result of the trade. Price impact costs, then, are statistical estimates that are more accurate for larger data samples.

One approach to estimating trade costs is to directly examine the complete record of market prices, tick by tick (see, for example, Breen, Hodrick, and Korajczyk [2002]). These data are noisy due to discrete prices, non-synchronous reporting of trades and quotes, and input er-

rors. Also, the record does not show orders placed, just those that eventually got executed (which may have been split up from the original, larger order). Research by Lee and Radhakrishna (2000) suggests empirical analysis should be done using aggregated samples of trades rather than individual trades at the tick-by-tick level.

Another approach is for portfolio managers to estimate a proprietary transaction cost model using their own trades and, if available, those of comparable managers. If generating a sufficient sample is feasible, this approach is ideal because the resulting model matches the stock characteristics, investment philosophy, and trading strategy of the individual portfolio manager. There is a large academic literature on measuring transaction costs. Further, models built from actual trading records provide a complementary source of information on market impact costs. (For empirical evidence on how transaction costs can vary across trade characteristics and how to predict transaction costs, see Chapter 23 of Litterman [2003].)

CONSTRUCTING PORTFOLIOS

In this section we discuss how to construct portfolios based on the forecasts described in the last section. In particular, we compare ad hoc, rule-based approaches to portfolio optimization. The first step in portfolio construction, however, is to specify the investment goals. While having good forecasts (as described in the previous section) is obviously important, the investor's goals define the portfolio management problem. These goals are usually specified by three major parameters: the benchmark, the risk/return target, and specific restrictions such as the maximum holdings in any single name, industry, or sector.

The benchmark represents the starting point for any active portfolio; it is the client's neutral position—a low-cost alternative to active

management in that asset class. For example, investors interested in holding large-cap U.S. stocks might select the S&P 500 or Russell 1000 as their benchmark, while investors interested in holding small-cap stocks might choose the Russell 2000 or the S&P 600. Investors interested in a portfolio of non-U.S. stocks could pick the FTSE 350 (United Kingdom), TOPIX (Japan), or MSCI EAFE (World minus North America) indexes. There are a large number of published benchmarks available, or an investor might develop a customized benchmark to represent the neutral position. In all cases, however, the benchmark should be a reasonably low-cost, investable alternative to active management.

Although some investors are content to merely match the returns on their benchmarks, most investors allocate at least some of their assets to active managers. The allocation of risk is done via *risk budgeting*. In equity portfolio management, active management means overweighting attractive stocks and underweighting unattractive stocks relative to their weights in the benchmark. (The difference between a stock's weight in the portfolio and its weight in the benchmark is called its active weight, where a positive active weight corresponds to an overweight position and a negative active weight corresponds to an underweight position.) Of course, there is always a chance that these active weighting decisions will cause the portfolio to underperform the benchmark, but one of the basic dictums of modern finance is that to earn higher returns, investors must accept higher risk—which is true of active returns as well as total returns.

A portfolio's *tracking error* measures its risk relative to a benchmark. Tracking error equals the time-series standard deviation of a portfolio's *active return* (which is the difference between the portfolio's return and that of the benchmark). A portfolio's *information ratio* equals its average active return divided by its tracking error. As a measure of return per unit of risk, the information ratio provides a conve-

nient way to compare strategies with different active risk levels.

An *efficient portfolio* is one with the highest expected return for a target level of risk—that is, it has the highest information ratio possible given the risk budget. In the absence of constraints, an efficient portfolio is one in which each stock's marginal contribution to expected return is proportional to its marginal contribution to risk. That is, there are no unintended risks, and all risks are compensated with additional expected returns. How can a portfolio manager construct such an efficient portfolio? Below we compare two approaches: (1) a rule-based system; and (2) portfolio optimization.

Building an efficient portfolio is a complex problem. To help simplify this complicated task, many portfolio managers use ad hoc, rule-based methods that partially control exposures to a small number of risk factors. For example, one common approach—called stratified sampling—ranks stocks within buckets formed on the basis of a few key risk factors, such as sector and size. The manager then invests more heavily in the highest-ranked stocks within each bucket, while keeping the portfolio's total weight in each bucket close to that of the benchmark. The resulting portfolio is close to neutral with respect to the identified risk factors (that is, sector and size) while overweighting attractive stocks and underweighting unattractive stocks.

Although stratified sampling may seem sensible, it is not very efficient. Numerous unintended risks can creep into the portfolio, such as an overweight in high-beta stocks, growth stocks, or stocks in certain subsectors. Nor does it allow the manager to explicitly consider trading costs or investment objectives in the portfolio construction problem. Portfolio optimization provides a much better method for balancing expected returns against different sources of risk, trade costs, and investor constraints. An optimizer uses computer algorithms to find the set of weights (or holdings) that maximize the portfolio's expected return

(net of trade costs) for a given level of risk. It minimizes uncompensated sources of risk, including sector and style biases. Fortunately, despite the complex math, optimizers require only the various forecasts we've already described and developed in the prior section.

Chapter 23 of Litterman (2003) demonstrates the benefits of optimization, comparing two portfolios: one constructed using stratified sampling and the other constructed using an optimizer. The *optimized portfolio* is designed to have the same predicted tracking error as the *rule-based portfolio*. The results show that (1) the optimized portfolio is more efficient in terms of its expected alpha and information ratio for the same level of risk, (2) risk is spread more broadly for the optimized portfolio compared to the rule-based portfolio, (3) more of the risk budget in the optimized portfolio is due to the factors that are expected to generate positive excess returns, and (4) the forecast beta for the optimized portfolio is closer to 1.0, as unintended sources of risk (such as the market timing) are minimized.

Another benefit of optimizers is that they can efficiently account for transaction costs, constraints, selected restrictions, and other account guidelines, making it much easier to create customized client portfolios. Of course, when using an optimizer to construct efficient portfolios, reliable inputs are essential. Data errors that add noise to the return, risk, and transaction cost forecasts can lead to portfolios in which these forecast errors are maximized. Instead of picking stocks with the highest actual expected returns, or the lowest actual risks or transaction costs, the optimizer takes the biggest positions in the stocks with the largest errors, namely, the stocks with the greatest overestimates of expected returns or the greatest underestimates of risks or transaction costs. A robust investment process will screen major data sources for outliers that can severely corrupt one's forecasts. Further, as described in the previous section, return forecasts should be adjusted for equilibrium views using the

Black-Litterman model to produce final return forecasts that are more consistent with risk estimates, and with each other. Finally, portfolio managers should impose sensible, but simple, constraints on the optimizer to help guard against the effects of noisy inputs. These constraints could include maximum active weights on individual stocks, industries, or sectors, as well as limitations on the portfolio's active exposure to factors such as size or market beta.

TRADING

Trading is the process of executing the orders derived in the portfolio construction step. To trade a list of stocks efficiently, investors must balance opportunity costs and execution price risk against market impact costs. Trading each stock quickly minimizes lost alpha and price uncertainty due to delay, but impatient trading incurs maximum market impact. However, trading more patiently over a longer period reduces market impact but incurs larger opportunity costs and short-term execution price risk. Striking the right balance is one of the keys to successful trade execution.

The concept of "striking a balance" suggests optimization. Investors can use a trade optimizer to balance the gains from patient trading (e.g., lower market-impact cost) against the risks (e.g., greater deviation between the execution price and the decision price; potentially higher short-term tracking error). Such an optimizer will tend to suggest aggressive trading for names that are liquid and/or have a large effect on portfolio risk, while suggesting patient trading for illiquid names that have less impact on risk. A trade optimizer can also easily handle most real-world trading constraints, such as the need to balance cash in each of many accounts across the trading period (which may last several days).

A trade optimizer can also easily accommodate the time horizon of a manager's views. That is, if a manager is buying a stock primarily for long-term valuation reasons, and the excess

return is expected to accrue gradually over time, then the optimizer will likely suggest a patient trading strategy (all else being equal). Conversely, if the manager is buying a stock in expectation of a positive earnings surprise tomorrow, the optimizer is likely to suggest an aggressive trading strategy (again, all else being equal). The trade optimizer can also be programmed to consider short-term return regularities, such as the tendency of stocks with dramatic price moves on one day to continue those moves on the next day before reversing the following day (see Heston, Korajczyk, and Sadka, 2010). Although these types of regularities may be too small to cover trading costs, and should not be used to initiate trades, they can be used to help minimize trading costs after an investor has independently decided to trade (see Engle and Ferstenberg, 2007).

To induce traders to follow the desired strategy (that is, that suggested by the trade optimizer), the portfolio manager needs to give the trader an appropriate benchmark, which provides guidance about how aggressively or patiently to trade. Two widely used benchmarks for aggressive trades are the closing price on the previous day and the opening price on the trade date. Because the values of these two benchmarks are measured prior to any trading, a patient strategy that delays trading heightens execution price risk by increasing the possibility of deviating significantly from the benchmark. Another popular execution benchmark is the volume-weighted average price (VWAP) for the stock over the desired trading period, which could be a few minutes or hours for an aggressive trade, or one or more days for a patient trade. However, the VWAP benchmark should only be used for trades that are not too large relative to total volume over the period; otherwise, the trader may be able to influence the benchmark against which he or she is evaluated.

Buy-side traders can increasingly make use of algorithmic trading, or computer algorithms that directly access market exchanges, to auto-

matically make certain trading decisions such as the timing, price, quantity, type, and routing of orders. These algorithms may dynamically monitor market conditions across time and trading venues, and reduce market impact by breaking large orders into smaller pieces, employing either limit orders or marketable limit orders, or selecting trading venues to submit orders, while closely tracking trading benchmarks. Algorithmic trading provides buy-side traders more anonymity and greater control over their order flow, but tends to work better for more liquid or patient trades.

Principal package trading is another way to lower transaction costs relative to traditional agency methods (see Kavajecz and Keim, 2005). Principal trades may be crossed with the principal's existing inventory positions, or allow the portfolio manager to benefit from the longer trading horizon and superior trading ability of certain intermediaries.

EVALUATING RESULTS AND UPDATING THE PROCESS

Once an investment process is up and running, it needs to be constantly reassessed and, if necessary, refined. The first step is to compare actual results to expectations; if realizations differ enough from expectations, process refinements may be necessary. Thus, managers need systems to monitor realized performance, risk, and trading costs and compare them to prior expectations.

A good performance monitoring system should be able to determine not only the degree of over- or under-performance, but also the sources of these excess returns. For example, a good performance attribution system might break excess returns down into those due to market timing (having a different beta than the benchmark), industry tilts, style differences, and stock selection. Such systems are available from a variety of third-party vendors. An even better system would allow the manager to

further disaggregate returns to see the effects of each of the proprietary signals used to forecast returns, as well as the effects of constraints and other portfolio requirements. And, of course, any system will be more accurate if it can account for daily trading and changes in portfolio exposures.

Investors should also compare realized risks to expectations. For example, Goldman Sachs has developed the concept of the green, yellow, and red zones to compare realized and targeted levels of risk (see Chapter 17 in Litterman, 2003). Essentially, if realized risk is within a reasonable band around the target (that is, the green zone), then one can assume the risk management techniques are working as intended and no action is required. If realized risk is further from the target (the yellow zone), the situation may require closer examination, and if realized risk is far from the target (the red zone), some action is usually called for.

Finally, it is important to monitor trading costs. Are they above or below the costs assumed when making trading decisions? Are they above or below competitors' costs? Are they too high in an absolute sense? If so, managers may need to improve their trade cost estimates, trading process, or both. There are many services that can report realized trade costs, but most are available with a significant lag, and are inflexible with respect to how they measure and report these costs. With in-house systems, however, managers can compare a variety of trade cost estimation techniques and get the feedback in a timely enough fashion to act on the results.

The critical question, of course, is what to do with the results of these monitoring systems: When do variations from expectations warrant refinements to the process? This will depend on the size of the variations and their persistence. For example, a manager probably would not throw out a stock-selection signal after one bad month—no matter how bad—but might want to reconsider after many years of poor performance, taking into consideration the eco-

nomic environment and any external factors that might explain the results. It is also important to compare the underperformance to historical simulations. Have similar periods occurred in the past, and if so, were they followed by improvements? In this case, the underperformance is part of the normal risk in that signal and no changes may be called for. If not, there may have been a structural change that might invalidate the signal going forward—for example, if the signal has become overly popular, it may no longer be a source of mispricing.

Similarly, the portfolio manager needs to consider the source of any differences between expectations and realizations. For example, was underperformance due to faulty signals, portfolio constraints, unintended risk, or random noise? The answer will determine the proper response. If constraints are to blame, they may be lifted—but only if doing so would not violate any investment guidelines or incur excessive risk. Alternatively, if the signals are to blame, the manager must decide if the deviations from expectations are temporary or more enduring. If it is just random noise, no action is necessary. Similarly, any differences between realized and expected risk could be due to poor risk estimates or poor portfolio construction, with the answer determining the response. Finally, excessive trading costs (versus expectations) could reflect poor trading or poor trade cost estimates, again with different implications for action.

In summary, ongoing performance, risk, and trade cost monitoring is an integral part of the equity portfolio management process and should get equal billing with forecasting, portfolio construction, and trading. Monitoring serves as both quality control and a source of new ideas and process improvements. The more sophisticated the monitoring systems, the more useful they are to the process. And although the implications of monitoring involve subtle judgments and careful analysis, better data can lead to better solutions.

KEY POINTS

- Two popular ways to manage equity portfolios are the traditional, or qualitative, approach and the quantitative approach.
- The equity investment process comprises four primary steps: (1) forecasting returns, risks, and transaction costs; (2) constructing portfolios that maximize expected risk-adjusted return net of transaction costs; (3) trading stocks efficiently; and (4) evaluating results and updating the process.
- There are four closely linked steps to building a quantitative equity return-forecasting model: (1) identifying a set of potential return forecasting variables, or signals; (2) testing the effectiveness of each signal, by itself and together with other signals; (3) determining the appropriate weight for each signal in the model; and (4) blending the model's views with market equilibrium to arrive at reasonable forecasts for expected returns.
- Most quantitative equity portfolio managers use a factor risk model in which individual variances and covariances are expressed as a function of a small set of stock characteristics such as industry membership, size, and leverage.
- Transaction costs consist of explicit costs, such as commissions and fees, and implicit costs, or market impact. The per-share cost of commissions and fees does not depend on the number of shares traded, whereas market impact costs increase on a per-share basis with the total number of shares traded.
- Tracking error measures a portfolio's risk relative to a benchmark. Tracking error equals the time-series standard deviation of a portfolio's active return, the difference between the portfolio's return and that of the benchmark.
- Information ratio is a measure of return per unit of risk, a portfolio's average active return divided by its tracking error.
- Two widely used ways to construct an efficient portfolio are stratified sampling, which

is a rule-based system, and portfolio optimization.

- To trade a list of stocks efficiently, investors must balance opportunity costs and execution price risk against market impact costs. Trading each stock quickly minimizes lost alpha and price uncertainty due to delay, but impatient trading incurs maximum market impact. Trading more patiently over a longer period reduces market impact but incurs larger opportunity costs and short-term execution price risk.
- Once an investment process is operational, it should be constantly reassessed and, if necessary, refined. Thus, managers need systems to monitor realized performance, risk, and trading costs and compare them to prior expectations.
- A good performance monitoring system should be able to determine the degree of over- or underperformance as well as the sources of these excess returns, such as market timing, industry tilts, style differences, and stock selection.

REFERENCES

- Breen, W., Hodrick, L. S., and Korajczyk, R. A. (2000). Predicting equity liquidity. *Management Science* 48, 4: 470–483.
- Dawes, R. M., Faust, E., and Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science* 243 (March 31): 1668–1674.
- Engle, R. F., and Ferstenberg, R. (2007). Execution risk. *The Journal of Portfolio Management*, 33–44.
- Fama, E. F., and French, K. R. (2008). Dissecting anomalies. *Journal of Finance*, 1653–1678.
- Graham, B., and Dodd, D. (1934). *Security Analysis*, 1st edition. New York: McGraw-Hill.
- Heston, S. L., Korajczyk, R. A., and Sadka, R., (2010). Intraday patterns in the cross-section of stock returns. *Journal of Finance*, 1369–1407.
- Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgment under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press.
- Kavajecz, K. A., and Keim, D. B. (2002). Packaging liquidity: Blind auctions and transaction cost

- efficiencies. *Journal of Financial and Quantitative Analysis* 40: 465–492.
- Lee, C., and Radhakrishna, B. (2000). Inferring investor behavior: Evidence from TORQ data. *Journal of Financial Markets* 3: 83–111.
- Litterman, R. (2003). *Modern Investment Management: An Equilibrium Approach*. Hoboken, NJ: John Wiley & Sons.
- Malkiel, B.G. (1995). Returns from investing in equity mutual funds, 1971 to 1991. *Journal of Finance* 50: 549–572.

Forecasting Stock Returns

FRANK J. FABOZZI, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

PETTER N. KOLM, PhD
Director of the Mathematics in Finance Masters Program and Clinical Associate Professor,
Courant Institute of Mathematical Sciences, New York University

SERGIO M. FOCARDI, PhD
Partner, The Intertek Group

Abstract: One of the key tasks in seeking to generate attractive returns is producing realistic and reasonable return expectations and forecasts. In the Markowitz mean-variance framework, an investor's objective is to choose a portfolio of securities that has the largest expected return for a given level of risk (as measured by the portfolio's variance). In the case of common stock, by return (or expected return) of a stock, we mean the change (or expected change) in the stock price over the period, plus any dividends paid, divided by the starting price. Of course, since we do not know the true values of the securities' expected returns and covariances, these must be estimated or forecasted. Equity portfolio managers have used various statistical models for forecasting returns and risk. These models, referred to as predictive return models, make conditional forecasts of expected returns using the current information set. Predictive return models include regressive models, linear autoregressive models, dynamic factor models, and hidden-variable models.

In contrast to forecasting events such as the weather, forecasting stock prices and returns is difficult because the predictions themselves will produce market movements that in turn provoke immediate changes in prices, thereby invalidating the predictions themselves. This leads to the concept of market efficiency: An efficient market is a market where all new information about the future behavior of prices is immediately impounded in the prices themselves and therefore exploits all information.

Actually the debate about the predictability of stock prices and returns has a long history.¹ More than 75 years ago, Cowles (1933) asked the question: "Can stock market forecasters forecast?" Armed with the state-of-the-art econometric tools at the time, Cowles analyzed the recommendations of stock market forecasters and concluded, "It is doubtful." Subsequent academic studies support Cowles's conclusion. However, the history goes further back. In 1900, a French mathematician, Louis Bachelier, in his doctoral dissertation in mathematical statistics

titled *Théorie de la Spéculation* (*The Theory of Speculation*), showed using mathematical techniques why the stock market behaves as it does.² He also provided empirical evidence based on the French capital markets at the turn of the century. He wrote:

Past, present, and even discounted future events are reflected in market price, but often show no apparent relation to price changes. . . . [A]rtificial causes also intervene: the Exchange reacts on itself, and the current fluctuation is a function, not only of the previous fluctuations, but also of the current state. The determination of these fluctuations depends on an infinite number of factors; it is, therefore, impossible to aspire to mathematical predictions of it. . . . [T]he dynamics of the Exchange will never be an exact science. (Bachelier, 1900)

In other words, according to Bachelier, stock price movements are difficult to forecast and even explain after the fact.

Despite this conclusion, the adoption of modeling techniques by asset management firms has greatly increased since the turn of the century. Models to predict expected returns are routinely used at asset management firms. In most cases, it is a question of relatively simple models based on factors or predictor variables. However, more statistical or econometric-oriented models are also being experimented with and adopted by some asset management firms, as well as what are referred to as nonlinear models based on specialized areas of statistics such as neural networks and genetic algorithms.

Historical data are often used for forecasting future returns as well as estimating risk. For example, a portfolio manager might proceed in the following way: Observing weekly or monthly returns, the portfolio manager might use the past five years of historical data to estimate the expected return and the covariances by the sample mean and sample covariances. The portfolio manager would then use these as inputs for mean-variance optimization, along with any ad hoc adjustments to reflect any views about expected returns on future performance. Unfortunately this historical approach

most often leads to counterintuitive, unstable, or merely “wrong” portfolios generated by the mean-variance optimization model. Better forecasts are necessary. Statistical estimates can be very noisy and typically depend on the quality of the data and the particular statistical techniques used to estimate the inputs. In general, it is desirable that an estimator of expected return and risk have the following properties:

- It provides a forward-looking forecast with some predictive power, not just a backward-looking historical summary of past performance.
- The estimate can be produced at a reasonable computational cost.
- The technique used does not amplify errors already present in the inputs used in the process of estimation.
- The forecast should be intuitive, that is, the portfolio manager should be able to explain and justify them in a comprehensible manner.

In this entry, we look at the issue of whether forecasting stock returns can be done so as to generate trading profits and excess returns. Because the issue about predictability of stock returns or prices requires an understanding of statistical concepts, we will provide a brief description of the relevant concepts in probability theory and statistics. We then discuss the different types of predictive return models that are used by portfolio managers.

THE CONCEPT OF PREDICTABILITY

To predict (or forecast) involves forming an expectation of a future event or future events. Since ancient times it has been understood that the notion of predicting the future is subject to potential inconsistencies. Consider what might happen if one receives a highly reliable prediction that tomorrow one will have a car accident driving to work. This might alter one’s behavior such that a decision is made not to go to

work. Hence, one's behavior will be influenced by the prediction, thus potentially invalidating the prediction. It is because of inconsistencies of this type that two economists in the mid 1960s, Paul Samuelson and Eugene Fama, arrived at the apparently paradoxical conclusion that "properly anticipated prices fluctuate randomly."³

The concept of forecastability rests on how one can forecast the future given the current state of knowledge. In probability theory, the state of knowledge on a given date is referred to as the *information set* known at that date. Forecasting is the relationship between the information set today and future events. By altering the information set, the forecast changes. However, the relationship between the information set and the future is fixed and immutable. Academicians and market practitioners adopt in finance theories this concept of forecastability. Prices or returns are said to be forecastable if the knowledge of the past influences our forecast of the future. For example, if the future returns of a firm's stock depend on the value of key financial ratios, then those returns are predictable. If the future returns of that stock do not depend on any variable known today, then returns are unpredictable.

As explained in the introduction to this entry, the merits of stock return forecasting is an ongoing debate. There are two beliefs that seem to be held in the investment community. First, predictable processes allow investors to earn excess returns. Second, unpredictable processes do not allow investors to earn excess returns. Neither belief is necessarily true. Understanding why will shed some light on the crucial issues in the debate regarding return modeling. The reasons can be summed up as follows. First, predictable processes do not necessarily produce excess returns if they are associated with unfavorable risk. Second, unpredictable expectations can be profitable if the expected value is favorable.

Because most of our knowledge is uncertain, our forecasts are also uncertain. Probability the-

ory provides the conceptual tools to represent and measure the level of uncertainty.⁴ Probability theory assigns a number—referred to as the "probability"—to every possible event. This number, the probability, might be interpreted in one of two ways. The first is that a probability is the "intensity of belief" that an event will occur, where a probability of 1 means certainty.⁵ The second interpretation is the one normally used in statistics: Probability is the percentage of times (i.e., frequency) that a particular event is observed in a large number of observations (or trials).⁶ This interpretation of probability is the frequentist interpretation, also referred to as the relative frequency concept of probability. Although it is this interpretation that is used in finance and the one adopted in this book, there are attempts to apply the subjective interpretation to financial decision making using an approach called the Bayesian approach.⁷

With this background, let's consider again the returns of some stock. Suppose that returns are unpredictable in the sense that future returns do not depend on the current information set. This does not mean that future returns are completely uncertain in the same sense in which the outcome of throwing a die is uncertain. Clearly, we cannot believe that every possible return on the stock is equally likely: There are upper and lower bounds for real returns in an economy. More important, if we collect a series of historical returns for a stock, a distribution of returns would be observed.

It is therefore reasonable to assume that our uncertainty is embodied in a probability distribution of returns. The absence of *predictability* means that the distribution of future returns does not change as a function of the current information set. More specifically, the distribution of future returns does not change as a function of the present and past values of prices and returns. This entails that the distribution of returns does not change with time. We can therefore state that (1) a price or return process is predictable if its probability distributions depend on the current information set, and (2) a price or

return process is unpredictable if its probability distributions are time-invariant.

Given the concept of predictability as we have just defined it, we can now discuss why prices and returns are difficult (or perhaps impossible) to predict. The key is that any prediction that might lead to an opportunity to generate a trading profit or an excess return tends to make that opportunity disappear. For example, suppose that the price of a stock is predicted to increase significantly in the next five trading days. A large price increase is a source of trading profit or excess return. As a consequence, if that prediction is widely shared by the investment community, investors will rush to purchase that stock. But the demand thus induced will make the stock's price rise immediately, thus eliminating the source of trading profit or excess return and invalidating the forecast.

Suppose that the predictions of stock returns were certain rather than uncertain. By a certain prediction it is meant a prediction that leaves no doubt about what will happen. For example, U.S. Treasury zero-coupon securities if held to maturity offer a known or certain prediction of returns because the maturity value is guaranteed by the full faith and credit of the U.S. government. Any forecast that leaves open the possibility that market forces will alter the forecast cannot be considered a certain forecast. If stock return predictions are certain, then simple arbitrage arguments would dictate that all stocks should have the same return. In fact, if stock returns could be predicted with certainty and if there were different returns, then investors would choose only those stocks with the highest returns.

Stock return forecasts are not certain; as we have seen, uncertain predictions are embodied in probability distributions. Suppose that we have a joint probability distribution of the returns of the universe of investable stocks. Investors will decide the rebalancing of their portfolios depending on their probabilistic predictions and their risk-return preferences. The problem we are discussing here is whether gen-

eral considerations of market efficiency are able to determine the mathematical form of price or return processes. In particular, we are interested in understanding if stock prices or returns are necessarily unpredictable.

The problem discussed in the literature is expressed roughly as follows. Suppose that returns are a series of random variables. These series will be fully characterized by the joint distributions of returns at any given time t and at any given set of different times. Suppose that investors know these distributions and that they select their portfolios according to specific rules that depend on these distributions. Can we determine the form of admissible processes, that is, of admissible distributions?

Ultimately, the objective in solving this problem is to avoid models that allow unreasonable inferences. Historically, three solutions have been proposed:

1. Returns fluctuate randomly around a given mean.
2. Returns are a fair game.
3. Returns are a fair game after adjusting for risk.

In statistical terminology, returns fluctuating randomly around a given mean refers to returns following *multivariate random walks*. A fair game means that returns are *martingales*. These concepts and their differences will be explained below. The first two proposed solutions are incorrect; the third is too general to be useful for asset management. Before we discuss the above models of prices, we digress to briefly explain some statistical concepts.

Statistical Concepts of Predictability and Unpredictability

Because we have stressed how we must rely on probability to understand the concepts of predictability and unpredictability, we will first explain the concepts of conditional probability, conditional expectation, independent and identically distributed random variables, strict

white noise, martingale difference sequence, and white noise. In addition, we have to understand the concept of an error term and an innovation.

Conditional probability and conditional expectation are fundamental in the probabilistic description of financial markets. A conditional probability of some random variable X is the probability for X given a particular value for another random variable Y is known. Similarly, a conditional probability distribution can be determined. For the conditional probability distribution, an expected value can be computed and is referred to as a conditional expected value or conditional mean or, more commonly, a conditional expectation.

The statistical concept independent and identically distributed variables (denoted by IID variables) means two conditions about probability distributions for random variables. First consider "independent." This means if we have a time series for some random variable, then at each time the random variable has a probability distribution. By independently distributed, it is meant that the probability distributions remain the same regardless of the history of past values for the random variable. "Identically" distributed means that all returns have the same distribution in every time period. These two conditions entail that, over time, the mean and the variance do not change from period to period. In the parlance of the statistician, we have a stationary time-series process.

A strict white noise is a sequence of IID variables that have a mean equal to zero and a finite variance. Hence, a strict white noise is unpredictable in the sense that the conditional probability distribution of the random variables is fixed and independent from the past. Because a strict white noise is unpredictable, expectations and higher moments are unpredictable. Moments are measures to summarize the probability distribution. The first four moments are expected value or mean (location), variance (dispersion), skewness

(asymmetry), and kurtosis (concentration in the tails). The higher moments of a probability distribution are those beyond the mean and variance, that is skewness and kurtosis.

A martingale difference sequence is a sequence of random variables that have a mean of zero that are uncorrelated such that their conditional expectations given the past values of the series is always zero. Because expectations and conditional expectations are both zero, in a martingale difference sequence, expectations are unpredictable. However, if higher moments exist, they might be predictable.

A white noise is a sequence of uncorrelated random variables with a mean of zero and a finite variance. Since the random variables are uncorrelated, in a white noise expectations are linearly unpredictable. Higher moments, if they exist, might be predictable. The key here is that they are unpredictable using a linear model. However, they may be predicted as nonlinear functions of past values. It is for this reason that certain statistical techniques that involve nonlinear functions such as neural networks have been used by some quantitative asset management firms to try to predict expectations.

Random Walks and Martingales

In the special case where the random variables are normally distributed, it can be proven that strict white noise, martingale difference sequence, and white noise coincide. In fact, two uncorrelated, normally distributed random variables are also independent.

We can now define what is meant by an arithmetic random walk, a martingale, and a strict arithmetic random walk that are used to describe the stochastic process for returns and prices as follows:

- An arithmetic random walk is the sum of white-noise terms. The mean of an arithmetic random walk is linearly unpredictable but might be predictable with nonlinear predictors. Higher moments might be predictable.

- A *martingale* is the sum of martingale difference sequence terms. The mean of a martingale is unpredictable (linearly and nonlinearly); that is, the expectation of a martingale coincides with its present value. Higher moments might be predictable.
- A strict random walk is the sum of strict white-noise terms. A strict random walk is unpredictable: Its mean, variance, and higher moments are all unpredictable.

Error Terms and Innovations

Any statistical process can be broken down into a predictable and an unpredictable component. The first component is that which can be predicted from the past values of the process. The second component is that which cannot be predicted. The component that cannot be predicted is called the innovation process. Innovation is not specifically related to a model, it is a characteristic of the process. Innovations are therefore unpredictable processes.

Now consider a model that is supposed to explain empirical data such as predicting future returns or prices. For a given observation, the difference between the value predicted by the model and the observation is called the residual. In econometrics, the residual is referred to as an error term or, simply, error of the model. It is not necessarily true that errors are innovations; that is, it is not necessarily true that errors are unpredictable. If errors are innovations, then the model offers the best possible explanation of data. If not, errors contain residual forecastability. The previous discussion is relevant because it makes a difference if errors are strict white noise, martingale difference sequences, or simply white noise.

More specifically, a random walk whose changes (referred to as increments) are non-normal white noise contains a residual structure not explained by the model both at the level of expectations and higher moments. If data follow a martingale model, then expectations are completely explained by the model but higher moments are not.

The Importance of the Statistical Concepts

We have covered a good number of complex statistical concepts. What's more, many of these statistical concepts are not discussed in basic statistics courses offered in business schools. So, why are these apparently arcane statistical considerations of practical significance to investors? The reason is that the properties of models that are used in attempting to forecast returns and prices depend on the assumptions made about "noise" in the data. For example, a linear model makes linear predictions of expectations and cannot capture nonlinear events such as the clustering of volatility that have been observed in real-world stock markets. It is therefore natural to assume that errors are white noise. In other models attempting to forecast returns and prices, however, different assumptions about noise need to be made; otherwise the properties of the model conflict with the properties of the noise term.

Now, the above considerations have important practical consequences when testing error terms to examine how well the models that will be described later in this entry perform. When testing a model, one has to make sure that the residuals have the properties that we assume they have. Thus, if we use a linear model, say a linear regression, we will have to make sure that residuals from time-series data are white noise; that is, that the residuals are uncorrelated over time. The correlation between the residuals at different times from a model based on time-series data is referred to as autocorrelation. In a linear regression using time-series data, the presence of autocorrelation violates the ordinary least squares assumption when estimating the parameters of the statistical model.⁸ In general, it will suffice to add lags to the set of predictor variables to remove the existence of autocorrelation of the residuals.⁹ However, if we have to check that residuals are martingale difference sequences or strict white noise, we will have to use more powerful tests. In addition, adding lags will not be sufficient to remove undesired properties of residuals. Models

will have to be redesigned. These effects are not marginal: They can have a significant impact on the profitability and performance of investment strategies.

A CLOSER LOOK AT PRICING MODELS

Armed with these concepts from statistics, let's now return to a discussion of pricing models. The first hypothesis on equity price processes that was advanced as a solution to the problem of forecastability was the random walk hypothesis. The strongest formulation assumes that returns are a sequence of IID variables, that is, a strict random walk. This means that, over time, the mean and the variance do not change from period to period. If returns are IID variables, it can be shown that the logarithms of prices follow a random walk and the prices themselves follow what is called a geometric random walk. The IID model is clearly a model without forecastability as the distribution of future returns does not depend on any information set known at the present moment. It does, however, allow stock prices to have a fixed drift.

There is a weaker form of the random walk hypothesis that only requires that returns at any two different times be uncorrelated. According to this weaker definition, returns are a sequence formed by a constant drift plus white noise. If returns are a white noise, however, they are not unpredictable. In fact, a white noise, although uncorrelated at every lag, might be predictable in the sense that its expectation might depend on the present information set.

At one time, it was believed that if one assumes investors make perfect forecasts, then the strict random walk model was the only possible model. However, this conclusion was later demonstrated to be incorrect by LeRoy (1973). He showed that the class of admissible models is actually much broader. That is, the strict random walk model is too restricted to be the

only possible model and proposed the use of the martingale model (i.e., the fair game model) that we explain next.

The idea of a martingale has a long history in gambling. Actually the word "martingale" originally meant a gambling strategy in which the gambler continually doubles his or her bets. In modern statistics, a martingale embodies the idea of a fair game where, at every bet, the gambler has exactly the same probability of winning or losing. In fact, as explained earlier in this entry, the martingale is a process where the expected value of the process at any future date is the actual value of the process. If a price process or a game is represented by a martingale, then the expectation of gains or losses is zero. As from our discussion, a random walk with uncorrelated increments is not necessarily a martingale as its expectations are only linearly unpredictable.

Technically, the martingale model applies to the logarithms of prices. Returns are the differences of the logarithms of prices. The martingale model requires that the expected value of returns is not predictable because it is zero or a fixed constant. However, there can be subtle patterns of forecastability for higher moments of the return distribution. Higher moments, to repeat, are those moments of a probability distribution beyond the expected value (mean) and variance, for example, skewness and kurtosis. In other words, the distribution of returns can depend on the present information set provided that the expected value of the distribution remains constant.

The martingale model does not fully take into consideration risk premiums because it allows higher moments of returns to vary while expected values remain constant. It cannot be a general solution to the problem of what processes are compatible with the assumptions that investors can make perfect probabilistic forecasts.

The definitive answer is due to Harrison and Kreps (1979) and Harrison and Pliska (1981, 1985). They demonstrated that stock prices

must indeed be martingales but after multiplication for a factor that takes into account risk. The conclusion of their work (which involves a very complicated mathematical model), however, is that a broad variety of predictable processes are compatible with the assumption that the market is populated by market agents capable of making perfect forecasts in a probabilistic sense. Predictability is due to the interplay of risk and return.

However, it is precisely due to the market being populated by market agents capable of making perfect forecasts, it is not necessarily true that successful predictions will lead to excess returns. For example, it is generally accepted that predicting volatility is easier than predicting returns. The usual explanation of this fact is that investors and portfolio managers are more interested in returns than in volatility. With the maturing of the quantitative methods employed by asset managers coupled with the increased emphasis placed on risk-return, risk and returns have become equally important. However, this does not entail that both risk and returns have become unpredictable. It is now admitted that it is possible to predict combinations of the two.

PREDICTIVE RETURN MODELS

Equity portfolio managers have used various statistical models for forecasting returns and risk. These models, referred to as *predictive return models*, make conditional forecasts of expected returns using the current information set. That information set could include past prices, company information, and financial market information such as economic growth or the level of interest rates.

Most predictive return models employed in practice are statistical models. More specifically, they use tools from the field of econometrics. We will provide a nontechnical review of econometric-based predictive return models below.

Predictive return models can be classified into four general types:¹⁰

1. *Regressive model*. This model involves the use of regression analysis where the variables used to predict returns (also referred to as predictors or explanatory variables) are the factors that are believed to impact returns.
2. *Linear autoregressive model*. In this model, the variables used to predict returns are the lagged returns (i.e., past returns).
3. *Dynamic factor model*. Models of this type use a mix of prices and returns.
4. *Hidden-variable model*. This type of model seeks to capture regime change.

Although these models use traditional econometric techniques and are the most commonly used in practice, in recent years other models based on the specialized area of machine learning have been proposed. The machine-learning approach in forecasting returns involves finding a model without any theoretical assumptions. This is done through a process of what is referred to as progressive adaptation. Machine-learning approaches, rooted in the fields of statistics and artificial intelligence (AI), include neural networks, decision trees, clustering, genetic algorithms, support vector machines, and text mining.¹¹ We will not describe machine-learning based predictive return models. However, in the 1990s, there were many exaggerated claims and hype about their potential value for forecasting stock returns that could completely revolutionize portfolio management. Consequently, they received considerable attention by the investment community and the media. It seems these claims never panned out.¹²

As a prerequisite for the adoption of a predictive return model, there are a number of key questions that a portfolio manager must address. These include:¹³

- What are the statistical properties of the model?
- How many predictor (explanatory) variables should be used in the model?

- What is the best statistical approach to estimate the model and is commercial software available for the task?
- How does one statistically test whether the model is valid?
- How can the consequences of errors in the choice of a model be mitigated?

The first and last questions rely on the statistical concepts that we described earlier. These questions are addressed in more technical-oriented equity investment management books.¹⁴ Consequently, we will limit our discussion in this entry to only the first question, describing the statistical properties of the four types of predictive return models. That is, we describe the fundamental statistical concepts behind these models and their economic meaning, but we omit the mathematical details.

Regressive Models

Regressive models of returns are generally based on linear regressions on factors. Factors are also referred to as predictors. Linear regression models are used in several aspects of portfolio management beyond that of return forecasting. For example, an equity analyst may use such models to forecast future sales of a company being analyzed.

Regressive models can be categorized as one of two fundamental kinds. The first is static regressive models. These models do not make predictions about the future but regress present returns on present factors. The second type is predictive regressive models. In such models future returns are regressed on present and past factors to make predictions. For both types of models, the statistical concepts and principles are the same. What differs is the economic meaning of each type of model.

Static Regressive Models

Static regressive models for predicting returns should be viewed as timeless relationships that are valid at any moment. They are not useful for predictive purposes because there is no time lag between the return and the factor. For example,

consider the empirical analogue of the CAPM as represented by the characteristic line given by the following regression model:

$$r_t - r_{ft} = \alpha_i + \beta_i [r_{Mt} - r_{ft}] + e_{it} \quad (1)$$

where

r_t = return on the stock in month t

r_{ft} = the risk-free rate in month t

r_{Mt} = the return on the market index (say S&P 500) in month t

e_t = the error term for the stock in month t

α and β = parameters for the stock to be estimated by the regression model

t = month ($t = 1, 2, \dots, T$)

The above model says that the conditional expectation of a stock's return at time t is proportional to the excess return of the market index at time t . This means that to predict the stock return at time $T + 1$, the portfolio manager must know the excess return of the market index at time $T + 1$, which is, of course, unknown at time $T + 1$. Predictions would be possible only if a portfolio manager could predict the excess return of the market index at time $T + 1$ (i.e., $r_{MT+1} - r_{fT+1}$).

There are also static multifactor models of return where the return at time t is based on the factor returns at time t . For example, suppose that there are N factors. Letting F_{nt} ($n = 1, 2, \dots, N$; $t = 1, 2, \dots, T$), then a regression model for a multifactor model for stock i (again dropping the subscript i for stock i) would be

$$\begin{aligned} r_t - r_{ft} = & \alpha + \beta_{F1}[r_{F1,t} - r_{ft}] \\ & + \beta_{F2}[r_{F2,t} - r_{ft}] + \dots \\ & + \beta_{FN}[r_{FN,t} - r_{ft}] + e_t \end{aligned} \quad (2)$$

where

r_t = return on the stock in month t

r_{ft} = the risk-free rate in month t

$r_{FN,t}$ = the return on factor N in month t

e_t = the error term for the stock in month t

α and β_{FN} 's = parameters for the stock to be estimated by the regression model
 $t = \text{month } (t = 1, 2, \dots, T)$

Thus, in order for a portfolio manager to build a portfolio or to compute portfolio risk measures using the above multifactor model for month $T + 1$, just as in the case of the characteristic line, some assumption about how to forecast the excess returns (i.e., $r_{FN,T+1} - r_{f,T+1}$) for each factor is required.

Predictive Regressive Models

In the search for models to predict returns, predictive regressive models have been developed. To explain predictive regressive models, consider some stock return and an assumed number of predictors. These predictors could be financial measures and market measures. A predictive linear regressive model assumes that the stock return at any given time t is a weighted average of its predictors at an earlier time plus a constant and some error. Hence, the information needed for predicting a stock's return does not require the forecasting of the predictor used in the regression model.

Predictive regressive models can also be defined by estimating a regression model where there are factors used as predictors at different lags. Such models, referred to as distributed lag models, have the advantage that they can capture the eventual dependence of returns not only on factors but also on the rate of change of factors. Here is the economic significance of such models. Suppose that a portfolio manager wants to create a predictive model based on, among other factors, "market sentiment." In practice, market sentiment is typically measured as a weighted average of analysts' forecasts. A reasonable assumption is that stock returns will be sensitive to the value of market sentiment but will be even more sensitive to changes in market sentiment. Hence, distributed lag models will be useful in this setting.

Linear Autoregressive Models

In a linear autoregressive model, a variable is regressed on its own past values. Past values are referred to as lagged values and when they are used as predictors in the model they are referred to as lagged variables. In the case of predictive return models, one of the lagged variables would be the past values of the return of the stock. If the model involves only the lagged variable of the stock return, it is called an *autoregressive model* (AR model). An AR model prescribes that the value of a variable at time t be a weighted average of the values of the same variable at times $t - 1, t - 2, \dots$, and so on (depending on number of lags) plus an error term. The weighting coefficients are the model parameters that must be estimated. If the model includes p lags, then p parameters must be estimated.

If there are other lagged variables in addition to the lagged variable representing the past values of the return on the stock included in the regression model, the model is referred to as a vector autoregressive model (VAR model). The model expresses each variable as a weighted average of its own lagged values plus the lagged values of the other variables. A VAR model with p lags is denoted by VAR(p) model. The benefit of a VAR model is that it can capture cross-autocorrelations; that is, a VAR model can model how values of a variable at time t are linked to the values of another variable at some other time. An important question is whether these links are causal or simply correlations.¹⁵

For a model to be useful, the number of parameters to be estimated needs to be small. In practice, the implementation of a VAR is complicated by the fact that such models can only deal with a small number of series. This is because when there is a large number of series—for example, the return processes for the individual stocks making up such aggregates as the S&P 500 Index—this would require a large number of parameters to be estimated. For example, if one wanted to model the daily returns

of the S&P 500 with a VAR model that included two lags, the number of parameters to estimate would be 500,000. To have at least as many data points as parameters, one would need at least four years of data, or 1,000 trading days, for each stock return process, which is $1,000 \times 500 = 500,000$ data points. Under these conditions, estimates would be extremely noisy and the estimated model would be meaningless.

Dynamic Factor Models

Unlike a VAR model, which involves regressing returns on factors but does not model the factors, a dynamic factor model assumes factors follow a VAR model and returns (or prices) are regressed on these factors. The advantage of such models is that unlike the large amount of data needed to estimate the large number of parameters in a VAR model, a dynamic factor model can significantly reduce the number of parameters to be estimated and therefore the amount of data needed.

Hidden-Variable Models

Hidden-variable models attempt to represent states of the market using hidden variables. Probably the best known hidden-variable model is the autoregressive conditional heteroscedasticity (ARCH) and generalized autoregressive conditional heteroscedasticity (GARCH) family. ARCH/GARCH models use an autoregressive process to model the volatility of another process. The result is a rich representation of the behavior of the model volatility.

Another category of hidden-variable models is the Markov switching–vector autoregressive (MS–VAR) family. These models do allow forecasting of expected returns. The simplest MS–VAR model is the Hamilton model.¹⁶ In economics, this model is based on two random walk models—one with a drift for periods of economic expansion and the other with a smaller drift for periods of economic recession. The switch between the two models is governed by a probability transition table that prescribes

the probability of switching from recession to expansion, and vice versa, and the probability of remaining in the same state.

IS FORECASTING MARKETS WORTH THE EFFORT?

In the end, all of this discussion leads to the question: What are the implications for portfolio managers and investors who are attempting or contemplating attempting building predictive return models? That is, how does this help portfolio managers and investors to decide if there is potentially sufficient benefit (i.e., trading profits and/or excess returns) in trying to extract information from market price data through quantitative modeling? There are three important points regarding this potential benefit.

The first, as stated by Fabozzi, Focardi, and Kolm (2006a, 11), is the following:

It is not true that progress in our ability to forecast will necessarily lead to a simplification in price and return processes. Even if investors were to become perfect forecasters, price and return processes might still exhibit complex patterns of forecastability in both expected values and higher moments, insofar as they might be martingales after dynamically adjusting for risk. No simple conclusion can be reached simply by assuming that investors are perfect forecasters: in fact, it is not true that the ability to forecast prices implies that prices are unpredictable random walks.

It is noteworthy that when the random walk hypothesis was first proposed in the academic community, it was the belief that the task of price forecasting efforts was a worthless exercise because prices were random walks. However, it seems reasonable to conclude that price processes will always be structured processes simply because investors are trying to forecast them. Modeling and sophisticated forecasting techniques will be needed to understand the risk-return trade-offs offered by the market.

The second point is that the idealized behavior of perfect forecasters does not have much to do with the actual behavior of real-world investors. The behavior of markets is the result,

not of perfectly rational market agents, but of the action of market agents with limited intelligence, limited resources, and subject to unpredictable exogenous events. Consequently, the action of market agents is a source of uncertainty in itself. As a result, there is no theoretical reason to maintain that the multivariate random walk is the most robust model.

Real-world investors use relatively simple forecasting techniques such as linear regressions. It is reasonable to believe that when real-world investors employ judgment, there is the possibility of making large forecasting errors. As the behavioral finance camp argues, the preoccupation with the idealized behavior of markets populated by perfect forecasters seems to be misguided. Theorists who defend the assumption that investors in the real world are perfect forecasters, believe that it is unreasonable to assume that investors make systematic mistakes. Proponents of this assumption claim that, on average, investors make correct forecasts.

However, the evidence suggests that this claim is not true. Investors can make systematic mistakes and then hit some boundary, the consequences of which can be extremely painful in terms of wealth accumulation as we saw in the late 1990s with the bursting of the technology, media, and telecommunications bubble. As Fabozzi, Focardi, and Kolm (2006a, 11) conclude:

A pragmatic attitude prevails. Markets are considered to be difficult to predict but to exhibit rather complex structures that can be (and indeed are) predicted, either qualitatively or quantitatively.

Finally, an important point is that predictability is not the only path to profitability/excess returns. Citing once again from Fabozzi, Focardi, and Kolm (2006a, 11–12):

If prices behaved as simple models such as the random walk model or the martingale, they could nevertheless exhibit high levels of persistent profitability. This is because these models are characterized by a fixed structure of expected returns. Actually, it is the time-invariance of expected returns coupled with the existence of risk premiums that makes these models unsuitable as long-term

models. . . . A model such as the geometric random walk model of prices leads to exponentially diverging expected returns. This is unrealistic in the long run, as it would lead to the concentration of all market capitalization in one asset. As a consequence, models such as the random walk model can only be approximate models over limited periods of time. This fact, in turn, calls attention to robust estimation methods. A random walk model is not an idealization that represents the final benchmark model: It is only a short-term approximation of what a model able to capture the dynamic feedbacks present in financial markets should be.

Hence, whether the random walk assumption is in fact the benchmark model of price processes must be addressed empirically. Yet, the view of portfolio managers is that markets offer patterns of predictability in returns, volatility (variance), and, possibly, higher moments. Because any such patterns might offer opportunities for realizing excess returns, a portfolio manager who ignores these patterns will be risking lost opportunities to enhance performance. As Fabozzi, Focardi, and Kolm (2006a, 24) state:

[S]imple random walk models with risk premiums are not necessarily the safest models. The joint assumptions that markets are unforecastable and that there are risk premiums is not necessarily the safest assumption.

KEY POINTS

- Despite the ongoing debate about the predictability of stock prices and returns, asset management firms have adopted statistical models of various levels of complexity for forecasting these values.
- The concept of forecastability rests on how one can forecast the future given the current information set known at that date.
- Prices or returns are said to be forecastable if the knowledge of the past influences our forecast of the future.
- The two beliefs that seem to be held in the investment community are (1) predictable processes allow investors to earn excess returns, and (2) unpredictable processes do not allow investors to earn excess returns.

- Predictable processes do not necessarily produce excess returns if they are associated with unfavorable risk, and unpredictable expectations can be profitable if the expected value is favorable.
- Probability theory is used in decision making to represent and measure the level of uncertainty.
- The absence of predictability means that the distribution of future returns does not change as a function of the present and past values of prices and returns.
- From this perspective, a price or return process is said to be predictable if its probability distributions depend on the current information set, and a price or return process is said to be unpredictable if its probability distributions do not vary over time. Using this concept of predictability, we can understand why prices and returns are difficult, perhaps even impossible, to predict.
- The key is that any prediction that might lead to an opportunity to generate a trading profit or an excess return tends to make that opportunity disappear. If stock return predictions are certain, then using simple arbitrage arguments would dictate that all stocks should have the same return. In fact, if stock returns could be predicted with certainty and if there were different returns, then investors would choose only those stocks with the highest returns.
- Because stock return forecasts are not certain, uncertain predictions are embodied in probability distributions.
- The problem faced by investors is whether general considerations of market efficiency are capable of determining the mathematical form of price or return processes. In particular, investors are interested in understanding if stock prices or returns are necessarily unpredictable.
- In solving this problem, the investor's objective is to shun models that permit unreasonable inferences. The following solutions have been proposed: (1) Returns fluctuate randomly around a given mean (i.e., returns follow multivariate random walks); (2) returns are a fair game (i.e., returns are martingales); and (3) returns are a fair game after adjusting for risk.
- Concepts from probability theory and statistics that are relevant in understanding return forecasting models are conditional probability, conditional expectation, independent and identically distributed random variables, strict white noise, martingale difference sequence, white noise, error terms, and innovations.
- An arithmetic random walk, a martingale, and a strict arithmetic random walk describe the stochastic process for returns and prices. If stock prices or returns follow an arithmetic random walk, the mean is linearly unpredictable but higher moments might be predictable.
- In the case of a martingale, the mean is unpredictable (linearly and nonlinearly), although higher moments might be predictable.
- If stock prices or returns follow a strict random walk, the mean, variance, and higher moments are all unpredictable.
- The statistical-based predictive return models used by portfolio managers make conditional forecasts of expected returns using the current information set: past prices, company information, and financial market information. These models are classified as regressive models, linear autoregressive models, dynamic factor models, and hidden-variable models.

NOTES

1. See Bernstein (2008).
2. The contributions of Bachelier are too exhaustive (and technical) to describe here. In addition to his study of the behavior of prices, his work in the area of random walks predated Albert Einstein's study of Brownian motion in physics by five years. His work in option pricing theory predated the well-known Black-Scholes option pricing model by 73 years.

3. See Samuelson (1965) and Fama (1965).
 4. See Bernstein (1998) for an account of the development of the concepts of risk and uncertainty from the beginning of civilization to modern risk management.
 5. The idea of probability as intensity of belief was introduced by Keynes (1921).
 6. The idea of probability as a relative frequency was introduced by von Mises (1921).
 7. See Rachev et al. (2007).
 8. More specifically, the presence of autocorrelation does not bias the estimated parameters of the model but results in biases in the standard errors of the estimated parameters, which are used in testing the goodness of fit of the model.
 9. Statements like this are intended as exemplifications but do not strictly embody sound econometric procedures. Adding lags has side effects, such as making estimations noisier, and cannot be used indiscriminately.
 10. Fabozzi, Focardi, and Kolm (2006a, 66).
 11. For a nontechnical discussion of these models, see Chapter 6 in Fabozzi, Focardi, and Kolm (2006a). For a more technical discussion see Fabozzi, Focardi, and Kolm (2006b).
 12. For discussion of the merits and limits of AI from a practical perspective, see Leinweber and Beinart (1996).
 13. Fabozzi, Focardi, and Kolm (2006a, 66).
 14. See, for example, Fabozzi, Focardi, and Kolm (2006b).
 15. For a discussion of the analysis of causality in VAR models, see Fabozzi, Focardi, and Kolm (2006b).
 16. Hamilton (1989).
- REFERENCES**
- Bachelier, L. (1900). Théorie de la spéculation. *Annales Scientifiques de l'École Normale Supérieure* 3, 17: 21–86.
- Bernstein, P. L. (1996). *Against the Gods: The Remarkable Story of Risk*. New York: John Wiley & Sons.
- Bernstein, P. L. (2008). Are stock prices predictable? In *Handbook of Finance*, vol. 3, edited by F. J. Fabozzi (pp. 273–380). Hoboken, NJ: John Wiley & Sons.
- Fabozzi, F. J., Focardi, S. M., and Kolm, P. N. (2006a). *Trends in Quantitative Finance*. Charlottesville, Va.: Research Foundation of the CFA Institute.
- Fabozzi, F. J., Focardi, S. M., and Kolm, P. N. (2006b). *Financial Modeling of the Equity Market: From CAPM to Cointegration*. Hoboken, NJ: John Wiley & Sons.
- Fama, E. F. (1965). The behavior of stock market prices. *Journal of Business* 38 (January): 34–105.
- Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57, 2: 357–384.
- Harrison, J. M., and Kreps, D. M. (1979). Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* 20: 381–408.
- Harrison, J. M., and Pliska, S. R. (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Process Application* 11: 215–260.
- Harrison, J. M., and Pliska, S. R. (1985). A stochastic calculus model of continuous trading: Complete markets. *Stochastic Process Application* 15: 313–316.
- Keynes, J. M. (1921). *Treatise on Probability*. London: Macmillan.
- Leinweber, D. J., and Beinart, Y. (1996). Little AI goes a long way on Wall Street. *Journal of Portfolio Management* 27, 2: 95–106.
- LeRoy, S. F. (1973). Risk aversion and the martingale property of stock prices. *International Economic Review* 14: 436–446.
- Rachev, S. T., Hsu, J., Bagasheva B., and Fabozzi, F. J. (2008). *Bayesian Methods in Finance*. Hoboken, NJ: John Wiley & Sons.
- Samuelson, P. A. (1965). Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review* 6, 2: 41–50.
- von Mises, R. (1928). *Wahrscheinlichkeitsrechnung, Statistik und Wahrheit*. Vienna: Julius Spring. (English edition translated in 1939 by J. Neyman, D. Scholl, and E. Rabinowitsch, *Probability, Statistics and Truth*. New York: Macmillan, 1939.)

Factor Models for Portfolio Construction

Factor Models

GUOFU ZHOU, PhD

Frederick Bierman and James E. Spears Professor of Finance, Olin Business School,
Washington University in St. Louis

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Asset pricing models seek to estimate the relationship between the factors that drive asset expected return. The factors that drive the expected returns are referred to as risk factors. Two well-known asset pricing models returns are the capital asset pricing model and the arbitrage pricing theory. The relationship between risk factors and expected return in these two equilibrium models is based on various assumptions. In practice, multifactor models are estimated from observed asset returns and sophisticated statistical techniques are employed to estimate the exposure of an asset to each factor.

Given a set of assets or asset classes, an important task in the practice of investment management is to understand and estimate their expected returns and the associated risks. Factor models are widely used by investors to link the risk exposures of the assets to a set of known or unknown factors. The known factors can be economic or political factors, industry factors or country factors, and the unknown factors are those that best describe the dynamics of the asset returns in the factor models, but they are not directly observable or easily interpreted by investors and have to be estimated from the data.

Applications of the mean-variance analysis and portfolio selection theories in general require the estimation of expected asset returns and their covariance matrix. Those market participants who can identify those true factors that

drive asset returns should have much better estimates of the true expected asset returns and the covariance matrix, and hence should be able to form a much better portfolio than otherwise possible. Hence, a lot of research and resources are devoted to analyzing factor models in practice by the investment community. There is an intellectual “arms race” to find the best portfolio strategies to outperform competitors.

Factor model estimation depends crucially on whether the factors are *identified* (known) and unidentified (latent), and depend on the sample size and the number of assets. In addition, factor models can be used not only for explaining asset returns, but also for predicting future returns. In this entry, we review first the factor models in the case of known and latent factors in order to provide a big picture, and then discuss the details of estimation.

ARBITRAGE PRICING THEORY

One of the fundamental problems in finance is to explain the cross-section differences in asset expected returns. Specifically, what factors can explain the observed differences? Those factors that systematically affect the differences in expected returns are therefore the risks that investors are compensated for. Hence, the term “factors” is interchangeable with the term “risk factors.”

The *arbitrage pricing theory* (APT), formulated by Ross (1976), posits that expected returns of assets are linearly related to K *systematic factors*, and the exposure to these factors is measured by factor betas; that is,

$$E[\tilde{r}_i] = r_f + \gamma_1\beta_{i1} + \cdots + \gamma_K\beta_{iK} \quad (1)$$

where β_{ik} is the beta or risk exposure on the k -th factor, and γ_k is the factor risk premium, for $k = 1, 2, \dots, K$.

Technically, the APT assumes a K -factor model for the return-generating process, that is, the asset returns are influenced by K factors in the economy via linear regression equations,

$$\tilde{r}_{it} - r_{ft} = \alpha_i + \beta_{i1}\tilde{f}_{1t} + \cdots + \beta_{iK}\tilde{f}_{Kt} + \tilde{\varepsilon}_{it} \quad (2)$$

where $\tilde{f}_1, \tilde{f}_2, \dots, \tilde{f}_K$ are the systematic factors that affect all the asset returns on the left-hand side, $i = 1, 2, \dots, N$; and $\tilde{\varepsilon}_{it}$ is the asset specific risk. Note that we have placed a tilde sign (\sim) over the random asset returns, factors, and specific risks. By so doing, we distinguish between factors (random) and their realizations (data), which are important for understanding the estimation procedure below.

Theoretically, under the assumption of no arbitrage, the asset pricing relation of the APT as given by equation (1) must be true as demonstrated by Ross. There are two important points to note. First, the return-generating process as given by equation (2) is fundamentally different from the asset pricing relation. The return-generating process is a statistical model used to measure the risk exposures of the asset returns. It does not require drawing any economic con-

clusion, nor does it say anything about what the expected returns on the assets should be. In other words, the α_i 's in the return-generating process can statistically be any numbers. Only when the no-arbitrage assumption is imposed can one claim the APT, which says that the α_i 's should be linearly related to their risk exposures (betas).

Second, the APT does not provide any specific information about what the factors are. Nor does the theory make any claims on the number of factors. It simply assumes that if the returns are driven by the factors, and if the smart investors know the betas (via learning or estimating), then an arbitrage portfolio, which requires no investment but yields a positive return, can be formed if the APT pricing relation is violated in the market. Hence, in equilibrium if there are no arbitrage opportunities, we should not observe deviations from the APT pricing relation.

TYPES OF FACTOR MODELS

In this section we describe the different types of factor models.

Known Factors

The simplest case of factor models is where the K factors are assumed known or observable, so that we have time-series data on them. In this case, the K -factor model for the return-generating process as given by equation (2) is a multiple regression for each asset and is a multivariate regression if all of the individual regressions are pooled together. For example, if one believes that the gross domestic product (GDP) is the driving force for a group of stock returns, one would have a one-factor model,

$$\tilde{r}_{it} - r_{ft} = \alpha_i + \beta_{i1}\tilde{GDP}_t + \tilde{\varepsilon}_{it}$$

The above equation corresponds to equation (1) with $K = 1$ and $f_1 = \tilde{GDP}$. In practice, one can obtain time-series data on both the asset returns and GDP, and then one can estimate regressions to obtain all the parameters, including in particular the expected returns.

Another popular one-factor model is the market model regression

$$\tilde{r}_{it} - r_{ft} = \alpha_i + \beta_{i1}(\tilde{r}_{mt} - r_{ft}) + \tilde{\varepsilon}_{it}$$

where \tilde{r}_{mt} is the return on a stock market index.

To understand the covariance matrix estimation, it will be useful to write the K -factor model in matrix form,

$$\tilde{R}_t = \alpha + \beta \tilde{f}_t + \tilde{\varepsilon}_t$$

or

$$\begin{bmatrix} \tilde{R}_{1t} \\ \vdots \\ \tilde{R}_{Nt} \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} + \begin{bmatrix} \beta_{11} & \cdots & \beta_{1K} \\ \vdots & \ddots & \vdots \\ \beta_{N1} & \cdots & \beta_{NK} \end{bmatrix} \begin{bmatrix} \tilde{f}_{1t} \\ \vdots \\ \tilde{f}_{Kt} \end{bmatrix} + \begin{bmatrix} \tilde{\varepsilon}_{1t} \\ \vdots \\ \tilde{\varepsilon}_{Nt} \end{bmatrix}$$

where

- \tilde{R}_t = an N -vector of asset excess returns
- α = an N -vector of the alphas
- β = an $N \times K$ of betas or factor loadings
- \tilde{f}_t = a K -vector of the factors
- $\tilde{\varepsilon}$ = an N -vector of the model residuals.

For example, we can write a model with $N = 3$ assets and $K = 2$ factors as

$$\begin{bmatrix} \tilde{R}_{1t} \\ \tilde{R}_{2t} \\ \tilde{R}_{3t} \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \end{bmatrix} \begin{bmatrix} \tilde{f}_{1t} \\ \tilde{f}_{2t} \end{bmatrix} + \begin{bmatrix} \tilde{\varepsilon}_{1t} \\ \tilde{\varepsilon}_{2t} \\ \tilde{\varepsilon}_{3t} \end{bmatrix}$$

Taking covariance on both sides of equation (2), we have the return covariance matrix

$$\Sigma = \beta' \Sigma_f \beta + \Sigma_\varepsilon \quad (3)$$

where Σ_f is the covariance matrix of the factors, and Σ_ε is the covariance matrix of the residuals. Σ_f can be estimated by using the sample covariance matrix from the historical returns. This works for Σ_ε too if N is small relative to T . However, when N is large relative to T , the sample covariance matrix of the residuals will be poorly behaved.

Usually an additional assumption that the residuals are uncorrelated is imposed, so that Σ_ε becomes a diagonal matrix and can then be estimated by using the sample variances of the residuals. Plugging in the estimates of all the parameters into the right-hand side of equation (3), we obtain the covariance matrix needed for applying mean-variance portfolio analysis.

In the estimation of a multifactor model, it is implicitly assumed that the number of time series observations T is far greater than K , the number of factors. Otherwise, the regressions will perform poorly. For the case in which K is close to T , some special treatments are needed. This will be addressed later in this entry.

Examples of Multifactor Models with Known Factors

Before discussing latent factors, let's briefly describe four *multifactor models* where known factors are used: (1) the Fama-French three-factor model (Fama and French, 1993), (2) the MSCI Barra fundamental factor model, (3) the Burmeister-Ibbotson-Roll-Ross (BIRR) macroeconomic factor model (Burmeister, Roll, and Ross, 1994), and (4) the Barclay Group Inc. factor model. The first three are equity factor models and the last is a bond factor model.

The widely used Fama-French three-factor model is a special case of equation (1) with $K = 3$,

$$\tilde{r}_{it} - r_{ft} = \alpha_i + \beta_{im}(\tilde{r}_{mt} - r_{ft}) + \beta_{is} \tilde{SMB}_t + \beta_{ih} \tilde{HML}_t + \tilde{\varepsilon}_{it}$$

where \tilde{r}_{mt} , as before, is the return on a stock market index, \tilde{SMB}_t and \tilde{HML}_t are two additional factors. SMB_t (small minus big) is defined as the difference between the returns on diversified portfolios of small and big stocks (where small and big are measured in terms of stock market capitalization), and HML_t (high minus low) is defined as the difference between the returns on diversified portfolios of high and low book value-to-market value (B/M) stocks.

The introduction of these factors by Fama and French is to better capture the systematic variation in average return for typical portfolios than when using a stock market index alone. These factors are supported by empirical studies and are consistent with classifying stocks in terms of growth and value.

Fundamental factor models use company and industry attributes and market data as “descriptors.” Examples are price/earnings ratios, book/price ratios, estimated earnings growth, and trading activity. The estimation of a fundamental factor model begins with an analysis of historical stock returns and descriptors about a company. In the MSCI Barra model, for example, the process of identifying the factors begins with monthly returns for hundreds of stocks that the descriptors must explain. Descriptors are not the “*r* factors” but instead they are the candidates for risk factors. The descriptors are selected in terms of their ability to explain stock returns. That is, all of the descriptors are potential risk factors but only those that appear to be important in explaining stock returns are used in constructing risk factors. Once the descriptors that are statistically significant in explaining stock returns are identified, they are grouped into “risk indexes” to capture related company attributes. For example, descriptors such as market leverage, book leverage, debt-to-equity ratio, and company’s debt rating are combined to obtain a risk index referred to as “leverage.” Thus, a risk index is a combination of descriptors that captures a particular attribute of a company. For example, in the MSCI Barra fundamental multifactor model, there are 13 risk indices and 55 industry groups. The 55 industry classifications are further classified into sectors.

In a macroeconomic factor model, the inputs to the model are historical stock returns and observable macroeconomic variables. In the BIRR macroeconomic multifactor model, the macroeconomic variables that have been pervasive in explaining excess returns and which are therefore included in the market are

- *The business cycle*: Changes in real output that are measured by percentage changes in the index of industrial production.
- *Interest rates*: Changes in investors’ expectations about future interest rates that are measured by changes in long-term government bond yields.
- *Investor confidence*: Expectations about future business conditions as measured by changes in the yield spread between high- and low-grade corporate bonds.
- *Short-term inflation*: Month-to-month jumps in commodity prices, such as gold or oil, as measured by changes in the consumer price index.
- *Inflationary expectations*: Changes in expectations of inflation as measured by changes in the short-term, risk-free nominal interest rate.

Additional variables, such as the real GDP growth and unemployment rates, are also among the macroeconomic factors used by asset managers in other macroeconomic multifactor models. Moreover, some asset managers also have identified technical variables, such as trading volume and market liquidity, as factors.

The Barclay Group Inc. (BGI) bond factor model (previously the Lehman bond factor model) uses two categories of systematic risk factors: term structure factors and non-term structure risk factors. The former include changes in the level of interest and changes in the shape of the yield curve. The non-term structure factors are sector risk, credit risk, optionality risk, and a series of risks associated with investing in mortgage-backed securities.

The search for factors is a never-ending task of asset managers. In practice, many popular investment software packages use dozens of factors. Some academic studies, such as Ludvigson and Ng (2007), use hundreds of them.

Latent Factors

While some applications use observed factors, some use entirely latent factors, that is, the view

that the factors f_t in the K -factor model,

$$\tilde{R}_t = \alpha + \beta \tilde{f}_t + \tilde{\varepsilon}_t$$

are not directly observable. An argument for the use of latent factors is that the observed factors may be measured with errors or have been already anticipated by investors. Without imposing what f_t are from our likely incorrect belief, we can statistically estimate the factors based on the factor model and data.

It is important to understand that in the field of statistics, there is statistical methodology known as “factor analysis” and the model generated is referred to as a “factor model.” Factor models as used by statisticians are statistical models that try to explain complex phenomena through a small number of basic causes or factors with the factors being latent. Factor models as used by statisticians serve two main purposes: (1) They reduce the dimensionality of models to make estimation possible, and/or (2) they find the true causes that drive data. In our discussion of multifactor models, we are using the statistical tool of factor analysis to try to determine the *latent factors* driving asset returns.

While the estimation procedures for determining the set of factors will be discussed in the next section, it will be useful to know some of the properties of the factor model here. The first property is that the factors are not uniquely defined in the model, but all sets of factors are linear combinations of each other. This is because if \tilde{f}_t is a set of factors, then, for any $K \times K$ invertible matrix A , we have

$$\tilde{R}_t = \alpha + \beta \tilde{f}_t + \tilde{\varepsilon}_t = \alpha + (\beta A^{-1})(A\tilde{f}_t) + \tilde{\varepsilon}_t \quad (4)$$

which says that if \tilde{f}_t with regression coefficients β (known as adding factor loadings in the context of factor models) explains asset returns well, so does $\tilde{f}_t^* = A\tilde{f}_t$ with loadings βA^{-1} . The linear transformation of $\tilde{f}_t, \tilde{f}_t^*$, is also known as a *rotation* of f_t .

The second property is that we can assume all the factors have zero mean (i.e., $E[\tilde{f}_t] = 0$). This

is because if $\mu_f = E[f_t]$, then the factor model can be written as

$$\tilde{R}_t = \alpha + \beta \tilde{f}_t + \tilde{\varepsilon}_t = (\alpha - \beta \mu_f) + \beta(\tilde{f}_t - \mu_f) + \tilde{\varepsilon}_t \quad (5)$$

If we rename $\alpha - \beta \mu_f$ as the new alphas, and $f_t - \mu_f$ as the new factors, then the new factors will have zero means, and the new factor model is statistically the same as the old one. Hence, without loss of generality, we will assume that the mean of the factors are zeros in our estimation in the next section.

Note that the return covariance matrix formula, equation (3) or

$$\Sigma = \beta' \Sigma_f \beta + \Sigma_\varepsilon \quad (6)$$

holds regardless of whether the factors are observable or latent. However, through factor rotation, we can make a new set of factors so as to have the identity covariance matrix. In this case with $\Sigma_f = I_K$, we say that the factor model is standardized, and the covariance equation then simply becomes

$$\Sigma = \beta' \beta + \Sigma_\varepsilon \quad (7)$$

In general, Σ_ε can have nonzero off-diagonal elements, implying that the residuals are correlated. If we assume that the residuals are uncorrelated, then Σ_ε becomes a diagonal matrix, and the factor model is known as a strict factor model. If we assume further that Σ_ε has equal diagonal elements, i.e., $\Sigma_\varepsilon = \sigma^2 I_N$ for some $\sigma > 0$ with I_N an N identity matrix, then the factor model is known as a normal factor model.

Both Types of Factors

Rather than taking the view of only observable factors or only latent factors, we can consider a more general factor model with both types of factors,

$$\tilde{R}_t = \alpha + \beta \tilde{f}_t + \beta_g \tilde{g}_t + \tilde{\varepsilon}_t \quad (8)$$

where \tilde{f}_t is a K -vector of latent factors, \tilde{g}_t is an L -vector of observable factors, and β_g are the betas associated with \tilde{g}_t . This model makes intuitive sense. If we believe a few fundamental or macroeconomic factors are the driving forces, they can be used to create the \tilde{g}_t vector. Since we may not account for all the possible factors, we need to add K unknown factors, which are to be estimated from the data.

The estimation of the above factor model given by equation (8) usually involves two steps. In the first step, a regression of the asset returns on the known factors is run in order to obtain $\hat{\beta}_g$, an estimate of β_g . This allows us to compute the residuals,

$$\hat{u}_t = R_t - \hat{\beta}_g g_t \quad (9)$$

that is, the difference of the asset returns from their fitted values by using the observed factors for all the time periods. Then, in the second step, a factor estimation approach is used to estimate the latent factors for \hat{u}_t ,

$$\tilde{u}_t = \alpha + \beta \tilde{f}_t + \beta_g \tilde{g}_t + \tilde{v}_t \quad (10)$$

where \tilde{u}_t is the random differences whose realized values are \hat{u}_t . The estimation method for this model is the same as estimating a latent factor model and will be detailed in the next section. With the factor estimates, we can treat the latent factors as known, and then use equation (8) to determine the expected asset returns and covariance matrix.

Predictive Factor Models

An important feature of factor models is that they use time t factors to explain time t returns. This is to estimate the long-run risk exposures of the assets, which are useful for both risk control and portfolio construction. On the other hand, portfolio managers are also very concerned about time-varying expected returns. In this case, they often use a predictive factor model such as the following to forecast the returns,

$$\tilde{R}_{t+1} = \alpha + \beta \tilde{f}_t + \beta_g \tilde{g}_t + \tilde{\varepsilon}_t \quad (11)$$

where as before \tilde{f}_t and \tilde{g}_t are the latent and observable factors, respectively. The single difference is that the earlier \tilde{R}_t is now replaced by \tilde{R}_{t+1} . Equation (11) uses time t factors to forecast future return \tilde{R}_{t+1} .

Computationally, the estimation of the predictive factor model is the same as for estimating the standard factor models. However, it should be emphasized that the regression R^2 , a measure of model fitting, is usually very good in the explanatory factor models. In contrast, if a predictive factor model is used to forecast the expected returns of various assets, the R^2 rarely exceeds 2%. This simply reflects the fact that assets returns are extremely difficult to predict in the real world. For example, Rapach, Strauss, Tu, and Zhou (2009) find that the R^2 are mostly less than 1% when forecasting industry returns using a variety of past economic variables and past industry returns.

FACTOR MODEL ESTIMATION

In this section, we provide first a step-by-step procedure for estimating the factor model based on the popular and implementable approach, *principal components analysis* (PCA), to which a detailed and intuitive introduction is provided in the last section of this entry. PCA is a statistical tool that is used by statisticians to determine factors with statistical learning techniques when factors are not observable. That is, given a variance-covariance matrix, a statistician can determine factors using the technique of PCA. Then, after learning the computational procedure, we provide an application to identify three factors for bond returns. Finally, we outline some alternative procedures for estimating the factor models and their extensions.

Computational Procedure

By our use of latent models, we need to consider only how to estimate the latent factors \tilde{f}_t from

the K -factor model,

$$\tilde{Y}_t = \beta \tilde{f}_t + \tilde{\varepsilon}_t \quad (12)$$

where

$$E(\tilde{f}_t) = 0, \quad E[\tilde{Y}_t] = 0$$

This version of the factor model is obtained in two steps. We de-mean first the factor f_t so that the alphas are the expected returns of the assets. Second, we de-mean again the asset returns. In other words, we let $\tilde{Y}_t = \tilde{R}_t - \alpha$.

In practice, suppose that we have return data on N risky assets over T time periods. Then the realizations of the random variable \tilde{Y}_t can be summarized by a matrix,

$$Y = \begin{pmatrix} Y_{11} & Y_{21} & \cdots & Y_{N1} \\ \vdots & \vdots & \vdots & \vdots \\ Y_{1T} & Y_{2T} & \cdots & Y_{NT} \end{pmatrix} \quad (13)$$

where each row is the N asset returns subtracting from their sample means at time t for $t = 1, 2, \dots, T$. Our task is to estimate the realizations (unobserved) on the K factors, \tilde{f}_t , over the T periods,

$$F = \begin{pmatrix} F_{11} & F_{21} & \cdots & F_{K1} \\ \vdots & \vdots & \vdots & \vdots \\ F_{1T} & F_{2T} & \cdots & F_{KT} \end{pmatrix} \quad (14)$$

We will now apply PCA estimation methodology.

There are two important cases, each of which calls for a different way of applying PCA. The first case is the one of traditional factor analysis in which N is treated as fixed, and T is allowed to grow. We will refer to this case as the “fixed N ” below. The second case is when N is allowed to grow but T is either fixed or allowed to grow. We will refer to this case simply as “large N .”

Case 1: Fixed N

In the case of fixed N , we have a relatively smaller number of assets and a relatively large sample size. Then the covariance matrix of the asset returns, which is the same as the covariance matrix of \tilde{Y}_t , can be estimated by the sam-

ple covariance matrix,

$$\Psi = \frac{Y'Y}{T} \quad (15)$$

which is an N by N matrix since Y is T by N . For example, if we think there are K (say $K = 5$) factors, we can use standard software to compute the first K eigenvectors of Ψ corresponding to the first K largest eigenvalues of Ψ , each of which is an N vector. Let $\hat{\beta}$ be the N by K matrix formed by these K eigenvectors. Then $\hat{\beta}$ will be an estimate of β . Based on this, the factors are estimated by

$$\hat{F}_t = Y_t \hat{\beta}, \quad t = 1, 2, \dots, T \quad (16)$$

where Y_t is the t -th row of Y , and \hat{F}_t is the estimate of F_t , the t -th row of F . The \hat{F}_t 's are the estimated realizations of the first K factors. Seber (1984) explains why the \hat{F}_t 's are good estimates of the true and unobserved factor realizations. However, theoretically, they, though close, will not necessarily converge to the true values, unless the factor model is normal, as T increases. Nevertheless, despite this problem, this procedure is widely used in practice.

Case 2: Large N

In the case of large N , we have a large number of assets. We now form a new matrix based on the product of Y with Y' ,

$$\Omega = \frac{YY'}{T} \quad (17)$$

which is a T by T matrix since Y is T by N . Given K , we use standard software to compute the first K eigenvectors of Ω corresponding to the first K largest eigenvalues of Ω , each of which is a T vector. Letting \hat{F} be the T by K matrix formed by these K eigenvectors, the PCA says that \hat{F} is an estimate of the true and unknown factor realizations F of equation (14), up to a linear transformation. Connor and Korajczyk (1986) provided the first study in the finance literature to apply the PCA as described above. The method is also termed “asymptotic PCA” since it allows the number of assets to increase without bound. In contrast, traditional PCA

keeps N fixed, while allowing the number of time periods, T , to be large.

Theoretically, if the true factor model is the strict factor model or is not much too different from it (i.e., the residual correlations are not too strong), Bai (2003) shows that \hat{F} converges to F up to a linear transformation when both T and N increase without limit. The estimation errors are of order the larger of $1/T$ or $1/\sqrt{N}$, and converge to zero as both T and N grow to infinity. However, when T is fixed, we need a stronger assumption that the true factor model is close to a normal model, then the estimation errors are of order of $1/\sqrt{N}$. Intuitively, at each time t , given that there are only a few factors to pricing so many assets, we should have enough information to back out the factors accurately.

Based on the estimated factors, the factor loadings are easily estimated from equation (12). For example, we can obtain the loadings for each asset by estimating the standard ordinary least squares (OLS) regression of the asset returns on the factors. Mathematically, this is equivalent to computing all the loadings from the formula

$$\hat{\beta}' = (\hat{F}'\hat{F})^{-1}\hat{F}'X \quad (18)$$

Under the same conditions above, $\hat{\beta}$ also converges to β up to a linear transformation.

The remaining question is how to determine K . In practice, this may be determined by trial and error depending on how different K 's perform in model fitting and in meeting the objectives where the model is applied. From an econometrics perspective, there is a simple solution in Case 2. Bai and Ng (2002) provide a statistical criterion

$$IC(K) = \log(V(K)) + K \left(\frac{N+T}{NT} \right) \log \left(\frac{NT}{N+T} \right) \quad (19)$$

where

$$V(K) = \sum_{i=1}^N \sum_{t=1}^T (Y_{it} - \hat{\beta}_{i1}\hat{f}_{1t} - \hat{\beta}_{i2}\hat{f}_{2t} - \dots - \hat{\beta}_{iK}\hat{f}_{Kt})^2 \quad (20)$$

For a given K , $V(K)$ is the sum of the fitted squared residual errors of the factor model across both asset and time. This is a measure of model fitting. The smaller the $V(K)$, the better the K -factor model in explaining the asset returns. So we want to choose such a K that minimizes $V(K)$. However, the more the factors, the smaller the $V(K)$, but at a cost of estimating more factors with greater estimation errors. Hence, we want to penalize too many factors. This is the same as the case in linear regressions where we also want to penalize too many regressors. The second term in equation (19) plays this role. It is an increasing function of K . Therefore, the trade-off between model fitting and estimation errors requires us to minimize the $IC(K)$ function. Theoretically, assuming that the factor model is indeed true for some fixed K^* , Bai and Ng show that the K that minimizes $IC(K)$ will converge to K^* as either N or T or both increase to infinity.

An Application to Bond Returns

To illustrate the procedure, consider an application of the PCA factor analysis to the excess returns on Treasury bonds with maturities 12, 18, 24, 30, 36, 42, 48, 54, 60, 120, and beyond 120 months. Hence, there are $N = 11$ assets. With monthly data from January 1980 to December 2008, available from the Center for Research in Security Prices of the University of Chicago's Graduate School of Business, we have a sample size of $T = 348$. Since N is small relative to T , this is a case of the fixed N .

Now

$$\Psi = \frac{Y'Y}{348}$$

is an 11 by 11 matrix. We can easily compute its eigenvalues and eigenvectors. The largest three eigenvalues are

$$(\lambda_1, \lambda_2, \lambda_3) = 10^{-2}(0.2403, 0.133, 0.012)$$

whose sum is more than 99% of the sum of all the eigenvalues. Thus, it is enough to consider $K = 3$ factors and use the first three

Table 1 Factor Loadings and Explanatory Power

	β_1	β_2	β_3	R^2 (F_1)	R^2 (F_1 and F_2)	R^2 (all three)
12 month	0.0671	-0.1418	0.4046	0.67	0.80	0.96
18 month	0.1118	-0.2057	0.4227	0.79	0.84	0.99
24 month	0.1524	-0.2455	0.3371	0.85	0.87	1.00
32 month	0.1932	-0.2876	0.3199	0.88	0.89	1.00
38 month	0.2269	-0.2851	0.2101	0.91	0.92	1.00
42 month	0.2523	-0.2621	-0.0813	0.94	0.94	0.99
48 month	0.2837	-0.2415	-0.2531	0.95	0.96	1.00
54 month	0.3072	-0.1920	-0.3762	0.97	0.97	1.00
60 month	0.3368	-0.1819	-0.3246	0.97	0.98	0.99
120 month	0.4038	0.0426	-0.1507	0.99	0.99	0.99
Over 120	0.5966	0.7173	0.2394	0.92	0.93	1.00

eigenvectors, PCAs, as proxies for the factors. Denote them as F_1 , F_2 and F_3 .

Consider now the regression of the 11 excess bond returns on the three factors,

$$R_{it} = \alpha_i + \beta_{i1}F_{1t} + \beta_{i2}F_{2t} + \beta_{i3}F_{3t} + \varepsilon_{it}$$

where $i = 1, 2, \dots, 11$. The regression R^2 s of using all the factors for each of the assets are reported in the last column of Table 1. All but one is 99% or above, confirming the eigenvalue analysis that three factors are sufficient, which explains almost all the variations of the bond returns. However, when only the first two are used, the R^2 s are smaller, but the minimum is still over 80%. When only the first factor is used, the R^2 s range from 67% on the first bond return to 99% on the 10th. Overall, the PCA factors are effective in explaining the asset returns.

The factor loadings or regression coefficients on the factors are also reported in Table 1. It is interesting that the loadings on the first factor are all positive. This implies that a positive realization of F_1 will have a positive effect on the returns of all the bonds. It is, however, clear that F_1 affects long-term bonds more than short-term bonds. As an approximation, F_1 is usually interpreted as a *level effect* or *parallel effect* that roughly shifts the returns on bonds across maturity.

The second factor, however, has a different pattern from the first. A positive realization

of F_1 will have a negative effect on short-term bonds and a positive effect on the long-term ones. This is equivalent to an increase in the slope of the bond returns across maturity (known as yield curve). Therefore, F_2 is commonly identified as a steepness factor.

Finally, a positive realization of F_3 will have a positive effect on both short- and long-term bonds, but a negative effect on the intermediate ones. Hence F_3 is usually interpreted as a curvature factor. Litterman and Scheinkman (1991) appears to have been one of the first to to apply the PCA to study bond returns and to have identified the above three factors. Although the data we used here are different, the three factors we computed share the same properties as those identified by them.

Alternative Approaches and Extensions

The standard statistical approach for estimating the factor model is the maximum likelihood (ML) method. Consider the factor model given by equation (12) where $E(\tilde{f}_t) = 0$, $E[\tilde{Y}_t] = 0$. The de-meaned returns and standardized factors are usually assumed to have normal distributions.

In addition, the factors are usually standardized so that $\Sigma_f = I_K$, and the residuals are assumed uncorrelated so that Σ_ε is diagonal. Then the log likelihood function, as the log density

function of the returns, is

$$\log L(\beta, \Sigma_\varepsilon) = -\frac{NT}{2} \log(2\pi) - \frac{T}{2} \log |\beta' \beta + \Sigma_\varepsilon| - \frac{1}{2} \sum_{t=1}^T Y_t (\beta' \beta + \Sigma_\varepsilon)^{-1} Y_t \quad (21)$$

The ML estimator of the parameters β and Σ_ε are those values that maximize the log likelihood function. Since β enters into the function in a complex nonlinear way, an analytical solution to the maximization problem is a very difficult problem. Numerically, it is still difficult if maximizing $\log L(\beta, \Sigma_\varepsilon)$ directly.

There is, however, a data-augmentation technique known as the *expectation maximization* (EM) *algorithm* that can be applied (see Lehmann and Modest, 1998). The EM algorithm can be effective in numerically solving the earlier maximization problem. The idea of the EM algorithm is simple. The key difficulty here is that the factors are unobserved. But conditional on the parameters and the factor model, we can learn them. Consider now that given the factors \tilde{f}_t , the log likelihood function conditional on f_t is

$$\log L_c(\beta, \Sigma_\varepsilon) = -\frac{NT}{2} \log(2\pi) - \frac{T}{2} \log |\Sigma_\varepsilon| - \frac{1}{2} \sum_{t=1}^T (Y_t - \beta f_t)' \Sigma_\varepsilon^{-1} (Y_t - \beta f_t) \quad (22)$$

Because it is conditional on f_t , the factor model is the usual linear regression. In other words, integrating out f_t from equation (22) yields the unconditional $\log L(\beta, \Sigma_\varepsilon)$. The beta estimates conditional on f_t are straightforward. They are the usual OLS regression coefficients, and the estimates for Σ_ε are the residual variances.

On the other hand, conditional on the parameters, we can learn the factors by using their conditional expected values obtained easily from their joint distribution with the returns. Hence, we can have an iterative algorithm. Starting from an initial guess of the factors, we maximize

the conditional likelihood function to obtain the OLS β and Σ_ε estimates, which is the M-step of the EM algorithm. Based on these estimates, we update a new estimate of f_t using their expected value. This is the EM algorithm's E-step. Using the new f_t , we learn new estimates of β and Σ_ε in the M-step. With the new estimates, we can again update the f_t . Iterating between the EM steps, the limits converge to the unconditional ML estimate and the factor estimates converge to the true ones.

As an alternative to the ML method, Geweke and Zhou (1996) propose a Bayesian approach, which treats all parameters as random variables. It works in a way similar to the EM algorithm. Conditional on parameters, we learn the factors, and conditional on the factors, we learn the parameters. Iterating after a few thousand times, we learn the entire joint distribution of the factors and parameters, which are all we need in a factor model. The advantage of the Bayesian approach is that it can incorporate prior information and can provide exact inference. In contrast, the ML method cannot use any priors, nor can it obtain the exact standard errors of both parameters and functions of interest due to the complexity of the factor model. Nardari and Scuggs (2007) extend the Bayesian approach to allow a more general model in which the covariance matrix can vary over time and the APT restrictions can be imposed.

Finally, we provide two important extensions of the factor model that are useful in practice. Note that the factors we discussed thus far assume identical and independently distributed returns and factors. These are known as *static factor models*. The first extension is *dynamic factor models*, which allow the factors to evolve over time according to a vector autoregression,

$$\tilde{f}_t = A_1 \tilde{f}_{t-1} + A_2 \tilde{f}_{t-2} + \cdots + A_m \tilde{f}_{t-m} + \tilde{v}_t \quad (23)$$

where the A 's are the regression coefficient matrices, m is the order of the autoregression that

determines how far past factor realizations still affect today's realizations, and v_t is the residual. In practice, many economic variables are highly persistent, and hence it will be important to incorporate this as above. (See Amengual and Watson [2007] for a discussion of estimation for dynamic factor models.)

The second extension is to allow the case with a large number of factors. Consider our earlier factor model

$$\tilde{R}_t = \alpha + \beta \tilde{f}_t + \beta_g \tilde{g}_t + \tilde{\varepsilon}_t \quad (24)$$

where \tilde{f}_t is a K vector of latent factors, \tilde{g}_t is an L vector of observable factors. The problem now is that L is large, about 100 or 200, for instance. This requires at least a few hundred or more time series observations for the regression of R_t on g_t to be well behaved, and this can cause a problem due to the lack of long-term time series data or due to concerns of stationarity. The idea is to break \tilde{g}_t into two sets, \tilde{g}_{1t} and \tilde{g}_{2t} , with the first having a few key variables and the second having the rest. We then consider the modified model

$$\tilde{R}_t = \alpha + \beta \tilde{f}_t + \beta_{g1} \tilde{g}_{1t} + \beta_{h1} \tilde{h}_t + \tilde{\varepsilon}_t \quad (25)$$

where $\tilde{h}(t)$ has a few variables too that represent a few major driving forces that summarize the potentially hundreds of variables of \tilde{g}_{2t} via another factor model,

$$\tilde{g}_{2t} = B \tilde{h}_t + \tilde{u}_t \quad (26)$$

where \tilde{u}_t is the residual. This second factor model provides a large dimension reduction that transforms the hundreds of variables into a few, which can be estimated by the PCA. In the end, we have only a few factors in equation (25), making the analysis feasible based on the methods we discussed earlier. Ludvigson and Ng (2007) appear to be the first to apply such a model in finance. They find that the model can effectively incorporate a few hundred variables so as to make a significant difference in understanding stock market predictability.

USE OF PRINCIPAL COMPONENTS ANALYSIS

Principal components analysis (PCA) is a widely used tool in finance. It is useful not only for estimating factor models as explained in this entry, but also for extracting a few driving variables in general out of many for the covariance matrix of asset returns. Hence, it is important to understand the statistical intuition behind it. To this end, we provide a simple introduction to it in the last section of the entry.

Perhaps the best way to understand the PCA is to go through an example in detail. Suppose there are two risky assets, whose returns are denoted by \tilde{r}_1 and \tilde{r}_2 , with covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 2.05 & 1.95 \\ 1.95 & 2.05 \end{bmatrix}$$

That is, we assume that they have the same variances of 2.05 and covariance of 1.95. Our objective is to find a linear combination of the two assets so that it has a large component in the covariance matrix, which will be clear below. For notation brevity, we assume first that the expected returns are zeros; that is,

$$E[\tilde{r}_1] = 0, \quad E[\tilde{r}_2] = 0$$

and will relax this assumption later.

Recall from linear algebra that we call any vector $(a_1, a_2)'$ satisfying

$$\Sigma \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

an eigenvector of Σ , and the associated λ the eigenvalue. In our example here, it is easy to verify that

$$\begin{bmatrix} 2.05 & 1.95 \\ 1.95 & 2.05 \end{bmatrix} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 4 \times \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and

$$\begin{bmatrix} 2.05 & 1.95 \\ 1.95 & 2.05 \end{bmatrix} \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 0.1 \times \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

so 4 and 0.1 are the eigenvalues, and $(1, 1)'$ and $(1, -1)'$ are the eigenvectors.

In practice, computer software is available to compute the eigenvalue and eigenvectors of any covariance matrix. The mathematical result is that for a covariance matrix of N assets, there are exactly N different eigenvectors and N associated positive eigenvalues (these eigenvalues can be equal in some cases). Moreover, the eigenvectors are orthogonal to each other; that is, their inner product or vector product is zero. In our example, it is clear that

$$(1, 1)' \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 1 - 1 = 0$$

It should be noted that the eigenvalue associated with each eigenvector is unique, but any scale of the eigenvector remains an eigenvector. In our example, it is obvious that a double of the first eigenvector, $(2, 2)'$, is also an eigenvector. However, the eigenvectors will be unique if we standardize them, making the sum of the elements 1. In our example,

$$A_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

are the standardized eigenvectors, which are obtained by scaling the earlier eigenvectors by $1/\sqrt{2}$. These are indeed standardized, since

$$A_1' A_1 = (1/\sqrt{2})^2 + (1/\sqrt{2})^2 = 1$$

$$A_2' A_2 = (1/\sqrt{2})^2 + (-1/\sqrt{2})^2 = 1$$

Now let us consider two linear combinations (or portfolios without imposing the weights summing to 1) of the two assets whose returns are \tilde{r}_1 and \tilde{r}_2 ,

$$\tilde{P}_1 = \frac{1}{\sqrt{2}}\tilde{r}_1 + \frac{1}{\sqrt{2}}\tilde{r}_2 = A_1' \tilde{R}$$

$$\tilde{P}_2 = \frac{1}{\sqrt{2}}\tilde{r}_1 - \frac{1}{\sqrt{2}}\tilde{r}_2 = A_2' \tilde{R}$$

where $\tilde{R} = (\tilde{r}_1, \tilde{r}_2)'$. Both \tilde{P}_1 and \tilde{P}_2 are called the *principal components* (PCs). There are three important and interesting mathematical facts about the PCs.

- *Fact 1.* The variances of the PCs are exactly equal to the eigenvalues corresponding to the eigenvectors used to form the PCs.

That is,

$$\text{Var}(\tilde{P}_1) = 4$$

$$\text{Var}(\tilde{P}_2) = 1$$

Note that the two PCs are random variables since they are the linear combination of random returns. So, their variances are well defined. The equalities to the eigenvalues can be verified directly.

- *Fact 2.* The returns can also be written as linear combinations of the PCs.

The PCs are defined as linear combinations of the returns. Inverting them, the returns are linear functions of the PCs, too. Mathematically, $\tilde{P} = A\tilde{R}$, and so $\tilde{R} = A^{-1}\tilde{P}$. Since A is orthogonal, $A^{-1} = A'$, thus $\tilde{R} = A'\tilde{P}$. That is, we have

$$\begin{aligned} \tilde{r}_1 &= \frac{1}{\sqrt{2}}\tilde{P}_1 + \frac{1}{\sqrt{2}}\tilde{P}_2 \\ \tilde{r}_2 &= \frac{1}{\sqrt{2}}\tilde{P}_1 - \frac{1}{\sqrt{2}}\tilde{P}_2 \end{aligned} \quad (26)$$

- *Fact 3.* The asset return covariance matrix can be decomposed as the sum of the products of eigenvalues with the cross products of eigenvectors.

Mathematically, it is known that

$$\begin{aligned} \Sigma &= [A_1, A_2] \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} [A_1, A_2]' \\ &= \lambda_1 A_1 A_1' + \lambda_2 A_2 A_2' = 4A_1 A_1' + 0.1A_2 A_2' \end{aligned}$$

which is also easy to verify in our example. The economic interpretation is that the total risk profile of the two assets, as captured by their covariance matrix, is a sum of two components. The first component is determined by the first PC, and the second is determined by the second PC. In other words, in the return linear combinations, equation (26), if we ignore P_2 , we will get only $\lambda_1 A_1 A_1'$, the first component in the covariance matrix decomposition, and only the

second if we ignore P_1 . We obtain the entire Σ if we ignore neither.

The purpose of the PCA is finally clear. Since 4 is 40 times as big as 0.1, the second component in the Σ decomposition has little impact, and hence may be ignored. Then, ignoring \tilde{P}_2 , we can write the returns simply as, based on equation (26),

$$\begin{aligned}\tilde{r}_1 &\approx (1/\sqrt{2})\tilde{P}_1 \\ \tilde{r}_2 &\approx (1/\sqrt{2})\tilde{P}_1\end{aligned}$$

This says that we can reduce the analysis of \tilde{r}_1 and \tilde{r}_2 by analyzing simple functions of \tilde{P}_1 . In this example, the result tells us that the two assets are almost the same. In practice, there may be hundreds of assets. By using PCA, we can reduce the dimensionality of the problem substantially to an analysis of perhaps a few, say five, PCs.

In general, when there are N assets with return $\tilde{R} = (\tilde{r}_1, \dots, \tilde{r}_N)'$, computer software can be used to obtain the N eigenvalues and N standardized eigenvectors. Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N \geq 0$ be the N eigenvalues in decreasing order, and $A_i = (a_{i1}, a_{i2}, \dots, a_{iN})'$ be the standardized eigenvector associated with λ_i , and A be an $N \times N$ matrix formed by the all the eigenvectors. Then, the i -th PC is defined as $\tilde{P}_i = A_i' \tilde{R}$, all of which can be computed in matrix form,

$$\tilde{P} = \begin{bmatrix} \tilde{P}_1 \\ \tilde{P}_2 \\ \vdots \\ \tilde{P}_N \end{bmatrix} = \begin{bmatrix} A_1' \tilde{R} \\ A_2' \tilde{R} \\ \vdots \\ A_N' \tilde{R} \end{bmatrix} = A' \tilde{R} \quad (27)$$

The decomposition for Σ is

$$\begin{aligned}\Sigma &= [A_1, \dots, A_N] \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_N \end{bmatrix} [A_1, \dots, A_N]' \\ &= \lambda_1 A_1 A_1' + \lambda_2 A_2 A_2' + \dots + \lambda_N A_N A_N'\end{aligned}$$

It is usually the case that, for some K , the first K eigenvalues are large, and the rest are too small and can then be ignored. In such situations, based on the first K PCs, we can approxi-

mate the asset returns by

$$\begin{aligned}\tilde{r}_1 &\approx a_{11}\tilde{P}_1 + a_{12}\tilde{P}_2 + \dots + a_{1K}\tilde{P}_K, \\ \tilde{r}_2 &\approx a_{21}\tilde{P}_1 + a_{22}\tilde{P}_2 + \dots + a_{2K}\tilde{P}_K, \\ &\vdots \\ \tilde{r}_N &\approx a_{N1}\tilde{P}_1 + a_{N2}\tilde{P}_2 + \dots + a_{NK}\tilde{P}_K\end{aligned} \quad (28)$$

In most studies, the K PCs may be interpreted as K factors that (approximately) derive the movements of all the N returns. Our earlier example is a case with $K = 1$ and $N = 2$.

In the above PCA discussion, the expected returns of the asset are assumed to be zero. If they are nonzero and given by a vector $(\mu_1, \mu_2, \dots, \mu_N)'$, Σ will remain the same, and so will the eigenvalues and eigenvectors. However, in this case we need to replace all the \tilde{r}_i 's in equation (27) by $\tilde{r}_i - \mu_i$'s and add μ_i 's on the right-hand side of equation (28). The interpretation will be, of course, the same as before.

In Case 1 of the factor model estimation (i.e., known or observable factors) discussed in the entry, the K PCs clearly provide a good approximation of the first K factors since they explain the asset variations the most given K . Moreover, in either Case 1 or Case 2 (latent factors), the PCA is equivalent to minimizing the model errors, as given by equation (20), by choosing both the loadings and factors, and hence the solution should be close to the true factors and loadings.

KEY POINTS

- The arbitrage pricing theory is a general multifactor model for pricing assets. The theory does not provide any specific information about what the factors are. Moreover, the APT does not make any claims on the number of factors either.
- The APT asserts that only taking the systematic risks is rewarded.
- The APT simply assumes that if the returns are driven by the factors, and if investors know the betas for the factors, then an arbitrage portfolio, which requires no investment

but yields a positive return, can be formed if the APT pricing relation is violated in the market. In equilibrium, therefore, if there are no arbitrage opportunities, deviations from the APT pricing relation should not be observed.

- In practice, factor models are widely used as a tool for estimating expected asset returns and their covariance matrix. The reason is that if investors can identify the factors that drive asset returns, they will have much better estimates of the true expected asset returns and the covariance matrix, and hence will be able to form a much better portfolio than otherwise possible.
- Factor model estimation depends crucially on (1) whether the factors are identified (known) and unidentified (latent), and (2) the sample size and the number of assets. Furthermore, factor models can be used not only for explaining asset returns, but also for predicting future returns.
- The simplest case of factor models is where the factors are assumed to be known or observable, so that time-series data are those factors can be used to estimate the model.
- In practice there are three commonly used equity multifactor models where known factors are used: (1) the Fama-French three-factor model, (2) the MSCI Barra fundamental factor model, and (3) the Burmeister-Ibbotson-Roll-Ross macroeconomic factor model. Fundamental factor models use company and industry attributes and market data as descriptors. In a macroeconomic factor model, the inputs to the model are historical stock returns and observable macroeconomic variables.
- An argument for the use of latent factors is that the observed factors may be measured with errors or have been already anticipated by investors. Without imposing what the factors are from likely incorrect beliefs, asset managers can statistically estimate the factors based on the factor model and data.
- Two important extensions of the static factor model used in practice are (1) dynamic fac-

tor models, which allow the factors to evolve over time according to a vector autoregression, and (2) allowance for a large number of factors. This second factor model provides a large dimension reduction that transforms the hundreds of variables into a few, which can be estimated by principal components analysis.

- Principal components analysis is a simple statistical approach that can be applied to estimate a factor model easily and effectively.

REFERENCES

- Amengual, D., and Watson, M. (2007). Consistent estimation of the number of dynamic factors in a large N and T panel. *Journal of Business and Economic Statistics* 25: 91–96.
- Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* 71: 135–172.
- Bai, J., and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* 70: 191–221.
- Burmeister, E., Roll, R., and Ross, S. A. (1994). A practitioner's guide to arbitrage pricing theory. In *A Practitioner's Guide to Factor Models* (pp.). Charlottesville, VA, Institute of Chartered Financial Analysts.
- Connor, G., and Korajczyk, R. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics* 15: 373–394.
- Fama, E. F., and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33, 1: 3–56.
- Geweke, J., and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies* 9: 557–587.
- Lehmann, B. N., and Modest, D. M. (1998). The empirical foundations of the arbitrage pricing theory. *Journal of Financial Economics* 21: 213–254.
- Litterman, R., and Scheinkman, J. (1991). Common factors affecting bond returns. *Journal of Fixed Income* 1: 54–61.
- Ludvigson, S. C., and Ng, S. (2007). The empirical risk-return relation: A factor analysis approach. *Journal of Financial Economics* 83: 171–222.

- Nardari, F., and Scruggs, J. T. (2007). Bayesian analysis of linear factor models with latent factors, multivariate stochastic volatility, and APT pricing restrictions. *Journal of Financial and Quantitative Analysis* 42: 857–892.
- Rapach, D. E., Strauss, J. K., Tu, J., and Zhou, G. (2009). Industry return predictability: Is it there out of sample? Working paper, Washington University, St. Louis.
- Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13: 341–360.
- Seber, G. A. F. (1984). *Multivariate Observations*, Wiley.

Principal Components Analysis and Factor Analysis

SERGIO M. FOCARDI, PhD

Partner, The Intertek Group

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: In investment management, multifactor risk modeling is the most common application of financial modeling. Multifactor risk models, or simply factor models, are linear regressions over a number of variables called factors. Factors can be exogenous variables or abstract variables formed by portfolios. Exogenous factors (or known factors) can be identified from traditional fundamental analysis or economic theory from macroeconomic factors. Abstract factors, also called unidentified or latent factors, can be determined with factor analysis or principal component analysis. Principal component analysis identifies the largest eigenvalues of the variance-covariance matrix or the correlation matrix. The largest eigenvalues correspond to eigenvectors that identify the entire market and sectors that correspond to industry classification. Factor analysis can be used to identify the structure of the latent factors.

Principal component analysis (PCA) and factor analysis are statistical tools that allow a modeler to (1) reduce the number of variables in a model (i.e., to reduce the dimensionality), and (2) identify if there is structure in the relationships between variables (i.e., to classify variables). In this entry, we explain PCA and factor analysis. We illustrate and compare both techniques using a sample of stocks. Because of its use in the estimation of factor models, we begin with a brief discussion of factor models.

FACTOR MODELS

Factor models are statistical models that try to explain complex phenomena through a small number of basic causes or factors. Factor models serve two main purposes: (1) They reduce the dimensionality of models to make estimation possible; and/or (2) they find the true causes that drive data. Factor models were introduced by Charles Spearman (1904), a leading psychologist who developed many concepts of modern psychometrics.

Spearman was particularly interested in understanding how to measure human intellectual abilities. In his endeavor to do so, he developed the first factor model, known as the Spearman model, a model that explains intellectual abilities through one common factor, the famous “general intelligence” g factor, plus another factor s , which is specific to each distinct ability. Spearman was persuaded that the factor g had an overwhelming importance. That is, he thought that any mental ability can be explained quantitatively through a common intelligence factor. According to this theory, outstanding achievements of, say, a painter, a novelist, and a scientist can all be ascribed to a common general intelligence factor plus a small contribution from specific factors.

Some 30 years later, Louis Leon Thurstone (1938) developed the first true multifactor model of intelligence. Thurstone was among the first to propose and demonstrate that there are numerous ways in which a person can be intelligent. Thurstone’s multiple-factors theory identified seven primary mental abilities.

One might question whether factors are only statistical artifacts or if they actually correspond to any reality. In the modern operational interpretation of science, a classification or a factor is “real” if we can make useful predictions using that classification. For example, if the Spearman theory is correct, we can predict that a highly intelligent person can obtain outstanding results in any field. Thus, a novelist could have obtained outstanding results in science. However, if many distinct mental factors are needed, people might be able to achieve great results in some field but be unable to excel in others.

In the early applications of factor models to psychometrics, the statistical model was essentially a conditional multivariate distribution. The raw data were large samples of psychometric tests. The objective was to explain these tests as probability distributions conditional on the value of one or more factors. In this way, one can make predictions of, for example, the

future success of young individuals in different activities.

In finance, factor models are typically applied to time series. The objective is to explain the behavior of a large number of stochastic processes, typically price, returns, or rate processes, in terms of a small number of factors. These factors are themselves stochastic processes. In order to simplify both modeling and estimation, most factor models employed in financial econometrics are static models. This means that time series are assumed to be sequences of temporally independent and identically distributed (IID) random variables so that the series can be thought of as independent samples extracted from one common distribution.

In financial econometrics, factor models are needed not only to explain data but to make estimation feasible. Given the large number of stocks presently available—in excess of 15,000—the estimation of correlations cannot be performed without simplifications. Widely used ensembles such as the S&P 500 or the MSCI Europe include hundreds of stocks and therefore hundreds of thousands of individual correlations. Available samples are insufficient to estimate this large number of correlations. Hence factor models are able to explain all pairwise correlations in terms of a much smaller number of correlations between factors.

Linear Factor Models Equations

Linear factor models are regression models of the following type:

$$X_i = \alpha_i + \sum_{j=1}^K \beta_{ij} f_j + \varepsilon_i$$

where

X_i = a set of N random variables

f_j = a set of K common factors

ε_i = the noise terms associated with each variable X_i

The β_{ij} 's are called the *factor loadings* or factor sensitivities; they express the influence of the j -th factor on the i -th variable.

In this formulation, factor models are essentially *static models*, where the variables and the factors are random variables without any explicit dependence on time. It is possible to add a *dynamic* to both the variables and the factors, but that is beyond the scope of our basic introduction in this entry.

As mentioned above, one of the key objectives of factor models is to reduce the dimensionality of the covariance matrix so that the covariances between the variables X_i are determined only by the covariances between factors. Suppose that the noise terms are mutually uncorrelated, so that

$$E(\varepsilon_i \varepsilon_j) = \begin{cases} 0, & i \neq j \\ \sigma_i^2, & i = j \end{cases}$$

and that the noise terms are uncorrelated with the factors, that is, $E(\varepsilon_i f_j) = 0, \forall i, j$. Suppose also that both factors and noise terms have a zero mean, so that $E(X_i) = \alpha_i$. Factor models that respect the above constraints are called strict factor models.

Let's compute the covariances of a strict factor model:

$$\begin{aligned} & E((X_i - \alpha_i)(X_j - \alpha_j)) \\ &= E\left(\left(\sum_{s=1}^K \beta_{is} f_s + \varepsilon_i\right)\left(\sum_{t=1}^K \beta_{jt} f_t + \varepsilon_j\right)\right) \\ &= E\left(\left(\sum_{s=1}^K \beta_{is} f_s\right)\left(\sum_{t=1}^K \beta_{jt} f_t\right)\right) + E\left(\left(\sum_{s=1}^K \beta_{is} f_s\right)(\varepsilon_j)\right) \\ &\quad + E\left((\varepsilon_i) \sum_{t=1}^K \beta_{jt} f_t\right) + E(\varepsilon_i \varepsilon_j) \\ &= \sum_{s,t} \beta_{is} E(f_s f_t) \beta_{jt} + E(\varepsilon_i \varepsilon_j) \end{aligned}$$

From this expression we can see that the variances and covariances between the variables X_i depend only on the covariances between the factors and the variances of the noise term.

We can express the above compactly in matrix form. Let's write a factor model in matrix form

as follows:

$$\mathbf{X} = \boldsymbol{\alpha} + \boldsymbol{\beta} \mathbf{f} + \boldsymbol{\varepsilon}$$

where

$\mathbf{X} = (X_1, \dots, X_N)'$ = the N -vector of variables

$\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_N)'$ = the N -vector of means

$\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_N)'$ = the N -vector of idiosyncratic noise terms

$\mathbf{f} = (f_1, \dots, f_K)'$ = the K -vector of factors

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{11} & \cdots & \beta_{1K} \\ \vdots & \ddots & \vdots \\ \beta_{N1} & \cdots & \beta_{NK} \end{bmatrix}$$

= the $N \times K$ matrix of factor loadings.

Let's define the following:

$\boldsymbol{\Sigma}$ = the $N \times N$ variance-covariance matrix of the variables \mathbf{X}

$\boldsymbol{\Omega}$ = the $K \times K$ variance-covariance matrix of the factors

$\boldsymbol{\Psi}$ = $N \times N$ variance-covariance matrix of the error terms $\boldsymbol{\varepsilon}$

If we assume that our model is a strict factor model, the matrix $\boldsymbol{\Psi}$ will be a diagonal matrix with the noise variances on the diagonal, that is,

$$\boldsymbol{\Psi} = \begin{pmatrix} \psi_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \psi_N^2 \end{pmatrix}$$

Under the above assumptions, we can express the variance-covariance matrix of the variables in the following way:

$$\boldsymbol{\Sigma} = \boldsymbol{\beta} \boldsymbol{\Omega} \boldsymbol{\beta}' + \boldsymbol{\Psi}$$

In practice, the assumption of a strict factor model might be too restrictive. In applied work, factor models will often be approximate factor models. (See, for example, Bai, 2003.) Approximate factor models allow idiosyncratic terms

to be weakly correlated among themselves and with the factors.

As many different factor models have been proposed for explaining stock returns, an important question is whether a factor model is fully determined by the observed time series. In a strict factor model, factors are determined up to a nonsingular linear transformation. In fact, the above matrix notation makes it clear that the factors, which are hidden, nonobservable variables, are not fully determined by the above factor model. That is, an estimation procedure cannot univocally determine the hidden factors and the factor loadings from the observable variables \mathbf{X} . In fact, suppose that we multiply the factors by any nonsingular matrix \mathbf{R} . We obtain other factors

$$\mathbf{g} = \mathbf{R}\mathbf{f}$$

with a covariance matrix

$$\mathbf{\Omega}_g = \mathbf{R}\mathbf{\Omega}\mathbf{R}^{-1}$$

and we can write a new factor model:

$$\begin{aligned}\mathbf{X} &= \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{f} + \boldsymbol{\varepsilon} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{R}^{-1}\mathbf{R}\mathbf{f} \\ &+ \boldsymbol{\varepsilon} = \boldsymbol{\alpha} + \boldsymbol{\beta}_g\mathbf{g} + \boldsymbol{\varepsilon}\end{aligned}$$

In order to solve this indeterminacy, we can always choose the matrix \mathbf{R} so that the factors \mathbf{g} are a set of orthonormal variables, that is, uncorrelated variables (the orthogonality condition) with unit variance (the normality condition). In order to make the model uniquely identifiable, we can stipulate that factors must be a set of orthonormal variables and that, in addition, the matrix of factor loadings is diagonal. Under this additional assumption, a strict factor model is called a *normal factor model*. Note explicitly that under this assumption, factors are simply a set of standardized independent variables. The model is still undetermined under rotation, that is multiplication by any nonsingular matrix such that $\mathbf{R}\mathbf{R}' = \mathbf{I}$.

In summary, a set of variables has a normal factor representation if it is represented by the following factor model:

$$\mathbf{X} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{f} + \boldsymbol{\varepsilon}$$

where factors are orthonormal variables and noise terms are such that the covariance matrix can be represented as follows:

$$\Sigma = \boldsymbol{\beta}\boldsymbol{\beta}' + \Psi$$

where $\boldsymbol{\beta}$ is the diagonal matrix of factor loadings and Ψ is a diagonal matrix.

How can we explain the variety of factor models proposed given that a strict factor model could be uniquely identified up to a factor linear transformation? As mentioned, the assumptions underlying strict factor models are often too restrictive and approximate factor models have to be adopted. Approximate factor models are uniquely identifiable only in the limit of an infinite number of series. The level of approximation is implicit in practical models of returns.

Types of Factors and Their Estimation

In financial econometrics, the factors used in factor models can belong to three different categories: macroeconomic factors, fundamental factors, and statistical factors. The first two are factor models that deal with known factors and will not be discussed here.

Note that factors defined through statistical analysis are linear combinations of the variables. That is, if the variables are asset returns, factors are portfolios of assets. They are hidden variables insofar as one does not know the weights of the linear combinations. However, once the estimation process is completed, statistical factors are always linear combinations of variables. If data have a strict factor structure, we can always construct linear combinations of the series (e.g., portfolios of returns) that are perfectly correlated with a set of factors. Often they can be given important economic interpretations. In the following sections we describe the theory and estimation methods of principal components analysis and factor analysis.

PRINCIPAL COMPONENTS ANALYSIS

Principal components analysis (PCA) was introduced by Harold Hotelling (1933). Hotelling proposed PCA as a way to determine factors with statistical learning techniques when factors are not exogenously given. Given a variance-covariance matrix, one can determine factors using the technique of PCA.

PCA implements a dimensionality reduction of a set of observations. The concept of PCA is the following. Consider a set of n stationary time series X_i , for example the 500 series of returns of the S&P 500. Consider next a linear combination of these series, that is, a portfolio of securities. Each portfolio P is identified by an n -vector of weights ω_P and is characterized by a variance σ_P^2 . In general, the variance σ_P^2 depends on the portfolio's weights ω_P . Lastly, consider a normalized portfolio, which has the largest possible variance. In this context, a normalized portfolio is a portfolio such that the squares of the weights sum to one.

If we assume that returns are IID sequences, jointly normally distributed with variance-covariance matrix σ , a lengthy direct calculation demonstrates that each portfolio's return will be normally distributed with variance

$$\sigma_P^2 = \omega_P^T \sigma \omega_P$$

The normalized portfolio of maximum variance can therefore be determined in the following way:

$$\text{Maximize } \omega_P^T \sigma \omega_P$$

subject to the normalization condition

$$\omega_P^T \omega_P = 1$$

where the product is a scalar product. It can be demonstrated that the solution of this problem is the eigenvector ω_1 corresponding to the largest eigenvalue λ_1 of the variance-covariance matrix σ . As σ is a variance-covariance matrix, the eigenvalues are all real.

Consider next the set of all normalized portfolios orthogonal to ω_1 , that is, portfolios com-

pletely uncorrelated with ω_1 . These portfolios are identified by the following relationship:

$$\omega_1^T \omega_P = \omega_P^T \omega_1 = 0$$

We can repeat the previous reasoning. Among this set, the portfolio of maximum variance is given by the eigenvector ω_2 corresponding to the second largest eigenvalue λ_2 of the variance-covariance matrix σ . If there are n distinct eigenvalues, we can repeat this process n times. In this way, we determine the n portfolios P_i of maximum variance. The weights of these portfolios are the orthonormal eigenvectors of the variance-covariance matrix σ . Note that each portfolio is a time series that is a linear combination of the original time series X_i . The coefficients are the portfolios' weights.

These portfolios of maximum variance are all mutually uncorrelated. It can be demonstrated that we can recover all the original return time series as linear combinations of these portfolios:

$$X_j = \sum_{i=1}^n \alpha_{j,i} P_i$$

Thus far we have succeeded in replacing the original n correlated time series X_j with n uncorrelated time series P_i with the additional insight that each X_j is a linear combination of the P_i . Suppose now that only p of the portfolios P_i have a significant variance, while the remaining $n - p$ have very small variances. We can then implement a dimensionality reduction by choosing only those portfolios whose variance is significantly different from zero. Let's call these portfolios factors F .

It is clear that we can approximately represent each series X_i as a linear combination of the factors plus a small uncorrelated noise. In fact we can write

$$X_j = \sum_{i=1}^p \alpha_{j,i} F_i + \sum_{i=p+1}^n \alpha_{j,i} P_i = \sum_{i=1}^p \alpha_{j,i} F_i + \varepsilon_j$$

where the last term is a noise term. Therefore to implement PCA one computes the eigenvalues and the eigenvectors of the variance-covariance matrix and chooses the eigenvalues

significantly different from zero. The corresponding eigenvectors are the weights of portfolios that form the factors. Criteria of choice are somewhat arbitrary.

Suppose, however, that there is a strict factor structure, which means that returns follow a strict factor model as defined earlier in this entry:

$$r = a + \beta f + \varepsilon$$

The matrix β can be obtained diagonalizing the variance-covariance matrix. In general, the structure of factors will not be strict and one will try to find an approximation by choosing only the largest eigenvalues.

Note that PCA works either on the variance-covariance matrix or on the correlation matrix. The technique is the same but results are generally different. PCA applied to the variance-covariance matrix is sensitive to the units of measurement, which determine variances and covariances. This observation does not apply to returns, which are dimensionless quantities. However, if PCA is applied to prices and not to returns, the currency in which prices are expressed matters; one obtains different results in different currencies. In these cases, it might be preferable to work with the correlation matrix.

We have described PCA in the case of time series, which is the relevant case in econometrics. However, PCA is a generalized dimensionality reduction technique applicable to any set of multidimensional observations. It admits a simple geometrical interpretation, which can be easily visualized in the three-dimensional case. Suppose a cloud of points in the three-dimensional Euclidean space is given. PCA finds the planes that cut the cloud of points in such a way as to obtain the maximum variance.

Illustration of Principal Components Analysis

Let's now show how PCA is performed. To do so, we used monthly observations for the following 10 stocks: Campbell Soup, General

Dynamics, Sun Microsystems, Hilton, Martin Marietta, Coca-Cola, Northrop Grumman, Mercury Interactive, Amazon.com, and United Technologies for the period from December 2000 to November 2005. Figure 1 shows the graphics of the 10 return processes.

As explained earlier, performing PCA is equivalent to determining the eigenvalues and eigenvectors of the covariance matrix or of the correlation matrix. The two matrices yield different results. We perform both exercises, estimating the principal components using separately the covariance and the correlation matrices of the return processes. We estimate the covariance with the empirical covariance matrix. Recall that the empirical covariance σ_{ij} between variables (X_i, X_j) is defined as follows:

$$\hat{\sigma}_{ij} = \frac{1}{T} \sum_{t=1}^T (X_i(t) - \bar{X}_i)(X_j(t) - \bar{X}_j)$$

$$\bar{X}_i = \frac{1}{T} \sum_{t=1}^T X_i(t), \quad \bar{X}_j = \frac{1}{T} \sum_{t=1}^T X_j(t)$$

Table 1 shows the covariance matrix.

Normalizing the covariance matrix with the standard deviations, we obtain the correlation matrix. Table 2 shows the correlation matrix. Note that the diagonal elements of the correlation matrix are all equal to one. In addition, a number of entries in the covariance matrix are close to zero. Normalization by the product of standard deviations makes the same elements larger.

Let's now proceed to perform PCA using the covariance matrix. We have to compute the eigenvalues and the eigenvectors of the covariance matrix. Table 3 shows the eigenvectors (panel A) and the eigenvalues (panel B) of the covariance matrix.

Each column of panel A of Table 3 represents an eigenvector. The corresponding eigenvalue is shown in panel B. Eigenvalues are listed in descending order; the corresponding eigenvectors go from left to right in the matrix of eigenvectors. Thus the leftmost eigenvector corresponds

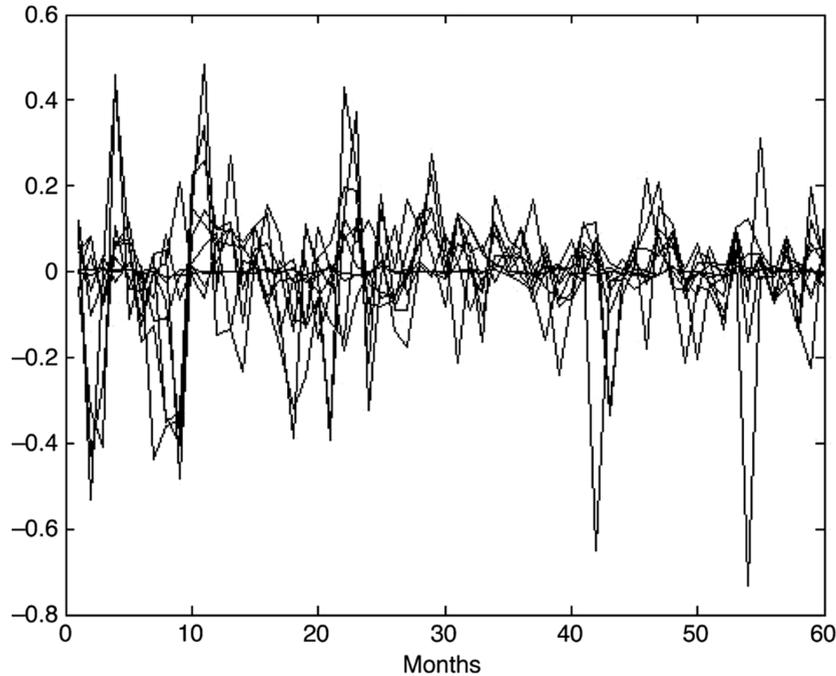


Figure 1 Graphics of the 10 Stock Return Processes

to the largest eigenvalue. Eigenvectors are not uniquely determined. In fact, multiplying any eigenvector for a real constant yields another eigenvector. The eigenvectors in Table 3 are normalized in the sense that the sum of the squares of each component is equal to 1. It can be easily checked that the sum of the squares of the elements in each column is equal to 1. This still leaves an indeterminacy, as we can change the

sign of the eigenvector without affecting this normalization.

As explained earlier, if we form portfolios whose weights are the eigenvectors, we can form 10 portfolios that are orthogonal (i.e., uncorrelated). These orthogonal portfolios are called *principal components*. The variance of each principal component will be equal to the corresponding eigenvector. Thus the first principal

Table 1 The Covariance Matrix of 10 Stock Returns

	SUNW	AMZN	MERQ	GD	NOC	CPB	KO	MLM	HLT	UTX
SUNW	0.02922	0.017373	0.020874	3.38E-05	-0.00256	-3.85E-05	0.000382	0.004252	0.006097	0.005467
AMZN	0.017373	0.032292	0.020262	5.03E-05	-0.00277	0.000304	0.001507	0.001502	0.010138	0.007483
MERQ	0.020874	0.020262	0.0355	-0.00027	-0.0035	-0.00011	0.003541	0.003878	0.007075	0.008557
GD	3.38E-05	5.03E-05	-0.00027	9.27E-05	0.000162	2.14E-05	-0.00015	3.03E-05	-4.03E-05	-3.32E-05
NOC	-0.00256	-0.00277	-0.0035	0.000162	0.010826	3.04E-05	-0.00097	0.000398	-0.00169	-0.00205
CPB	-3.85E-05	0.000304	-0.00011	2.14E-05	3.04E-05	7.15E-05	2.48E-05	-7.96E-06	-9.96E-06	-4.62E-05
KO	0.000382	0.001507	0.003541	-0.00015	-0.00097	2.48E-05	0.004008	-9.49E-05	0.001485	0.000574
MLM	0.004252	0.001502	0.003878	3.03E-05	0.000398	-7.96E-06	-9.49E-05	0.004871	0.00079	0.000407
HLT	0.006097	0.010138	0.007075	-4.03E-05	-0.00169	-9.96E-06	0.001485	0.00079	0.009813	0.005378
UTX	0.005467	0.007483	0.008557	-3.32E-05	-0.00205	-4.62E-05	0.000574	0.000407	0.005378	0.015017

Note: Sun Microsystems (SUNW), Amazon.com (AMZN), Mercury Interactive (MERQ), General Dynamics (GD), Northrop Grumman (NOC), Campbell Soup (CPB), Coca-Cola (KO), Martin Marietta (MLM), Hilton (HLT), United Technologies (UTX).

Table 2 The Correlation Matrix of the Same 10 Return Processes

	SUNW	AMZN	MERQ	GD	NOC	CPB	KO	MLM	HLT	UTX
SUNW	1	0.56558	0.64812	0.020565	-0.14407	-0.02667	0.035276	0.35642	0.36007	0.26097
AMZN	0.56558	1	0.59845	0.029105	-0.14815	0.20041	0.1325	0.11975	0.56951	0.33983
MERQ	0.64812	0.59845	1	-0.14638	-0.17869	-0.06865	0.29688	0.29489	0.37905	0.37061
GD	0.020565	0.029105	-0.14638	1	0.16217	0.26307	-0.24395	0.045072	-0.04227	-0.02817
NOC	-0.14407	-0.14815	-0.17869	0.16217	1	0.034519	-0.14731	0.054818	-0.16358	-0.16058
CPB	-0.02667	0.20041	-0.06865	0.26307	0.034519	1	0.046329	-0.01349	-0.0119	-0.04457
KO	0.035276	0.1325	0.29688	-0.24395	-0.14731	0.046329	1	-0.02147	0.23678	0.07393
MLM	0.35642	0.11975	0.29489	0.045072	0.054818	-0.01349	-0.02147	1	0.11433	0.047624
HLT	0.36007	0.56951	0.37905	-0.04227	-0.16358	-0.0119	0.23678	0.11433	1	0.44302
UTX	0.26097	0.33983	0.37061	-0.02817	-0.16058	-0.04457	0.07393	0.047624	0.44302	1

Note: Sun Microsystems (SUNW), Amazon.com (AMZN), Mercury Interactive (MERQ), General Dynamics (GD), Northrop Grumman (NOC), Campbell Soup (CPB), Coca-Cola (KO), Martin Marietta (MLM), Hilton (HLT), United Technologies (UTX).

component (i.e., the portfolio corresponding to the first eigenvalue), will have the maximum possible variance and the last principal component (i.e., the portfolio corresponding to the last eigenvalue) will have the smallest variance. Figure 2 shows the graphics of the principal components of maximum and minimum variance.

The 10 principal components thus obtained are linear combinations of the original series, $\mathbf{X} = (X_1, \dots, X_N)'$ that is, they are obtained by multiplying \mathbf{X} by the matrix of the eigenvectors. If the eigenvalues and the corresponding eigenvectors are all distinct, as it is the case in our illustration, we can apply the inverse

Table 3 Eigenvectors and Eigenvalues of the Covariance Matrix

Panel A: Eigenvectors											
	1	2	3	4	5	6	7	8	9	10	
1	-0.50374	0.50099	0.28903	-0.59632	-0.01824	-0.01612	0.22069	-0.08226	0.002934	-0.00586	
2	-0.54013	-0.53792	0.51672	0.22686	-0.06092	0.25933	-0.10967	-0.12947	0.020253	0.016624	
3	-0.59441	0.32924	-0.4559	0.52998	0.051976	0.015346	0.010496	0.21483	-0.01809	-0.00551	
4	0.001884	-0.00255	0.018107	-0.01185	0.013384	0.01246	-0.01398	0.01317	-0.86644	0.4981	
5	0.083882	0.10993	0.28331	0.19031	0.91542	-0.06618	0.14532	-0.02762	0.011349	-0.00392	
6	-0.00085	-0.01196	0.016896	0.006252	-0.00157	0.01185	-0.00607	-0.02791	-0.49795	-0.86638	
7	-0.0486	-0.02839	-0.1413	0.19412	-0.08989	-0.35435	0.31808	-0.8387	-0.01425	0.027386	
8	-0.07443	0.19009	0.013485	-0.06363	0.11133	-0.22666	-0.90181	-0.27739	0.010908	0.002932	
9	-0.20647	-0.36078	-0.01067	-0.1424	0.038221	-0.82197	0.052533	0.35591	-0.01155	-0.01256	
10	-0.20883	-0.41462	-0.5835	-0.46223	0.3649	0.27388	-0.02487	-0.14688	0.001641	-0.00174	
Panel B: Eigenvalues											
1	0.0783										
2	0.0164										
3	0.0136										
4	0.0109										
5	0.0101										
6	0.0055										
7	0.0039										
8	0.0028										
9	0.0001										
10	0.0001										

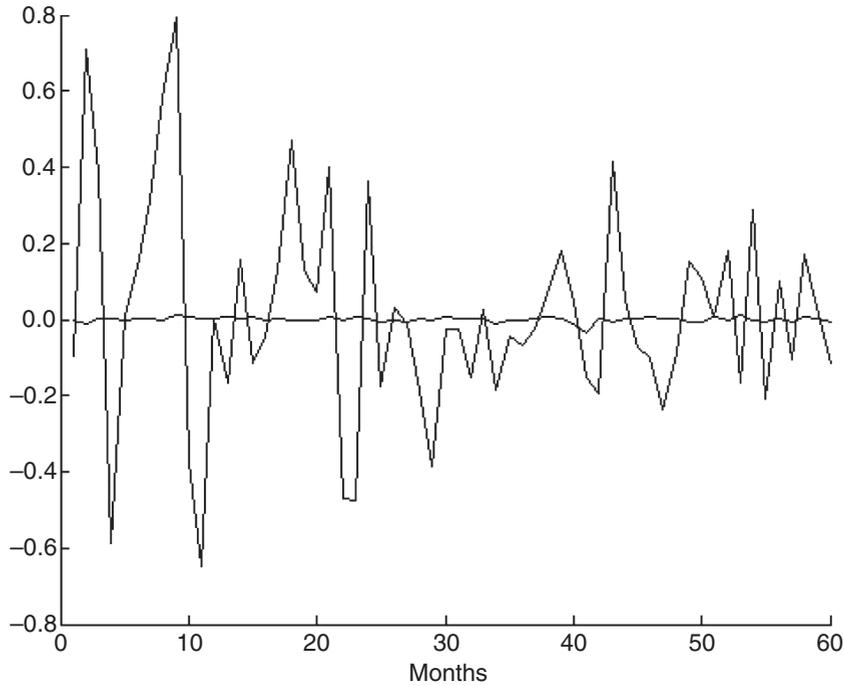


Figure 2 Graphic of the Portfolios of Maximum and Minimum Variance Based on the Covariance Matrix

transformation and recover the X as linear combinations of the principal components.

PCA is interesting if, in using only a small number of principal components, we nevertheless obtain a good approximation. That is, we use PCA to determine principal components but we use only those principal components that have a large variance as factors of a factor model. Stated otherwise, we regress the original series X onto a small number of principal components. In this way, PCA implements a dimensionality reduction as it allows one to retain only a small number of components. By choosing as factors the components with the largest variance, we can explain a large portion of the total variance of X .

Table 4 shows the total variance explained by a growing number of components. Thus the first component explains 55.2784% of the total variance, the first two components explain 66.8507% of the total variance, and so on. Obviously 10 components explain 100% of the

total variance. The second, third, and fourth columns of Table 5 show the residuals of the Sun Microsystems return process with 1, 5, and all 10 components, respectively. There is a large gain from 1 to 5, while the gain from 5 to all 10 components is marginal.

Table 4 Percentage of the Total Variance Explained by a Growing Number of Components Based on the Covariance Matrix

Principal Component	Percentage of Total Variance Explained
1	55.2784%
2	66.8508
3	76.4425
4	84.1345
5	91.2774
6	95.1818
7	97.9355
8	99.8982
9	99.9637
10	100.0000

Table 5 Residuals of the Sun Microsystems Return Process with 1, 5, and All Components Based on the Covariance Matrix and the Correlation Matrix

Month/Year	Residuals Based on Covariance Matrix			Residuals Based on Correlation Matrix		
	1 Principal Component	5 Principal Components	10 Principal Components	1 Principal Component	5 Principal Components	10 Principal Components
Dec. 2000	0.069044	0.018711	1.53E-16	0.31828	0.61281	-2.00E-15
Jan. 2001	-0.04723	-0.02325	1.11E-16	-0.78027	-0.81071	1.78E-15
Feb. 2001	-0.03768	0.010533	-1.11E-16	-0.47671	0.04825	2.22E-16
March 2001	-0.16204	-0.02016	2.50E-16	-0.47015	-0.82958	-2.78E-15
April 2001	-0.00819	-0.00858	-7.63E-17	-0.32717	-0.28034	-5.00E-16
May 2001	0.048814	-0.00399	2.08E-17	0.36321	0.016427	7.22E-16
June 2001	0.21834	0.025337	-2.36E-16	1.1437	1.37	7.94E-15
July 2001	-0.03399	0.02732	1.11E-16	-0.7547	0.35591	1.11E-15
Aug. 2001	0.098758	-0.00146	2.22E-16	1.0501	0.19739	-8.88E-16
Sept. 2001	0.042674	0.006381	-5.55E-17	0.40304	0.28441	2.00E-15
Oct. 2001	0.038679	-0.00813	-5.55E-17	0.50858	0.17217	4.44E-16
Nov. 2001	-0.11967	-0.01624	1.11E-16	-0.89512	-0.8765	-7.77E-16
Dec. 2001	-0.19192	0.030744	1.67E-16	-1.001	0.047784	-1.55E-15
Jan. 2002	-0.13013	-0.00591	5.55E-17	-1.1085	-0.68171	-1.33E-15
Feb. 2002	0.003304	0.017737	0	-0.05222	0.20963	-9.99E-16
March 2002	-0.07221	0.012569	5.55E-17	-0.35765	0.13344	2.22E-16
April 2002	-0.08211	-0.00916	2.78E-17	-0.38222	-0.47647	-2.55E-15
May 2002	-0.05537	-0.02103	0	-0.45957	-0.53564	4.22E-15
June 2002	-0.15461	0.004614	1.39E-16	-1.0311	-0.54064	-3.33E-15
July 2002	0.00221	0.013057	8.33E-17	0.24301	0.37431	-1.89E-15
Aug. 2002	-0.12655	0.004691	5.55E-17	-0.8143	-0.30497	2.00E-15
Sept. 2002	-0.07898	0.039666	5.55E-17	-0.25876	0.64902	-6.66E-16
Oct. 2002	0.15839	0.003346	-1.11E-16	0.98252	0.53223	-1.78E-15
Nov. 2002	-0.11377	0.013601	1.67E-16	-0.95263	-0.33884	-2.89E-15
Dec. 2002	-0.06957	0.012352	1.32E-16	-0.10309	0.029623	-4.05E-15
Jan. 2003	0.14889	-0.00118	-8.33E-17	1.193	0.73723	5.00E-15
Feb. 2003	-0.03359	-0.02719	-4.16E-17	-0.02854	-0.38331	4.05E-15
March 2003	-0.05314	-0.00859	2.78E-17	-0.38853	-0.40615	-2.22E-16
April 2003	0.10457	-0.01442	-2.22E-16	0.73075	0.097101	-1.11E-15
May 2003	0.078567	0.022227	-5.55E-17	0.52298	0.63772	-7.77E-16
June 2003	-0.1989	-0.02905	1.39E-16	-1.4213	-1.3836	-3.55E-15
July 2003	-0.0149	-0.00955	0	0.13876	-0.1059	3.44E-15
Aug. 2003	-0.12529	-0.00528	8.33E-17	-0.73819	-0.51792	9.99E-16
Sept. 2003	0.10879	-0.00645	-8.33E-17	0.69572	0.25503	-2.22E-15
Oct. 2003	0.07783	0.01089	-2.78E-17	0.36715	0.45274	-1.11E-15
Nov. 2003	0.038408	-0.01181	-5.55E-17	0.11761	-0.13271	3.33E-16
Dec. 2003	0.18203	0.012593	-1.39E-16	1.2655	0.98182	3.77E-15
Jan. 2004	0.063885	-0.00042	6.94E-18	0.33717	0.038477	0
Feb. 2004	-0.12552	-0.00225	1.11E-16	-0.70345	-0.49379	0
March 2004	-0.01747	0.016836	0	-0.1949	0.35348	-1.94E-16
April 2004	0.015742	0.013764	4.16E-17	0.2673	0.46969	-5.77E-15
May 2004	-0.03556	-0.02072	-6.94E-17	-0.60652	-0.68268	0
June 2004	0.14325	0.008155	-1.94E-16	0.54463	0.59768	3.22E-15
July 2004	0.030731	-0.00285	-4.16E-17	0.13011	0.028779	7.08E-16
Aug. 2004	0.032719	-0.00179	-5.55E-17	0.26793	0.18353	2.05E-15
Sept. 2004	0.083238	0.003664	0	0.58186	0.29544	3.77E-15
Oct. 2004	0.11722	-0.00356	-1.39E-16	0.77575	0.38959	2.22E-16
Nov. 2004	-0.04794	-0.00088	0	-0.47706	-0.35464	-3.13E-15
Dec. 2004	-0.1099	-0.01903	1.11E-16	-0.69439	-0.64663	-2.22E-16
Jan. 2005	0.0479	-0.00573	2.08E-17	0.24203	-0.04065	-4.45E-16
Feb. 2005	-0.015	0.003186	1.39E-17	-0.07198	0.054412	3.28E-15
March 2005	0.005969	-0.0092	-4.16E-17	0.035251	-0.02106	3.83E-15
April 2005	-0.00742	-0.01241	-4.16E-17	-0.09335	-0.42659	-1.67E-16
May 2005	0.14998	-0.01126	6.25E-17	1.0219	0.034585	-9.05E-15
June 2005	-0.05045	-0.00363	3.47E-17	-0.25655	-0.1229	-4.66E-15
July 2005	0.065302	-0.00421	-5.20E-17	0.56136	0.16602	3.08E-15
Aug. 2005	0.006719	-0.01174	1.39E-17	0.09319	-0.22119	-2.00E-15
Sept. 2005	0.12865	-0.00259	-8.33E-17	0.95602	0.33442	3.50E-15
Oct. 2005	-0.01782	0.011827	-8.33E-17	-0.2249	0.27675	1.53E-15
Nov. 2005	0.026312	-7.72E-05	-1.39E-17	0.26642	0.19725	1.67E-15

Table 6 Eigenvectors and Eigenvalues of the Correlation Matrix

Panel A: Eigenvectors										
	1	2	3	4	5	6	7	8	9	10
1	-0.4341	0.19295	-0.26841	0.040065	-0.19761	0.29518	-0.11161	-0.11759	-0.72535	-0.14857
2	-0.45727	0.18203	0.20011	0.001184	0.013236	0.37606	0.05077	0.19402	0.47275	-0.55894
3	-0.47513	-0.03803	-0.16513	0.16372	-0.01282	0.19087	-0.08297	-0.38843	0.37432	0.61989
4	0.06606	0.63511	0.18027	-0.16941	-0.05974	-0.24149	-0.66306	-0.14342	0.092295	0.02113
5	0.17481	0.33897	-0.21337	0.14797	0.84329	0.23995	0.091628	-0.07926	-0.06105	0.001886
6	-0.00505	0.42039	0.57434	0.40236	-0.15072	-0.05018	0.48758	-0.07382	-0.15788	0.19532
7	-0.18172	-0.397	0.28037	0.58674	0.26063	-0.26864	-0.38592	-0.16286	-0.11336	-0.24105
8	-0.1913	0.26851	-0.55744	0.32448	-0.09047	-0.58736	0.20083	0.19847	0.15935	-0.13035
9	-0.40588	-0.0309	0.20884	-0.20157	0.29193	-0.16641	-0.08666	0.67897	-0.1739	0.37201
10	-0.32773	-0.05042	0.14067	-0.51858	0.24871	-0.41444	0.30906	-0.4883	-0.06781	-0.17077

Panel B: Eigenvalues	
1	3.0652
2	1.4599
3	1.1922
4	0.9920
5	0.8611
6	0.6995
7	0.6190
8	0.5709
9	0.3143
10	0.2258

We can repeat the same exercise for the correlation matrix. Table 6 shows the eigenvectors (panel A) and the eigenvalues (panel B) of the correlation matrix. Eigenvectors are normalized as in the case of the covariance matrix.

Table 7 shows the total variance explained by a growing number of components. Thus the first component explains 30.6522% of the total variance, the first two components explain

Table 7 Percentage of the Total Variance Explained by a Growing Number of Components Using the Correlation Matrix

Principal Component	Percentage of Total Variance Explained
1	30.6522%
2	45.2509
3	57.1734
4	67.0935
5	75.7044
6	82.6998
7	88.8901
8	94.5987
9	97.7417
10	100.0000

45.2509% of the total variance, and so on. Obviously 10 components explain 100% of the total variance. The increase in explanatory power with the number of components is slower than in the case of the covariance matrix.

The proportion of the total variance explained grows more slowly in the correlation case than in the covariance case. Figure 3 shows the graphics of the portfolios of maximum and minimum variance. The ratio between the two portfolios is smaller in this case than in the case of the covariance.

The last three columns of Table 6 show the residuals of the Sun Microsystems return process with 1, 5, and all components based on the correlation matrix. Residuals are progressively reduced, but at a lower rate than with the covariance matrix.

PCA and Factor Analysis with Stable Distributions

In the previous sections we discussed PCA and factor analysis without making any explicit

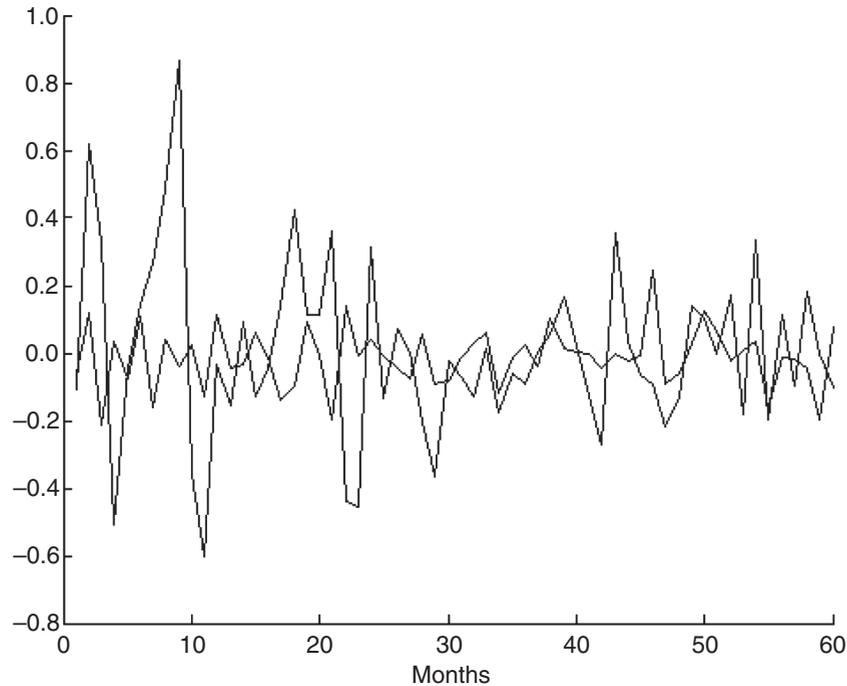


Figure 3 Graphic of the Portfolios of Maximum and Minimum Variance Based on the Correlation Matrix

reference to the distributional properties of the variables. These statistical tools can be applied provided that all variances and covariances exist. Therefore applying them does not require, per se, that distributions are normal, but only that they have finite variances and covariances. Variances and covariances are not robust but are sensitive to outliers. Robust equivalents of variances and covariances exist. In order to make PCA and factor analysis insensitive to outliers, one could use robust versions of variances and covariances and apply PCA and factor analysis to these robust estimates.

In many cases, however, distributions might exhibit fat tails and infinite variances. In this case, large values cannot be trimmed but must be taken into proper consideration. However, if variances and covariances are not finite, the least squares methods used to estimate factor loadings cannot be applied. In addition, the concept of PCA and factor analysis as illustrated in the previous sections cannot be

applied. In fact, if distributions have infinite variances, it does not make sense to determine the portfolio of maximum variance as all portfolios will have infinite variance and it will be impossible, in general, to determine an ordering based on the size of variance.

Both PCA and factor analysis as well as the estimation of factor models with infinite-variance error terms are at the forefront of econometric research.

FACTOR ANALYSIS

Thus far, we have seen how factors can be determined using principal components analysis. We retained as factors those principal components with the largest variance. In this section, we consider an alternative technique for determining factors: *factor analysis* (FA). Suppose we are given T independent samples of a random vector $\mathbf{X} = (X_1, \dots, X_N)'$. In the most common cases in financial econometrics, we will be given

T samples of a multivariate time series. However, factor analysis can be applied to samples extracted from a generic multivariate distribution. To fix these ideas, suppose we are given N time series of stock returns at T moments, as in the case of PCA.

Assuming that the data are described by a strict factor model with K factors, the objective of factor analysis (FA) consists of determining a model of the type

$$\mathbf{X} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{f} + \boldsymbol{\varepsilon}$$

with covariance matrix

$$\boldsymbol{\Sigma} = \boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{\Psi}$$

The estimation procedure is performed in two steps. In the first step, we estimate the covariance matrix and the factor loadings. In the second step, we estimate factors using the covariance matrix and the factor loadings.

If we assume that the variables are jointly normally distributed and temporally IID, we can estimate the covariance matrix with maximum likelihood methods. Estimation of factor models with maximum likelihood methods is not immediate because factors are not observable. Iterative methods such as the *expectation maximization* (EM) algorithm are generally used.

After estimating the matrices $\boldsymbol{\beta}$ and $\boldsymbol{\Psi}$ factors can be estimated as linear regressions. In fact, assuming that factors are zero means (an assumption that can always be made), we can write the factor model as

$$\mathbf{X} - \boldsymbol{\alpha} = \boldsymbol{\beta}\mathbf{f} + \boldsymbol{\varepsilon}$$

which shows that, at any given time, factors can be estimated as the regression coefficients of the regression of $(\mathbf{X} - \boldsymbol{\alpha})$ onto $\boldsymbol{\beta}$. Using the standard formulas of regression analysis, we can now write factors, at any given time, as follows:

$$\hat{\mathbf{f}}_t = \left(\hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\Psi}}^{-1} \hat{\boldsymbol{\beta}} \right)^{-1} \hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\Psi}}^{-1} (\mathbf{X}_t - \hat{\boldsymbol{\alpha}})$$

The estimation approach based on maximum likelihood estimates implies that the number of factors is known. In order to determine the number of factors, a heuristic procedure consists of iteratively estimating models with a

growing number of factors. The correct number of factors is determined when estimates of q factors stabilize and cannot be rejected on the basis of p probabilities. A theoretical method for determining the number of factors was proposed by Bai and Ng (2002).

The factor loadings matrix can also be estimated with ordinary least squares (OLS) methods. The OLS estimator of the factor loadings coincides with the principal component estimator of factor loadings. However, in a strict factor model, OLS estimates of the factor loadings are inconsistent when the number of time points goes to infinity but the number of series remains finite, unless we assume that the idiosyncratic noise terms all have the same variance.

The OLS estimators, however, remain consistent if we allow both the number of processes and the time to go to infinity. Under this assumption, as explained by Bai (2003), we can also use OLS estimators for approximate factor models.

In a number of applications, we might want to enforce the condition $\alpha = 0$. This condition is the condition of asset of arbitrage. OLS estimates of factor models with this additional condition are an instance of constrained OLS methods.

An Illustration of Factor Analysis

Let's now show how factor analysis is performed. To do so, we will use the same 10 stocks and return data for December 2000 to November 2005 that we used to illustrate principal components analysis.

As just described, to perform factor analysis, we need estimate only the factor loadings and the idiosyncratic variances of noise terms. We assume that the model has three factors. Table 8 shows the factor loadings. Each row represents the loadings of the three factors corresponding to each stock. The last column of the table shows the idiosyncratic variances.

The idiosyncratic variances are numbers between 0 and 1, where 0 means that the variance

Table 8 A Factor Loadings and Idiosyncratic Variances

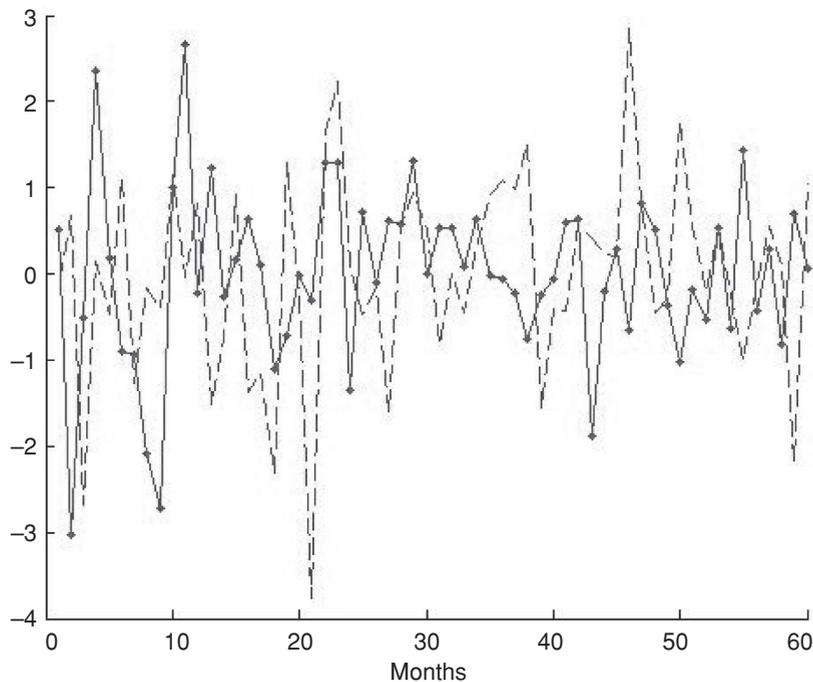
	Factor Loadings			Variance
	β_1	β_2	β_3	
SUNW	0.656940	0.434420	0.27910	0.301780
AMZN	0.959860	-0.147050	-0.00293	0.057042
MERQ	0.697140	0.499410	-0.08949	0.256570
GD	0.002596	-0.237610	0.43511	0.754220
NOC	-0.174710	-0.119960	0.23013	0.902130
CPB	0.153360	-0.344400	0.13520	0.839590
KO	0.170520	0.180660	-0.46988	0.717500
MLM	0.184870	0.361180	0.28657	0.753250
HLT	0.593540	0.011929	-0.18782	0.612300
UTX	0.385970	0.144390	-0.15357	0.806590

is completely explained by common factors and 1 that common factors fail to explain variance.

The p -value turns out to be 0.6808 and therefore fails to reject the null of three factors. Estimating the model with 1 and 2 factors we obtain much lower p -values while we run into numerical difficulties with 4 or more factors. We can therefore accept the null of three factors. Figure 4 shows the graphics of the three factors.

PCA AND FACTOR ANALYSIS COMPARED

The two illustrations of PCA and FA are relative to the same data and will help clarify the differences between the two methods. Let's first observe that PCA does not imply, per se, any specific restriction on the process. Given a nonsingular covariance matrix, we can always

**Figure 4** Graph of the three factors

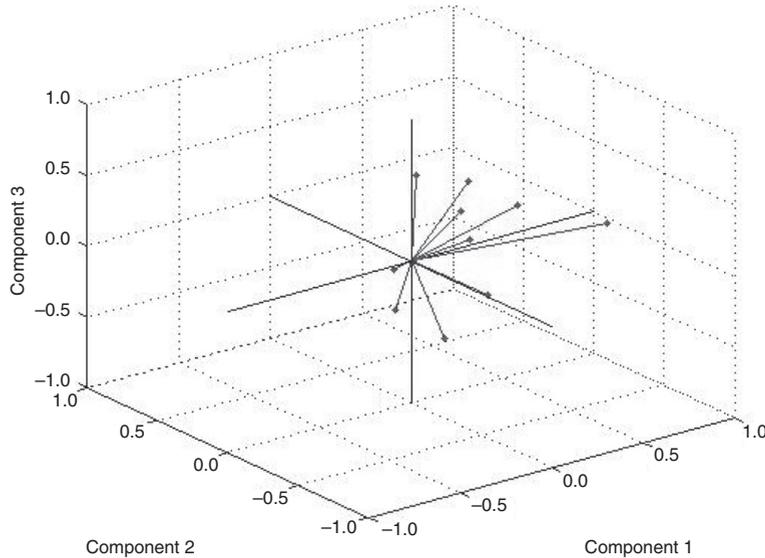


Figure 5 Graphical Representation of Factor Loadings

perform PCA as an exact linear transformation of the series. When we consider a smaller number of principal components, we perform an approximation that has to be empirically justified. For example, in our PCA illustration, the first three components explain 76% of the total variance (based on the covariance matrix; see Table 4).

Factor analysis, on the other hand, assumes that the data have a strict factor structure in the sense that the covariance matrix of the data can be represented as a function of the covariances between factors plus idiosyncratic variances. This assumption has to be verified, otherwise the estimation process might yield incorrect results.

In other words, PCA tends to be a dimensionality reduction technique that can be applied to any multivariate distribution and that yields incremental results. This means that there is a trade-off between the gain in estimation from dimensionality reduction and the percentage of variance explained. Consider that PCA is not an estimation procedure: It is an exact linear transformation of a time series. Estimation comes into play when a reduced number of princi-

pal components is chosen and each variable is regressed onto these principal components. At this point, a reduced number of principal components yields a simplified regression, which results in a more robust estimation of the covariance matrix of the original series though only a fraction of the variance is explained.

Factor analysis, on the other hand, tends to reveal the exact factor structure of the data. That is, FA tends to give an explanation in terms of what factors explain what processes. Factor rotation can be useful both in the case of PCA and FA. Consider FA. In our illustration, to make the factor model identifiable, we applied the restriction that factors are orthonormal variables. This restriction, however, might result in a matrix of factor loadings that is difficult to interpret.

For example, if we look at the loading matrix in Table 8, there is no easily recognizable structure, in the sense that the time series is influenced by all factors. Figure 5 shows graphically the relationship of the time series to the factors. In this graphic, each of the 10 time series is represented by its three loadings.

We can try to obtain a better representation through factor rotation. The objective is to

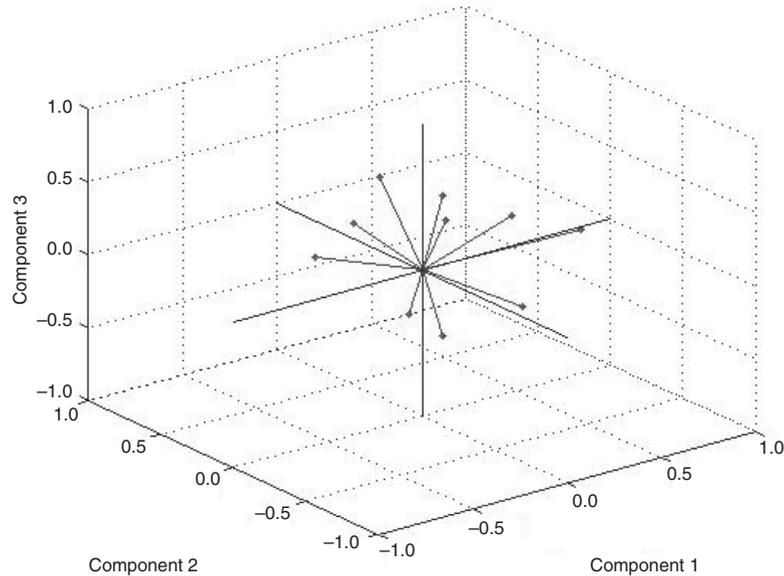


Figure 6 Relationship of Time Series to the Factors after Rotation

create factors such that each series has only one large loading and thus is associated primarily with one factor. Several procedures have been proposed for doing so. For example, if we rotate factors using the “promax” method, we obtain factors that are no longer orthogonal but that often have a better explanatory power. Figure 6 shows graphically the relationship of time series to the factors after rotation. The association of the series to a factor is more evident. This fact can be seen from the matrix of new factor loadings in Table 9, which shows how nearly each stock has one large loading.

Table 9 Factor Loadings after Rotation

	F1	F2	F3
SUNW	0.214020	0.750690	0.101240
AMZN	0.943680	0.127310	0.104990
MERQ	0.218340	0.578050	-0.294340
GD	0.163360	0.073269	0.544220
NOC	-0.070130	-0.003990	0.278000
CPB	0.393120	-0.178070	0.301920
KO	0.032397	-0.100020	-0.545120
MLM	-0.137130	0.561640	0.123670
HLT	0.513660	0.048842	-0.168290
UTX	0.229400	0.133510	-0.204650

KEY POINTS

- Principal component analysis (PCA) and factor analysis are statistical tools used in financial modeling to reduce the number of variables in a model (i.e., to reduce the dimensionality) and to identify a structure in the relationships between variables.
- Factor models seek to explain complex phenomena via a small number of basic causes or factors. In finance these models are typically applied to time series.
- The objective of a factor model in finance is to explain the behavior of a large number of stochastic processes typically price, returns, or rate processes in terms of a small number of factors (which themselves are stochastic processes). In financial modeling, factor models are needed not only to explain data but to make estimation feasible.
- Linear factor models are regression models. The coefficients are referred to as factor loadings or factor sensitivities, and they represent the influence of a factor on some variable.
- Principal components analysis is a tool to determine factors with statistical learning

techniques when factors are not exogenously given. PCA implements a dimensionality reduction of a set of observations.

- Performing PCA is equivalent to determining the eigenvalues and eigenvectors of the covariance matrix or of the correlation matrix.
- Factor analysis is an alternative technique for determining factors. The estimation procedure is performed in two steps: (1) estimate the covariance matrix and the factor loadings, and (2) estimate factors using the covariance matrix and the factor loadings.
- The covariance matrix can be estimated with maximum likelihood methods, assuming that the variables are jointly normally distributed and temporally independently and identically distributed. The estimation of models with maximum likelihood methods is not immediate because factors are not observable, and consequently iterative methods such as

the expectation maximization (EM) algorithm are generally used.

REFERENCES

- Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* 71: 135–171.
- Bai, J., and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* 70: 191–221.
- Hendrickson, A. E., and White, P. O. (1964). Pro-max: A quick method for rotation to orthogonal oblique structure. *British Journal of Statistical Psychology* 17: 65–70.
- Hotelling, H. (1933). Analysis of a complex of statistical variables with principal components. *Journal of Educational Psychology* 27: 417–441.
- Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology* 15: 201–293.
- Thurstone, L. L. (1938). *Primary Mental Abilities*. Chicago: University of Chicago Press.

Multifactor Equity Risk Models and Their Applications

ANTHONY LAZANAS, PhD
Managing Director, Barclays

ANTÓNIO BALDAQUE DA SILVA, PhD
Managing Director, Barclays

ARNE D. STAAL, PhD
Director, Barclays

CENK URAL, PhD
Vice President, Barclays

Abstract: Multifactor risk models seek to estimate and characterize the risk of a portfolio, either in absolute value or when compared against a benchmark. Risk is typically decomposed into a systematic and an idiosyncratic component. Systematic risk captures the exposures the portfolio has to broad risk factors. For equity portfolios these are typically countries, industries, fundamental (e.g., size), or technical (e.g., momentum). The portfolio systematic risk depends on its exposure to these risk factors, the volatility of the factors, and how they correlate with each other. Idiosyncratic risk captures the uncertainty associated with news affecting only individual issuers in the portfolio. This risk can be diversified by decreasing the importance of individual issuers in the portfolio. Intuitive multifactor risk models can provide relevant information regarding the major sources of risk in the portfolio. This information can be used to understand the important imbalances of the portfolio and guide the portfolio manager in constructing or rebalancing the portfolio. It can also be used in interpreting results from return attribution or scenario analysis.

Risk management is an integral part of the portfolio management process. *Risk models* are central to this practice, allowing managers to quantify and analyze the risk embedded in their

portfolios. Risk models provide managers insight into the major sources of risk in a portfolio, helping them to control their exposures and understand the contributions of different

The authors would like to thank Andy Sparks, Anuj Kumar, and Chris Sturhahn of Barclays for their help and comments.

portfolio components to total risk. They help portfolio managers in their decision-making process by providing answers to important questions such as: How does my small-cap exposure affect portfolio risk? Does my underweight in diversified financials hedge my overweight in banks? Risk models are also widely used in various other areas such as in portfolio construction, performance attribution, and scenario analysis.

In this entry, we discuss the structure of multifactor equity risk models, types of factors used in these models, and describe certain estimation techniques. We also illustrate the use of equity risk factor models in various applications, namely the analysis of portfolio risk, portfolio construction, scenario analysis, and performance attribution.

Throughout this entry, we will be using the Barclays Global Risk Model¹ for illustration purposes. For completeness, we also refer to other approaches one can take to construct such a model.

MOTIVATION

In this section, we discuss the motivation behind the multifactor equity risk models. Let's assume that a portfolio manager wants to estimate and analyze the volatility of a large portfolio of stocks. A straightforward idea would be to compute the volatility of the historical returns of the portfolio and use this measure to forecast future volatility. However, this framework does not provide any insight into the relationships between different securities in the portfolio or the major sources of risk. For instance, it does not assist a portfolio manager interested in diversifying her portfolio or constructing a portfolio that has better risk-adjusted performance.

Instead of estimating the portfolio volatility using historical portfolio returns, one could utilize a different strategy. The portfolio return is a function of stock returns and the market weights of these stocks in the portfolio. Us-

ing this, the forecasted volatility of the portfolio (σ_p) can be computed as a function of the weights (w) and the covariance matrix (Σ_s) of stock returns in the portfolio:

$$\sigma_p^2 = w^T \cdot \Sigma_s \cdot w$$

This covariance matrix can be decomposed into individual stock volatilities and the correlations between stock returns. Volatilities measure the riskiness of individual stock returns and correlations represent the relationships between the returns of different stocks. Looking into these correlations and volatilities, the portfolio manager can gain insight into her portfolio, namely the riskiness of different parts of the portfolio or how the portfolio can be diversified. As we outlined above, to estimate the portfolio volatility we need to estimate the correlation between each pair of stocks. Unfortunately, this means that the number of parameters to be estimated grows quadratically with the number of stocks in the portfolio.² For most practical portfolios, the relatively large number of stocks makes it difficult to estimate the relationship between stock returns in a robust way. Moreover, this framework uses the history of individual stock returns to forecast future stock volatility. However, stock characteristics are dynamic and hence using returns from different time periods may not produce good forecasts.³ Finally, the analysis does not provide much insight regarding the broad factors influencing the portfolio. These drawbacks constitute the motivation for the multifactor risk models detailed in this entry.

One of the major goals of multifactor risk models is to describe the return of a portfolio using a smaller set of variables, called factors. These factors should be designed to capture broad (systematic) market fluctuations, but should also be able to capture specific nuances of individual portfolios. For instance, a broad U.S. market factor would capture the general movement in the equity market, but not the varying behavior across industries. If our portfolio is heavily biased toward particular

industries, the broad U.S. market factor may not allow for a good representation of our portfolio's return.

In the context of factor models, the total return of a stock is decomposed into a systematic and an idiosyncratic component. Systematic return is the component of total return due to movements in common risk factors, such as industry or size. On the other hand, idiosyncratic return can be described as the residual component that cannot be explained by the systematic factors. Under these models, the idiosyncratic return is uncorrelated across issuers. Therefore, correlations across securities are driven by their exposures to the systematic risk factors and the correlation between those factors.

The following equation demonstrates the systematic and the idiosyncratic components of total stock return:

$$r_s = L_s \cdot F + \varepsilon_s$$

The systematic return for security s is the product of the loadings of that security (L_s , also called sensitivities) to the systematic risk factors and the returns of these factors (F). The idiosyncratic return is given by ε_s . Under these models, the portfolio volatility can be estimated as

$$\sigma_p^2 = L_p^T \cdot \Sigma_F \cdot L_p + w^T \cdot \Omega \cdot w$$

Models represented by equations of this form are called linear factor models. Here L_p represents the loadings of the portfolio to the risk factors (determined as the weighted average of individual stock loadings), and Σ_F is the covariance matrix of factor returns. w is the vector of security weights in the portfolio, and Ω is the covariance matrix of idiosyncratic stock returns. Due to the uncorrelated nature of these returns, this covariance matrix is diagonal: all elements outside its diagonal are zero. As a result, the idiosyncratic risk of the portfolio is diversified away as the number of securities in the portfolio increases. This is the diversification benefit attained when combining uncorrelated exposures.

For most practical portfolios, the number of factors is significantly smaller than the number of stocks in the portfolio. Therefore, the number of parameters in Σ_F is much smaller than in Σ_S , leading to a generally more robust estimation. Moreover, the factors can be designed in a way that they are relatively more stable than individual stock returns, leading to models with potentially better predictability.

Another important advantage of using linear factor models is the detailed insight they provide into the structure and properties of portfolios. These models characterize stock returns in terms of systematic factors that (can) have intuitive economic interpretations. Linear factor models can provide important insights regarding the major systematic and idiosyncratic sources of risk and return. This analysis can help managers to better understand their portfolios and can guide them through the different tasks they perform, such as rebalancing, hedging, or the tilting of their portfolios. The Barclays Global Risk Model—the model used for illustration throughout this entry—is an example of such a linear factor model.

EQUITY RISK FACTOR MODELS

The design of a linear factor model usually starts with the identification of the major sources of risk embedded in the portfolios of interest. For an equity portfolio manager who invests in various markets across the globe, the major sources of risk are typically country, industry membership, and other fundamental or technical exposures such as size, value, and momentum. The relative significance of these components varies across different regions. For instance, for regional equity risk models in developed markets, industry factors tend to be more important than country factors, although in periods of financial distress country factors become more significant. On the other hand, for emerging markets models the country factor is still considered to be the most important

source of risk. For regional models, the relative significance of industry factors depends on the level of financial integration across different local markets in that region. The importance of these factors is also time-varying, depending on the particular time period of the analysis. For instance, country risk used to be a large component of total risk for European equity portfolios. However, country factors have been generally losing their significance in this context due to financial integration in the region as a result of the European Union and a common currency, the euro. This is particularly true for larger European countries. Similarly, the relative importance of industry factors is higher over the course of certain industry-led crises, such as the dot-com bubble burst (2000–2002) and the 2007–2009 banking and credit crisis. As we will see, the relative importance of different risk factors varies also with the particular design and the estimation process chosen to calibrate the model.

A typical global or regional equity risk model has the following structure:

$$r_i = \beta_i^{\text{MKT}} \cdot F^{\text{MKT}} + \beta_i^{\text{IND}} \cdot F^{\text{IND}} + \beta_i^{\text{CNT}} \cdot F^{\text{CNT}} + \sum_{j=1}^n \ell_{ij} \cdot F_j^{\text{FT}} + \varepsilon_i$$

where

- r_i = the rate of return for stock i
- F^{MKT} = the market factor
- F^{IND} = the industry factor corresponding to stock i
- F^{CNT} = the country factor corresponding to stock i
- β_i = the exposure (beta) of the stock to the corresponding factor
- F^{FT} = the set of fundamental and technical factors
- ℓ_{ij} = the loading of stock i to factor F_j^{FT}
- ε_i = the residual return for stock i

There are different ways in which these factors can be incorporated into an equity risk model. The choice of a particular model

affects the interpretation of the factors. For instance, consider a model that has only market and industry factors. Industry factors in such a model would represent industry-specific moves net of the market return. On the other hand, if we remove the market factor from the equation, the industry factors now incorporate the overall market effect. Their interpretation would change, with their returns now being close to market value-weighted industry indexes. Country-specific risk models are a special case of the previous representation where the country factor disappears and the market factor is represented by the returns of the countrywide market. Macroeconomic factors are also used in some equity risk models, as discussed later.

The choice of estimation process also influences the interpretation of the factors. As an example, consider a model that has only industry and country factors. These factors can be estimated jointly in one step. In this case, both factors represent their own effect net of the other ones. On the other hand, these factors can be estimated in a multistep process—e.g., industry factors estimated in the first step and then the country factors estimated in the second step, using residual returns from the first step. In this case, the industry factors have an interpretation close to the market value-weighted industry index returns, while the country factors would now represent a residual country average effect, net of industry returns. We discuss this issue in more detail in the following section.

Model Estimation

In terms of the estimation methodology, there are three major types of multi-factor equity risk models: cross-sectional, time series, and statistical. All three of these methodologies are widely used to construct linear factor models in the equity space.⁴ In cross-sectional models, loadings are known and factors are estimated. Examples of loadings used in these models are industry

membership variables and fundamental security characteristics (e.g., the book-to-price ratio). Individual stock returns are regressed against these security-level loadings in every period, delivering estimation of factor returns for that period. The interpretation of these estimated factors is usually intuitive, although dependent on the estimation procedure and on the quality of the loadings. In time-series models, factors are known and loadings are estimated. Examples of factors in these models are financial or macroeconomic variables, such as market returns or industrial production. Time series of individual equity returns are regressed against the factor returns, delivering empirical sensitivities (loadings or betas) of each stock to the risk factors. In these models, factors are constructed and not estimated, therefore, their interpretation is straightforward. In statistical models (e.g., principal component analysis), both factors and loadings are estimated jointly in an iterative fashion. The resulting factors are statistical in nature, not designed to be intuitive. That being said, a small set of the statistical factors can be (and usually are) correlated with broad economic factors, such as the market. Table 1 summarizes some of the characteristics of these models.

An important advantage of cross-sectional models is that the number of parameters to be

estimated is generally significantly smaller as compared to the other two types of models. On the other hand, cross-sectional models require a much larger set of input data (company-specific loadings). Cross-sectional models tend to be relatively more responsive as loadings can adjust faster to changing market conditions. There are also hybrid models, which combine cross-sectional and time-series estimation in an iterative fashion; these models allow the combination of observed and estimated factors. Finally, statistical models require only a history of security returns as input to the process. They tend to work better when economic sources of risk are hard to identify and are primarily used in high-frequency applications.

As we mentioned in the previous section, the estimation process is a major determinant in the interpretation of factors. Estimating all factors jointly in one-step regression allows for a natural decomposition of total variance in stock returns. However it also complicates the interpretation of factors as each factor now represents its own effect net of all other factors. Moreover, multicollinearity problems arise naturally in this set-up, potentially delivering lack of robustness to the estimation procedure and leading to unintuitive factor realizations. This problem can be serious when using factors that are highly correlated.

Table 1 Cross-Sectional, Time-Series, and Statistical Factor Models

Model	Cross-Sectional	Time-Series	Statistical
Input set	Security-specific loadings and returns	Factor and security returns	Security returns
Factors and loadings	Factors are estimated using the known loadings (e.g., industry beta or momentum score)	Factors are known (e.g., market or industrial production) and loadings are estimated (e.g., industry or momentum betas)	Both factors and loadings are estimated
Interpretation	Clean interpretation of loadings; generally intuitive interpretation of factors	Straightforward interpretation of factors	Factors may have no intuitive interpretation
Number of parameters	(No. of factors) × (No. of time periods)	(No. of securities) × (No. of factors)	(No. of securities) × (No. of factors)

An alternative in this case is to use a multistep estimation process where different sets of factors are estimated sequentially, in separate regressions. In the first step, stock returns are used in a regression to estimate a certain set of factors, and then residual returns from this step are used to estimate the second step factors, and so on. The choice of the order of factors in such estimation influences the nature of the factors and their realizations. This choice should be guided by the significance and the desired interpretation of the resulting factors. The first-step factors have the most straightforward interpretation as they are estimated in isolation from all other factors using raw stock returns. For instance, in a country-specific equity risk model where there are industry, fundamental and technical factors, the return series of industry factors would be close to the industry index returns if they are estimated in isolation in the first step. This would not be the case if all industry, fundamental, and technical factors are estimated in the same step.

An important input to the model estimation is the security weights used in the regressions. There is a variety of techniques employed in practice but generally more weight is assigned to less volatile stocks (usually represented by larger companies). This enhances the robustness of the factor estimates as stocks from these companies tend to have relatively more stable return distributions.

Types of Factors

In this section, we analyze in more detail the different types of factors typically used in equity risk models. These can be classified under five major categories: market factors, classification variables, firm characteristics, macroeconomic variables, and statistical factors.

Market Factors

A market factor can be used as an observed factor in a time-series setting (e.g., in the cap-

ital asset pricing model, the market factor is the only systematic factor driving returns). As an example, for a U.S. equity factor model, S&P 500 can be used as a market factor and the loading to this factor—market beta—can be estimated by regressing individual stock returns to the S&P 500. On the other hand, in a cross-sectional setting, the market factor can be estimated by regressing stock returns to their market beta for each time period (this beta can be empirical—estimated via statistical techniques—or set as a dummy loading, usually 1). When incorporated into a cross-sectional regression with other factors, it generally works as an intercept, capturing the broad average return for that period. This changes the interpretation of all other factors to returns relative to that average (e.g., industry factor returns would now represent industry-specific moves net of market).

Classification Variables

Industry and country are the most widely used classification variables in equity risk models. They can be used as observed factors in time-series models via country/industry indexes (e.g., return series of GICS indexes⁵ can be used as observed industry factors). In a cross-sectional setting, these factors are estimated by regressing stock returns to industry/country betas (either estimated or represented as a 0/1 dummy loading). These factors constitute a significant part of total risk for a majority of equity portfolios, especially for portfolios tilted toward specific industries or countries.

Firm Characteristics

Factors that represent firm characteristics can be classified as either fundamental or technical factors. These factors are extensively used in equity risk models; exposures to these factors represent tilts towards major investment themes such as size, value, and momentum. Fundamental factors generally employ a mix of

accounting and market variables (e.g., accounting ratios) and technical factors commonly use return and volume data (e.g., price momentum or average daily volume traded).

In a time-series setting, these factors can be constructed as representative long-short portfolios (e.g., Fama-French factors). As an example, the value factor can be constructed by taking a long position in stocks that have a large book to price ratio and a short position in the stocks that have a small book to price ratio. On the other hand, in a cross-sectional setup, these factors can be estimated by regressing the stock returns to observed firm characteristics. For instance, a book to price factor can be estimated by regressing stock returns to the book to price ratios of the companies. In practice, fundamental and technical factors are generally estimated jointly in a multivariate setting.

A popular technique in the cross-sectional setting is the standardization of the characteristic used as loading such that it has a mean of zero and a standard deviation of one. This implies that the loading to the corresponding factor is expressed in relative terms, making the exposures more comparable across the different fundamental/technical factors. Also, similar characteristics can be combined to form a risk index and then this index can be used to estimate the relevant factor (e.g., different value ratios such as earnings to price and book to price can be combined to construct a value index, which would be the exposure to the value factor). The construction of an index from similar characteristics can help reduce the problem of multicollinearity referred to above. Unfortunately, it can also dilute the signal each characteristic has, potentially reducing its explanatory power. This trade-off should be taken into account while constructing the model. The construction of fundamental factors and their loadings requires careful handling of accounting data. These factors tend to become more significant for portfolios that are hedged with respect to the market or industry exposures.

Macroeconomic Variables

Macroeconomic factors, representing the state of the economy, are generally used as observed factors in time-series models. Widely used examples include interest rates, commodity indexes, and market volatility (e.g., the VIX index). These factors tend to be better suited for models with a long horizon. For short to medium horizons, they tend to be relatively insignificant when included in a model that incorporates other standard factors such as industry. The opposite is not true, suggesting that macro factors are relatively less important for these horizons. This does not mean that the macroeconomic variables are not relevant in explaining stock returns; it means that a large majority of macroeconomic effects can be captured through the industry factors. Moreover, it is difficult to directly estimate stock sensitivities to slow-moving macroeconomic variables. These considerations lead to the relatively infrequent use of macro variables in short to medium horizon risk models.⁶

Statistical Factors

Statistical factors are very different in nature from all the aforementioned factors as they do not have direct economic interpretation. They are estimated using statistical techniques such as principal component analysis where both factors and loadings are estimated jointly in an iterative fashion. Their interpretation can be difficult, yet in certain cases they can be re-mapped to well-known factors. For instance, in a principal component analysis model for the U.S. equity market, the first principal component would represent the U.S. market factor. These models tend to have a relatively high in-sample explanatory power with a small set of factors and the marginal contribution of each factor tends to diminish significantly after the first few factors. Statistical factors can also be used to capture the residual risk in a model with economic factors. These factors tend to work better when there are unidentified sources of risk such as in the case of high-frequency models.

Other Considerations in Factor Models

Various quantitative and qualitative measures can be employed to evaluate the relative performance of different model designs. Generically, better risk models are able to forecast more accurately the risk of different types of portfolios across different economic environments. Moreover, a better model allows for an intuitive analysis of the portfolio risk along the directions used to construct and manage the portfolio. The relative importance of these considerations should frame how we evaluate different models.

A particular model is defined by its estimation framework and the selection of its factors and loadings. Typically, these choices are evaluated jointly, as the contributions of specific components are difficult to measure in practice. Moreover, decisions on one of these components (partially) determine the choice of the others. For instance, if a model uses fundamental firm characteristics as loadings, it also uses estimated factors—more generally, decisions on the nature of the factors determine the nature of the loadings and vice-versa.

Quantitative measures of factor selection include the explanatory power or significance of the factor, predictability of the distribution of the factor, and correlations between factors. On a more qualitative perspective, portfolio managers usually look for models with factors and loadings that have clean and intuitive interpretation, factors that correspond to the way they think about the asset class, and models that reflect their investment characteristics (e.g., short vs. long horizon, local vs. global investors).

Idiosyncratic Risk

Once all systematic factors and loadings are estimated, the residual return can be computed as the component of total stock return that cannot be explained by the systematic factors. Idiosyncratic return—also called residual, nonsystematic, or name-specific return—can

be a significant component of total return for individual stocks, but tends to become smaller for portfolios of stocks as the number of stocks increases and concentration decreases (the aforementioned diversification effect). The major input to the computation of idiosyncratic risk is the set of historical idiosyncratic returns of the stock. Because the nature of the company may change fast, a good idiosyncratic risk model should use only recent and relevant idiosyncratic returns. Moreover, recent research suggests that there are other conditional variables that may help improve the accuracy of idiosyncratic risk estimates. For instance, there is substantial evidence that the market value of a company is highly correlated with its idiosyncratic risk, where larger companies exhibit relatively smaller idiosyncratic risk. The use of such variables as an extra adjustment factor can improve the accuracy of idiosyncratic risk estimates.

As mentioned before, idiosyncratic returns of different issuers are assumed to be uncorrelated. However, different securities from the same issuer can show a certain level of co-movement, as they are all exposed to specific events affecting their common issuer.

Interestingly, this co-movement is not perfect or static. Certain news can potentially affect the different securities issued by the same company (e.g., equity, bonds, or equity options) in different ways. Moreover, this relationship changes with the particular circumstances of the firm. For instance, returns from securities with claims to the assets of the firm should be more highly correlated if the firm is in distress. A good risk model should be able to capture these phenomena.

APPLICATIONS OF EQUITY RISK MODELS

Multifactor equity risk models are employed in various applications such as the quantitative analysis of portfolio risk, hedging unwanted exposures, portfolio construction, scenario

analysis, and performance attribution. In this section we discuss and illustrate some of these applications.

Portfolio managers can be divided broadly into indexers (those that measure their returns relative to a benchmark index) and absolute return managers (typically hedge fund managers). In between stand the enhanced indexers, those that are allowed to deviate from the benchmark index in order to express views, presumably leading to superior returns. All are typically subject to a risk budget that prescribes how much risk they are allowed to take to achieve their objectives: minimize transaction costs and match the index return for the pure indexers, maximize the net return for the enhanced indexers, or maximize absolute return for absolute return managers. In all of these cases, the manager has to merge all her views and constraints into a final portfolio.

The investment process of a typical portfolio manager involves several steps. Given the investment universe and objective, the steps usually consist of portfolio construction, risk prediction, and performance evaluation. These steps are iterated throughout the investment cycle over each rebalancing period. The examples in this section are constructed following these steps. In particular, we start with a discussion on the portfolio construction process for three equity portfolio managers with different goals: The first aims to track a benchmark, the second to build a momentum portfolio, and the third to implement sector views in a portfolio. We conduct these exercises through a risk-based portfolio optimization approach at a monthly rebalancing frequency. For the index-tracking portfolio example, we then conduct a careful evaluation of its risk exposures and contributions to ensure that the portfolio manager's views and intuition coincide with the actual portfolio exposures. Once comfortable with the positions and the associated risk, the portfolio is implemented. At the end of the monthly investment cycle, the performance of the portfolio and return contributions of its different compo-

nents can be evaluated using performance attribution.

Scenario analysis can be employed in both the portfolio construction and the risk evaluation phases of the portfolio process. This exercise allows the manager to gain additional intuition regarding the exposures of her portfolio and how it may behave under particular economic circumstances. It usually takes the form of stress testing the portfolio under historical or hypothetical scenarios. It can also reveal the sensitivity of the portfolio to particular movements in economic and financial variables not explicitly considered during the portfolio construction process. The last application in this entry illustrates this kind of analysis.

Throughout our discussion, we use a suite of global cash equity risk models available through POINT[®], the Barclays portfolio analytics and modeling platform.⁷

Portfolio Construction

Broadly speaking there are two main approaches to portfolio construction: a formal quantitative optimization-based approach and a qualitative approach that is based primarily on manager intuition and skill. There are many variations within and between these two approaches. In this section, we focus on risk-based optimization using a linear factor model. We do not discuss other more qualitative or nonrisk-based approaches (e.g., a stratified sampling). A common objective in a risk-based optimization exercise is the minimization of volatility of the portfolio, either in isolation or when evaluated against a benchmark. In the context of multifactor risk models, total volatility is composed of a systematic and an idiosyncratic component, as described above. Typically, both of these components are used in the objective function of the optimization problems. We demonstrate three different portfolio construction exercises and discuss how equity factor models are employed in this endeavor. The examples were constructed using the POINT[®] Optimizer.⁸ All

optimization problems were run as of July 30, 2010.

Tracking an Index

In our first example, we study the case of a portfolio manager whose goal is to create a portfolio that tracks a benchmark equity index as closely as possible, using a limited number of stocks. This is a very common problem in the investment industry since most assets under management are benchmarked to broad market indexes. Creating a benchmark-tracking portfolio provides a good starting point for implementing strategic views relative to that benchmark. For example, a portfolio manager might have a mandate to outperform a benchmark under particular risk constraints. One way to implement this mandate is to dynamically tilt the tracking portfolio toward certain investment styles based on views on the future performance of those styles at a particular point in the business cycle.

Consider a portfolio manager who is benchmarked to the S&P 500 index and aims to build a tracking portfolio composed of long-only positions from the set of S&P 500 stocks. Because of transaction cost and position management limitations, the portfolio manager is restricted to a maximum number of 50 stocks in the tracking portfolio. Her objective is to minimize the tracking error volatility (TEV) between her portfolio and the benchmark. Tracking error volatility can be described as the volatility of the return differential between the portfolio and the benchmark (i.e., measures a typical movement in this net position). A portfolio's TEV is commonly referred to as the risk or the (net) volatility of the portfolio.

As mentioned before, the total TEV is decomposed into a systematic TEV and an idiosyncratic TEV. Moreover, because these two components are assumed to be independent,

$$\begin{aligned} \text{Total TEV} &= \sqrt{\text{Systematic TEV}^2 + \text{Idiosyncratic TEV}^2} \end{aligned}$$

Table 2 Total Risk of Index-Tracking Portfolio vs. the Benchmark (bps/month)

Attribute	Realized Value
Total TEV	39.6
Idiosyncratic TEV	35.8
Systematic TEV	16.9

The minimization of systematic TEV is achieved by setting the portfolio's factor exposures (net of benchmark) as close to zero as possible, while respecting other potential constraints of the problem (e.g., maximum number of 50 securities in the portfolio). The minimization of idiosyncratic volatility is achieved through the diversification of the portfolio holdings.

Table 2 illustrates the total risk for portfolio versus benchmark that comes out of the optimization problem. We see that total TEV of the net position is 39.6 bps/month with 16.9 bps/month of systematic TEV and 35.8 bps/month of idiosyncratic TEV. If the portfolio manager wants to reduce her exposure to name-specific risk, she can increase the upper bound on the number of securities picked by the optimizer to construct the optimal portfolio (increasing the diversification effect). Another option would be to increase the relative weight of idiosyncratic TEV compared to the systematic TEV in the objective function. The portfolio resulting from this exercise would have smaller idiosyncratic risk but, unfortunately, would also have higher systematic risk. This trade-off can be managed based on the portfolio manager's preferences.

Figure 1 depicts the distribution of the position amount for individual stocks in the portfolio. We can see that the portfolio is well diversified across the 50 constituent stocks with no significant concentrations in any of the individual positions. The largest stock position is 4.1%, about three times larger than the smallest holding. Later in this entry, we analyze the risk of this particular portfolio in more detail.

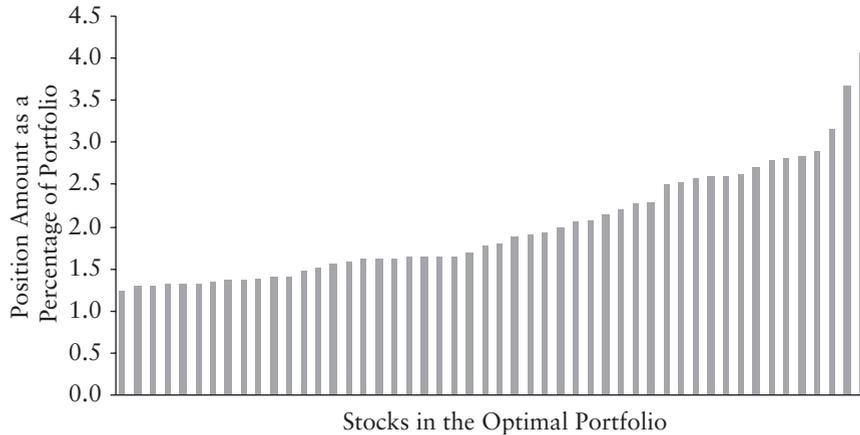


Figure 1 Position Amount of Individual Stocks in the Optimal Tracking Portfolio

Constructing a Factor-Mimicking Portfolio

Factor-mimicking portfolios allow portfolio managers to capitalize on their views on various investment themes. For instance, the portfolio manager may forecast that small-cap stocks will outperform large-cap stocks or that value stocks will outperform growth stocks in the near future. By constructing long-short factor-mimicking portfolios, managers can place positions in line with their views on these investment themes without taking explicit directional views on the broader market.

Considering another example, suppose our portfolio manager forecasts that recent winner (high momentum) stocks will outperform recent losers (low momentum). To implement her

views, she constructs two portfolios, one with winner stocks and one with loser stocks (100 stocks from the S&P 500 universe in each portfolio). She then takes a long position in the winners portfolio and a short position in the losers portfolio. While a sensible approach, a long-short portfolio constructed in this way would certainly have exposures to risk factors other than momentum. For instance, the momentum view might implicitly lead to unintended sector bets. If the portfolio manager wants to understand and potentially limit or avoid these exposures, she needs to perform further analysis. The use of a risk model will help her substantially.

To illustrate this point, table 3 shows one of POINT[®]'s risk model outputs—the 10

Table 3 Largest Risk Factor Exposures for the Momentum Winners/Losers Portfolio (bps/month)

Factor Name	Sensitivity/ Exposure	Net Exposure	Factor Volatility	Contribution to TEV
EQUITIES DEVELOPED MARKETS				
U.S. Equity Energy	Empirical beta	-0.094	651	25.3
U.S. Equity Materials	Empirical beta	-0.045	808	15.9
U.S. Equity CYC Media	Empirical beta	0.027	759	-9.9
U.S. Equity FIN Banks	Empirical beta	0.088	900	13.0
U.S. Equity FIN Diversified Financials	Empirical beta	-0.108	839	39.6
U.S. Equity FIN Real Estate	Empirical beta	0.100	956	-19.0
U.S. Equity TEC Software	Empirical beta	-0.057	577	17.2
U.S. Equity TEC Semiconductors	Empirical beta	-0.029	809	9.9
U.S. Equity Corporate Default Probability	CDP	-0.440	76	23.2
U.S. Equity Momentum (9m)	Momentum	1.491	73	74.9

largest risk factor exposures by their contribution to TEV (last column in the table) for the initial long-short portfolio. While momentum has the largest contribution to volatility, other risk factors also play a significant role. As a result, major moves in risk factors other than momentum can have a significant—and potentially unintended—impact on the portfolio's return.

Given this information, suppose our portfolio manager decides to avoid these exposures to the extent possible. She can do that by setting all exposures to factors other than momentum to zero (these type of constraints may not always be feasible and one may need to relax them to achieve a solution). Moreover, because she wants the portfolio to represent a pure systematic momentum effect, she seeks to minimize idiosyncratic risk. There are many ways to implement these additional goals, but increasingly portfolio managers are turning to risk models (using an optimization engine) to construct their portfolios in a robust and convenient way. She decides to set up an optimization problem where the objective function is the minimization of idiosyncratic risk. The tradable universe is the set of S&P 500 stocks and the portfolio is constructed to be dollar-neutral. This problem also incorporates the aforementioned factor exposure constraints.

The resulting portfolio (not shown) has exactly the risk factor exposures that were specified in the problem constraints. It exhibits a relatively low idiosyncratic TEV. Figure 2 depicts the largest 10 positions on the long and short sides of the momentum factor-mimicking portfolio; we see that there are no significant individual stock concentrations.

Implementing Sector Views

For our final portfolio construction example, let's assume we are entering a recessionary environment. An equity portfolio manager forecasts that the consumer staples sector will outperform the consumer discretionary sector in the near future, so she wants to create a portfolio to capitalize on this view. One simple idea would be to take a long position in the consumer staples sector (NCY: noncyclical) and a short position in the consumer discretionary sector (CYC: cyclical) by using, for example, sector ETFs. Similar to the previous example, this could result in exposures to risk factors other than the industry factors. Table 4 illustrates the exposure of this long-short portfolio to the risk factors in the POINT[®] U.S. equity risk model. As we can see in the table, the portfolio has significant net exposures to certain fundamental and technical factors, such as share turnover.

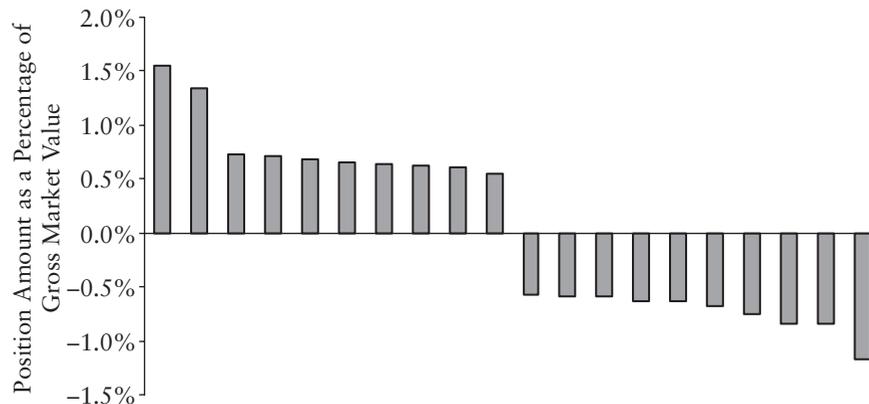


Figure 2 Largest 10 Positions on Long and Short Sides for the Momentum Portfolio

Table 4 Factor Exposures and Contributions for Consumer Staples vs. Consumer Discretionary Portfolio (bps/month)

Factor Name	Sensitivity/ Exposure	Net Exposure	Factor Volatility	Contribution to TEV
CURRENCY				
USD (U.S. dollar)	Market weight (%)	0.00	0	0.0
EQUITIES DEVELOPED MARKETS				
U.S. Equity CYC Automobiles	Empirical beta	-0.069	1,086	60.3
U.S. Equity CYC Consumer Durables	Empirical beta	-0.093	822	59.1
U.S. Equity CYC Consumer Services	Empirical beta	-0.140	690	71.1
U.S. Equity CYC Media	Empirical beta	-0.292	759	172.8
U.S. Equity CYC Retailing	Empirical beta	-0.317	745	185.1
U.S. Equity NCY Retailing	Empirical beta	0.226	404	-44.9
U.S. Equity NCY Food	Empirical beta	0.546	418	-96.4
U.S. Equity NCY Household	Empirical beta	0.236	415	-55.3
U.S. Equity Total Yield	Total yield	0.269	36	-3.7
U.S. Equity Corporate Default Probability	CDP	-0.201	76	9.0
U.S. Equity Share Turnover Rate	Share turnover	-0.668	59	-10.3
U.S. Equity Momentum (9m)	Momentum	-0.144	73	-5.6
U.S. Equity Discretionary Accruals	Accruals	-0.020	31	-0.2
U.S. Equity Market Value	Size	0.193	111	1.7
U.S. Equity Realized Volatility	Realized volatility	-0.619	97	5.5
U.S. Equity Earnings to Price	Earnings-Price	0.024	44	0.0
U.S. Equity Book to Price	Book-Price	-0.253	40	5.8
U.S. Equity Earnings Forecast	Earnings forecast	0.038	67	-0.4

Suppose the portfolio manager decides to limit exposures to fundamental and technical factors. We can again use the optimizer to construct a long-short portfolio, with an exposure (beta) of 1 to the consumer staples sector and a beta of -1 to the consumer discretionary sector. To limit the exposure to fundamental and technical risk factors, we further impose the exposure to each of these factors to be between -0.2 and 0.2.⁹ We also restrict the portfolio to be dollar neutral, and allow for only long positions in the consumer staples stocks and for only short positions in consumer discretionary stocks. Finally, we restrict the investment universe to the members of the S&P 500 index.¹⁰

The resulting portfolio consists of 69 securities (approximately half of discretionary and staples stocks in S&P 500) with 31 long positions in the consumer staples stocks and 38 short positions in consumer discretionary stocks. Table 5 depicts the factor exposures for this portfolio. As we can see in the table, the sum

of the exposures to the industry factors is 1 for the consumer staples stocks and -1 for the consumer discretionary stocks. Exposures to fundamental and technical factors are generally significantly smaller when compared to the previous table, limiting the adverse effects of potential moves in these factors. Interestingly, no stocks from the automobiles industry are selected in the optimal portfolio, potentially due to excessive idiosyncratic risk of firms in that particular industry. The contribution to volatility from the cyclical sector is higher than that from the non-cyclical sector, due to higher volatility of industry factors in the former.

The bounds used for the fundamental and technical factor exposures in the portfolio construction process were set to force a reduction in the exposure to these factors. However, there is a trade-off between having smaller exposures and having smaller idiosyncratic risk in the final portfolio. The resolution of this trade-off depends on the preferences of the portfolio

Table 5 Factor Exposures and Contributions for the Optimal Sector View Portfolio (bps/month)

Factor Name	Sensitivity/ Exposure	Net Exposure	Factor Volatility	Contribution to TEV
EQUITIES DEVELOPED MARKETS				
U.S. Equity CYC Consumer Durables	Empirical beta	-0.118	822	80.6
U.S. Equity CYC Consumer Services	Empirical beta	-0.222	690	124.9
U.S. Equity CYC Media	Empirical beta	-0.242	759	151.0
U.S. Equity CYC Retailing	Empirical beta	-0.417	745	264.2
U.S. Equity NCY Retailing	Empirical beta	0.287	404	-67.9
U.S. Equity NCY Food	Empirical beta	0.497	418	-112.6
U.S. Equity NCY Household	Empirical beta	0.216	415	-56.2
U.S. Equity Total Yield	Total yield	-0.059	36	0.8
U.S. Equity Corporate Default Probability	CDP	-0.042	76	1.7
U.S. Equity Share Turnover Rate	Share turnover	-0.196	59	-3.9
U.S. Equity Momentum (9m)	Momentum	-0.138	73	-4.7
U.S. Equity Discretionary Accruals	Accruals	-0.014	31	-0.1
U.S. Equity Market Value	Size	-0.011	111	-0.1
U.S. Equity Realized Volatility	Realized volatility	-0.199	97	-0.1
U.S. Equity Earnings to Price	Earnings-Price	0.027	44	0.0
U.S. Equity Book to Price	Book-Price	-0.070	40	1.4
U.S. Equity Earnings Forecast	Earnings forecast	0.085	67	-1.1

manager. When the bounds are more restrictive, we are also decreasing the feasible set of solutions available to the problem and therefore potentially achieving a higher idiosyncratic risk (remember that the objective is the minimization of idiosyncratic risk). In our example, the idiosyncratic TEV of the portfolio increases from 119 bps/month (before the optimization) to 158 bps/month on the optimized portfolio. This change is the price paid for the ability to limit certain systematic risk factor exposures.

Analyzing Portfolio Risk Using Multifactor Models

Now that we have seen examples of using multifactor equity models for portfolio construction and briefly discussed their risk outcomes, we take a more in-depth look at portfolio risk. Risk analysis based on multifactor models can take many forms, from a relatively high-level aggregate approach to an in-depth analysis of the risk properties of individual stocks and groups of stocks. Multifactor equity risk models provide

the tools to perform the analysis of portfolio risk in many different dimensions, including exposures to risk factors, security factor contributions to total risk, analysis at the ticker level, and scenario analysis. In this section, we provide an overview of such detailed analysis using the S&P 500 index tracker example we created in the previous section.

Recall from Table 2 that the TEV of the optimized S&P 500 tracking portfolio was 39.6 bps/month, composed mostly of idiosyncratic risk (35.8 bps/month) and a relatively small amount of systematic risk (16.9 bps/month). To analyze further the source of these numbers, we first compare the holdings of the portfolio with those of the benchmark and then study the impact of the mismatch to the risk of the net position (Portfolio-Benchmark). The first column in Table 6 shows the net market weights (NMW) of the portfolio at the sector level (GICS level 1). Our portfolio appears to be well balanced with respect to the benchmark from a net market weight perspective. The largest market value discrepancies are an overweight in information

Table 6 Net Market Weights and Risk Contributions by Sector (bps/month)

	Net Market Weight (%)	Contribution to TEV (CTEV)		
		Systematic	Idiosyncratic	Total
Total	0.0	7.2	32.7	39.8
Energy	1.4	1.3	4.4	5.7
Materials	-2.1	1.0	1.3	2.3
Industrials	2.1	0.3	3.8	4.1
Consumer discretionary	-3.6	1.7	4.7	6.3
Consumer staples	-0.5	0.5	2.2	2.6
Health care	-3.3	1.3	2.2	3.4
Financials	0.1	0.6	7.0	7.5
Information tech	5.2	0.6	5.5	6.1
Telecom services	2.4	0.2	0.8	1.0
Utilities	-1.7	-0.1	0.9	0.8

technology (+5.2%) and an underweight in consumer discretionary (-3.6%) and health care (-3.3%) companies. However, the sector with the largest contribution to overall risk (contribution to TEV, or CTEV) is financials (7.5 bps/month). This may seem unexpected, given the small NMW of this sector (0.1%). This result is explained by the fact that contributions to risk (CTEV) are dependent on the net market weight of an exposure, its risk and also the correlation between the different exposures. Looking into the decomposition of the CTEV, the table also shows that most of the total contribution to risk from financials is idiosyncratic (7.0 bps/month). This result is due to the

small number of securities our portfolio has in this sector and the underlying high volatility of these stocks. In short, the diversification benefits across financial stocks are small in our portfolio: We could potentially significantly reduce total risk by constructing our financials exposure using more names. Note that this analysis is only possible with a risk model.

Table 7 highlights additional risk measures by sector. What we see in the first column is the isolated TEV, that is, the risk associated with the stocks in that particular sector only. On an isolated basis, the information technology sector has the highest risk in the portfolio. This top position in terms of isolated risk does not

Table 7 Additional Risk Measures by Sector

	Isolated TEV (bps/month)	Liquidation Effect on TEV (bps/month)	TEV Elasticity (x100) (bps)	Systematic Beta (bps)
Total	39.64	-39.64	100.00	1.00
Energy	13.94	-3.38	14.25	0.89
Materials	16.94	1.29	5.74	1.25
Industrials	20.99	1.41	10.27	1.20
Consumer discretionary	29.34	4.25	15.89	1.11
Consumer staples	10.70	-1.20	6.59	0.70
Health care	17.37	0.37	8.56	0.65
Financials	20.77	-2.19	18.93	1.34
Information tech	31.90	6.20	15.30	0.99
Telecom services	11.58	0.67	2.53	0.76
Utilities	10.30	0.56	1.93	0.79

translate into the highest contribution to overall portfolio risk, as we saw in Table 6. The discrepancy between isolated risk numbers and contributions to risk is explained by the correlation between the exposures and allows us to understand the potential hedging effects present across our portfolio. The liquidation effect reported in the table represents the change in TEV when we completely hedge that particular position, that is, enforce zero net exposure to any stock in that particular sector. Interestingly, eliminating our exposure to information technology stocks would actually increase our overall portfolio risk by 6.2 bps/month. This happens because the overweight in this sector is effectively hedging out risk contributions from other sectors. If we eliminate this exposure, the portfolio balance is compromised. The TEV elasticity reported gives an additional perspective regarding how the TEV in the portfolio changes when we change the exposure to that sector. Specifically, it tells us the percentage change in TEV for each 1% change in our exposure to that particular sector. For example, if we double our exposure to the energy sector, our TEV would increase by 14.25% (from 39.6 bps/month to 45.2 bps/month). Finally, the report estimates the portfolio to have a beta of 1.00 to the benchmark, which is, of course, in line with our index tracking objective. The beta statistic measures the comovement between the systematic risk drivers of the portfolio and the benchmark and should be interpreted only as that. In particular, a low portfolio beta (relative to the benchmark) does not imply low portfolio risk. It signals relatively low systematic co-movement between the two universes or a relatively high idiosyncratic risk for the portfolio. For example, if the sources of systematic risk from the portfolio and the benchmark are distinct, the portfolio beta is close to zero. The report also provides the systematic beta associated with each sector. For instance, we see that a movement of 1% in the benchmark leads to a 1.34% return in the financials component of

our portfolio. As expected, consumer staples and health care are two low beta industries, as they tend to be more stable through the business cycle.¹¹

Although important, the information we examined so far is still quite aggregated. For instance, we know from Table 6 that a large component of idiosyncratic risk comes from financials. But what names are contributing most? What are the most volatile sectors? How are systematic exposures distributed within each sector? Risk models should be able to provide answers to all these questions, allowing for a detailed view of the portfolio's risk exposures and contributions. As an example, Table 8 displays all systematic risk factors the portfolio or the benchmark loads onto. It also provides the portfolio, benchmark, and net exposures for each risk factor, the volatility of each of these factors, and their contributions to total TEV. The table shows that the net exposures to the risk factors are generally low, meaning that the tracking portfolio has small active exposures. This finding is in line with the evidence from Table 2, where we see that the systematic risk is small (16.9 bps/month). If we look into the contributions of individual factors to total TEV, the table shows that the top contributors are the size, share turnover, and realized volatility factors. The optimal index tracking portfolio tends to be composed of very large-cap names within the specified universe, and that explains the net positive loading to the market value (size) factor. This portfolio tilt is due to the generally low idiosyncratic risk large companies have. This is seen favorably by the optimization engine, as it tries to minimize idiosyncratic risk. This same tilt would explain our net exposure to both the share turnover and realized volatility factors, as larger companies tend to have lower realized volatility and share turnover too. Interestingly, industry factors have relatively small contributions to TEV, even though they exhibit significantly higher volatilities. This results from the fact that the optimization

Table 8 Factor Exposures and Contributions for the Tracking Portfolio vs. S&P 500 (bps/month)

Factor Name	Sensitivity/ Exposure	Portfolio Exposure	Benchmark Exposure	Net Exposure	Factor Volatility	Contribution to TEV
CURRENCY						
USD (U.S. dollar)	Market weight (%)	100.00	100.00	0.00	0	0.00
EQUITIES DEVELOPED MARKETS						
U.S. Equity Energy	Empirical beta	0.10	0.10	0.00	651	0.17
U.S. Equity Materials	Empirical beta	0.01	0.03	-0.02	808	0.59
U.S. Equity IND Capital Goods	Empirical beta	0.11	0.08	0.03	723	-0.66
U.S. Equity IND Commercial	Empirical beta	0.00	0.01	-0.01	640	0.13
U.S. Equity IND Transportation	Empirical beta	0.01	0.02	-0.01	739	0.12
U.S. Equity CYC Automobiles	Empirical beta	0.00	0.01	-0.01	1,086	0.35
U.S. Equity CYC Consumer Durables	Empirical beta	0.01	0.01	0.00	822	-0.09
U.S. Equity CYC Consumer Services	Empirical beta	0.00	0.01	-0.01	690	0.47
U.S. Equity CYC Media	Empirical beta	0.02	0.03	-0.01	759	0.42
U.S. Equity CYC Retailing	Empirical beta	0.03	0.03	-0.01	745	0.28
U.S. Equity NCY Retailing	Empirical beta	0.02	0.03	-0.01	404	0.11
U.S. Equity NCY Food	Empirical beta	0.04	0.06	-0.02	418	0.32
U.S. Equity NCY Household	Empirical beta	0.05	0.03	0.02	415	-0.25
U.S. Equity HLT Health Care	Empirical beta	0.02	0.04	-0.02	518	0.82
U.S. Equity HLT Pharmaceuticals	Empirical beta	0.06	0.07	-0.02	386	0.31
U.S. Equity FIN Banks	Empirical beta	0.03	0.03	0.00	900	-0.01
U.S. Equity FIN Diversified Financials	Empirical beta	0.08	0.08	0.01	839	-0.16
U.S. Equity FIN Insurance	Empirical beta	0.02	0.04	-0.02	712	0.54
U.S. Equity FIN Real Estate	Empirical beta	0.03	0.01	0.02	956	-0.19
U.S. Equity TEC Software	Empirical beta	0.10	0.09	0.02	577	-0.61
U.S. Equity TEC Hardware	Empirical beta	0.08	0.07	0.00	645	-0.08
U.S. Equity TEC Semiconductors	Empirical beta	0.05	0.02	0.02	809	-0.49
U.S. Equity Telecommunication	Empirical beta	0.05	0.03	0.02	458	-0.37
U.S. Equity Utilities	Empirical beta	0.02	0.04	-0.01	554	0.17
U.S. Equity Total Yield	Total yield	0.12	0.05	0.07	36	0.04
U.S. Equity Corporate Default Probability	GDP	-0.15	-0.07	-0.08	76	0.22
U.S. Equity Share Turnover Rate	Share turnover	-0.20	0.01	-0.21	59	1.28
U.S. Equity Momentum (9m)	Momentum	-0.02	-0.03	0.01	73	0.02
U.S. Equity Discretionary Accruals	Accruals	-0.03	0.02	-0.05	31	-0.14
U.S. Equity Market Value	Size	0.29	0.20	0.09	111	1.16
U.S. Equity Realized Volatility	Realized volatility	-0.21	-0.08	-0.13	97	2.38
U.S. Equity Earnings to Price	Earnings-Price	0.09	0.04	0.05	44	0.19
U.S. Equity Book to Price	Book-Price	-0.06	-0.03	-0.03	40	0.02
U.S. Equity Earnings Forecast	Earnings forecast	0.08	0.05	0.03	67	0.12
U.S. Equity Other Market Volatility	Market weight	1.00	1.00	0.00	17	0.00

engine specifically targets these factors because of their high volatility and is successful in minimizing net exposure to industry factors in the final portfolio.

Finally, remember from Table 2 that the largest component of the portfolio risk comes from name-specific exposures. Therefore, it is important to be aware of which individual stocks in our portfolio contribute the most to overall risk. Table 9 shows the set of stocks in our portfolio with the largest idiosyncratic risk. The portfolio manager can use this information as a screening device to filter out undesirable positions with high idiosyncratic risk and to make sure her views on individual firms translate into risk as expected. In particular, the list in the table should only include names about which the portfolio manager has strong views, either positive—expressed with positive NMW—or negative—in which case we would expect a short net position.

It should be clear from the above examples that although the factors used to measure risk are predetermined in a linear factor model, there is a large amount of flexibility on the way the risk numbers can be aggregated and reported. Instead of sectors, we could have grouped risk by any other classification of individual stocks, for example, by regions or market capitalization. This allows the risk to be reported using the same investment philosophy underlying the portfolio construction process¹² regardless of the underlying factor model.

There are also many other risk analytics available, not mentioned in this example, that give additional detail about specific risk properties of the portfolio and the constituents. We have only discussed total, systematic, and idiosyncratic risk (which can be decomposed into risk contributions on a flexible basis), and referred to isolated and liquidation TEV, TEV elasticity, and portfolio beta. Most users of multifactor risk models will find their own preferred approach to risk analysis through experience.

Performance Attribution

Now that we discussed portfolio construction and risk analysis as the first steps of the investment process, we give a brief overview of performance attribution, an ex post analysis of performance typically conducted at the end of the investment horizon. Performance attribution analysis provides an evaluation of the portfolio manager's performance with respect to various decisions made throughout the investment process. The underperformance or outperformance of the portfolio manager when compared to the benchmark can be due to different reasons, including effective sector allocation, security selection, or tilting the portfolio toward certain risk factors. Attribution analysis aims to unravel the major sources of this performance differential. The exercise allows the portfolio manager to understand how her particular views—translated into net

Table 9 Individual Securities and Idiosyncratic Risk Exposures

Company Name	Portfolio Weight (%)	Benchmark Weight (%)	Net Weight (%)	Idiosyncratic TEV (bps/month)
Vornado Realty Trust	2.80	0.13	2.67	7.42
Kohls Corp	1.41	0.15	1.26	6.58
Bank of America Corp	2.71	1.41	1.29	6.16
Conocophillips	2.29	0.82	1.47	6.03
Roper Industries Inc	1.62	0.06	1.56	5.98
Walt Disney Co	2.26	0.66	1.60	5.48
Honeywell International Inc.	2.58	0.33	2.25	5.48
Cincinnati Financial Corp	1.88	0.05	1.83	5.35
Goldman Sachs	0.00	0.78	-0.78	5.25

exposures—performed during the period and reveals whether some of the portfolio's performance was the result of unintended bets.

There are three basic forms of attribution analysis used for equity portfolios. These are return decomposition, factor model-based attribution, and style analysis. In the return decomposition approach, the performance of the portfolio manager is generally attributed to top-down allocation (e.g., currency, country, or sector allocation) in a first step, followed by a bottom-up security selection performance analysis. This is a widely used technique among equity portfolio managers.

Factor model-based analysis attributes performance to exposures to risk factors such as industry, size, and financial ratios. It is relatively more complicated than the previous approach and is based on a particular risk model that needs to be well understood. For example, let's assume that a portfolio manager forecasts that value stocks will outperform growth stocks in the near future. As a result, the manager tilts the portfolio toward value stocks as compared to the benchmark, creating an active exposure to the value factor. In an attribution framework without systematic factors, such sources of performance cannot be identified and hence may be inadvertently attributed to other reasons. Factor model-based attribution analysis adds value by incorporating these factors (representing major investment themes) explicitly into the return decomposition process and by identifying additional sources of performance represented as active exposures to systematic risk factors.

Style analysis, on the other hand, is based on a regression of the portfolio return to a set of style benchmarks. It requires very little information (e.g., we do not need to know the contents of the portfolio), but the outcome depends significantly on the selection of style benchmarks. It also assumes constant loadings to these styles across the regression period, which may be unrealistic for managers with somewhat dynamic allocations.

Factor-Based Scenario Analysis

The last application we review goes over the use of equity risk factor models in the context of scenario analysis. Many investment professionals utilize scenario analysis in different shapes and forms for both risk and portfolio construction purposes. Factor-based scenario analysis is a tool that helps portfolio managers in their decision-making process by providing additional intuition on the behavior of their portfolio under a specified scenario. A scenario can be a historical episode, such as the equity market crash of 1987, the war in Iraq, or the 2008 credit crisis. Alternatively, scenarios can be defined as a collection of hypothetical views (e.g., user-defined scenarios) in a variety of forms such as a view on a given portfolio or index (e.g., S&P 500 drops by 20%) or a factor (e.g., U.S. equity-size factor moves by 3 standard deviations) or correlation between factors (e.g., increasing correlations across markets in episodes of flight to quality). In this section, we use the POINT[®] Factor-Based Scenario Analysis Tool to illustrate how we can utilize factor models to perform scenario analysis.

Before we start describing the example, let's take an overview of the mechanics of the model. It allows for the specification of user views on returns of portfolios, indexes, or risk factors. When the user specifies a view on a portfolio or index, this is translated into a view on risk factor realizations, through the linear factor model framework.¹³ These views are combined with ones that are directly specified in terms of risk factors. It is important to note that the portfolio manager does not need to specify views on all risk factors, and typically has views only on a small subset of them. Once the manager specifies this subset of original views, the next step is to expand these views to the whole set of factors. The scenario analysis engine achieves this by estimating the most likely realization of all other factors—given the factor realizations on which views are specified—using the risk model covariance matrix. Once all factor

Table 10 Index Returns under Scenario 1 (VIX jumps by 50%)

Universe	Type	Measure	Unit	Result
S&P 500	Equity index	Return	%	-7.97
FTSE U.K. 100	Equity index	Return	%	-9.34
DJ EURO STOXX 50	Equity index	Return	%	-11.63
NIKKEI 225	Equity index	Return	%	-4.99
MSCI-AC ASIA PACIFIC EX JAPAN	Equity index	Return	%	-10.33
MSCI-EMERGING MARKETS	Equity index	Return	%	-9.25

realizations are populated, the scenario outcome for any portfolio or index can be computed by multiplying their specific exposures to the risk factors by the factor realizations under the scenario. The tool provides a detailed analysis of the portfolio behavior under the specified scenario.

We illustrate this tool using two different scenarios: a 50% shift in the U.S. equity market volatility—represented by the VIX index—(scenario 1) and a 50% jump in the European credit spreads (scenario 2).¹⁴ We use a set of equity indexes from across the globe to illustrate the impact of these two scenarios. We run the scenarios as of July 30, 2010, which specifies the date both for the index loadings and the covariance matrix used. Base currency is set to U.S. dollars (USD) and hence index returns presented below are in USD.

Table 10 shows the returns of the chosen equity indexes under the first scenario. We see that all indexes experience significant negative returns with Euro Stoxx plummeting the most

and Nikkei experiencing the smallest drop. To understand these numbers better, let's look into the contributions of different factors to these index returns.

Table 11 illustrates return contributions for four of these equity indexes under scenario 1. Specifically, for each index, it decomposes the total scenario return into return coming from different factors each index has exposure to. In this example, all currency factors are defined with respect to USD. Moreover, equity factors are expressed in their corresponding local currencies and can be described as broad market factors for their respective regions.

Not surprisingly, Table 11 shows that the majority of the return contributions for selected indexes come from the reaction of equity market factors to the scenario. However, foreign exchange (FX) can also be a significant portion of total return for some indexes, such as in the case of the Euro Stoxx (-4.8%). Nikkei experiences a relatively smaller drop in USD terms, majorly due to a positive contribution coming from the

Table 11 Return Contributions for Equity Indexes under Scenario 1 (in %)

Group	Factor	S&P 500	FTSE U.K. 100	DJ EURO STOXX 50	NIKKEI 225
FX	GBP		-1.77		
FX	JPY				1.21
FX	EUR		-0.38	-4.80	
Equity	U.S. equity	-7.97			
Equity	U.K. equity		-6.67		
Equity	Japan equity				-6.20
Equity	EMG equity		-0.09		
Equity	Continental Europe equity		-0.43	-6.83	
Total		-7.97	-9.34	-11.63	-4.99

Table 12 Factor Returns and Z-Scores under Scenario 1

Group	Name	Measure	Unit	Value	Std. Dev.	Z-Score
Equity	U.K. equity	Return	%	-7.85	4.99	-1.57
Equity	U.S. equity	Return	%	-8.61	6.06	-1.42
Equity	Continental Europe equity	Return	%	-7.12	5.04	-1.41
Equity	Japan equity	Return	%	-5.96	4.73	-1.26
Equity	EMG equity	Return	%	-8.50	6.88	-1.24
FX	EUR	Return	%	-4.80	3.93	-1.22
FX	GBP	Return	%	-1.93	3.42	-0.56
FX	JPY	Return	%	1.21	3.39	0.36

JPY FX factor. This positive contribution is due to the safe haven nature of Japanese yen in case of flight to quality under increased risk aversion in global markets.

Table 12 demonstrates the scenario-implied factor realizations (“value”), factor volatilities, and the Z-scores for the risk factors given in Table 11. The Z-score of the factor quantifies the effect of the scenario on that specific factor. It is computed as

$$z = \frac{r}{\sigma_r}$$

where r is the return of the factor in the scenario and σ_r is the standard deviation of the factor. Hence, the Z-score measures how many standard deviations a factor moves in a given scenario. Table 12 lists the factors by increasing Z-score under scenario 1. The U.K. equity factor experiences the largest negative move, at -1.57 standard deviations. FX factors experience relatively smaller movements. JPY is the only factor with a positive realization due to the aforementioned characteristic of the currency.

In the second scenario, we shift European credit spreads by 50% (a 3.5-sigma event) and explore the effect of credit market swings on the equity markets. As we can see in Table 13, all equity indexes experience significant returns, in line with the severity of the scenario.¹⁵ The result also underpins the strong recent co-movement between the credit and equity markets. The exception is again the Nikkei that realizes a relatively smaller return.

Table 14 provides the return, volatility, and the Z-score of certain relevant factors under scenario 2. As expected, the major mover on the equity side is the continental Europe equity factor, followed by the United Kingdom. Given the recent strong correlations between equity and credit markets across the globe, the table suggests that a 3.5 standard deviation shift in the European spread factor results in a 2 to 3 standard deviation movement of global equity factors.

The two examples above illustrate the use of factor models in performing scenario analysis to achieve a clear understanding of how a portfolio may react under different circumstances.

Table 13 Index Returns under Scenario 2 (EUR Credit Spread Jumps by 50%)

Universe	Type	Measure	Unit	Result
S&P 500	Equity index	Return	%	-13.03
FTSE U.K. 100	Equity index	Return	%	-18.62
DJ EURO STOXX 50	Equity index	Return	%	-19.68
NIKKEI 225	Equity index	Return	%	-8.92
MSCI-AC ASIA PACIFIC EX JAPAN	Equity index	Return	%	-18.40
MSCI-EMERGING MARKETS	Equity index	Return	%	-16.83

Table 14 Factor Returns and Z-Scores under Scenario 2

Group	Name	Measure	Unit	Value	Std. Dev.	Z-Score
Equity	Continental Europe equity	Return	%	-14.02	5.04	-2.78
Equity	U.K. equity	Return	%	-13.05	4.99	-2.62
Equity	Japan equity	Return	%	-11.53	4.73	-2.44
Equity	U.S. equity	Return	%	-14.09	6.06	-2.33
Equity	EMG equity	Return	%	-15.93	6.88	-2.32
FX	GBP	Return	%	-6.54	3.42	-1.91
FX	EUR	Return	%	-6.23	3.93	-1.59
FX	JPY	Return	%	3.07	3.39	0.90

KEY POINTS

- Multifactor equity risk models provide detailed insight into the structure and properties of portfolios. These models characterize stock returns in terms of systematic factors and an idiosyncratic component. Systematic factors are generally designed to have intuitive economic interpretation and they represent common movements across securities. On the other hand, the idiosyncratic component represents the residual return due to stock-specific events.
- Systematic factors used in equity risk models can be broadly classified under five categories: market factors, classification variables, firm characteristics, macroeconomic variables, and statistical factors.
- Relative significance of systematic risk factors depends on various parameters such as the model horizon, region/country for which the model is designed, existence of other factors, and the particular time period of the analysis. For instance, in the presence of industry factors, macroeconomic factors tend to be insignificant for short to medium horizon equity risk models whereas they tend to be more significant for long-horizon models. Moreover, for developed equity markets, industry factors are typically more significant as compared to the country factors. The latter are still the dominant effect for emerging markets.
- Choice of the model and the estimation technique affect the interpretation of factors. For instance, in the existence of a market factor, industry factors represent industry-specific movements net of market. If there is no market factor, their interpretation is very close to market value-weighted industry indexes.
- Multifactor equity risk models can be classified according to how their loadings and factors are specified. The most common equity factor models specify loadings based on classification (e.g., industry) and fundamental or technical information, and estimate factor realizations every period. Certain other models take factors as known (e.g., returns on industry indexes) and estimate loadings based on time-series information. A third class of models is based purely on statistical approaches without concern for economic interpretation of factors and loadings. Finally, it is possible to combine these approaches and construct hybrid models. Each of these approaches has its own specific strengths and weaknesses.
- A good multifactor equity risk model provides detailed information regarding the exposures of a complex portfolio and can be a valuable tool for portfolio construction and risk management. It can help managers construct portfolios tracking a particular benchmark, express views subject to a given risk budget, and rebalance a portfolio

while avoiding excessive transaction costs. Further, by identifying the exposures where the portfolio has the highest risk sensitivity it can help a portfolio manager reduce (or increase) risk in the most effective way.

- Performance attribution based on multifactor equity risk models can give ex post insight into how the portfolio manager's views and corresponding investments translated into actual returns.
- Factor-based scenario analysis provides portfolio managers with a powerful tool to perform stress testing of portfolio positions and gain insight into the impact of specific market events on portfolio performance.

NOTES

1. The Barclays Global Risk Model is available through POINT[®], Barclays portfolio management tool. It is a multicurrency cross-asset model that covers many different asset classes across the fixed income and equity markets, including derivatives in these markets. At the heart of the model is a covariance matrix of risk factors. The model has more than 500 factors, many specific to a particular asset class. The asset class models are periodically reviewed. Structure is imposed to increase the robustness of the estimation of such a large covariance matrix. The model is estimated from historical data. It is calibrated using extensive security-level historical data and is updated on a monthly basis.
2. As an example, if the portfolio has 10 stocks, we need to estimate 45 parameters, with 100 stocks we would need to estimate 4,950 parameters.
3. This is especially the case over crisis periods where stock characteristics can change dramatically over very short periods of time.
4. Fixed income managers typically use cross-sectional type of models.
5. GICS is the Global Industry Classification Standard by Standard & Poor's, a widely used classification scheme by equity portfolio managers.
6. An application of macro variables in the context of risk factor models is as follows. First, we get the sensitivities of the portfolio to the model's risk factors. Then we project the risk factors into the macro variables. We then combine the results from these two steps to get the indirect loadings of the portfolio to the macro factors. Therefore, instead of calculating the portfolio sensitivities to macro factors by aggregating individual stock macro sensitivities—that are always hard to estimate—we work with the portfolio's macro loadings, estimated indirectly from the portfolio's risk factor loadings as described above. This indirect approach may lead to statistically more robust relationships between portfolio returns and macro variables.
7. The equity risk model suite in POINT consists of six separate models across the globe: the United States, United Kingdom, Continental Europe, Japan, Asia (excluding Japan), and global emerging markets equity risk models (for details see Silva, Staal, and Ural, 2009). It incorporates many unique features related to factor choice, industry and fundamental exposures, and risk prediction.
8. See Kumar (2010).
9. The setting of these exposures and its trade-offs are discussed later in this entry.
10. As POINT[®] U.S. equity risk model incorporates industry level factors, a unit exposure to a sector is implemented by restricting exposures to different industries within that sector to sum up to 1. Also, note that as before, the objective in the optimization problem is the minimization of idiosyncratic TEV to ensure that the resulting portfolio represents systematic—not idiosyncratic—effects.

11. Note that we can sum the sector betas into the portfolio beta, using portfolio sector weights (not net weights) as weights in the summation.
12. For a detailed methodology on how to perform this customized analysis, see Silva (2009).
13. Specifically, we can back out factor realizations from the portfolio or index returns by using their risk factor loadings.
14. For reference, as of July 30, 2010, scenario 1 would imply the VIX would move from 23.5 to 35.3 and scenario 2 would imply that the credit spread for the Barclays European Credit Index would change from 174 bps to 261 bps.
15. The same scenario results in a -8.12% move in the Barclays Euro Credit Index.

REFERENCES

- Kumar, A. (2010). The POINT optimizer. *Barclays Publication*, June.
- Silva, A. B. (2009). Risk attribution with custom-defined risk factors. *Barclays Publication*, August.
- Silva, A. B., Staal, A. D., and Ural, C. (2009). The US equity risk model. *Barclays Publication*, July.

Factor-Based Equity Portfolio Construction and Analysis

PETTER N. KOLM, PhD

Director of the Mathematics in Finance Masters Program and Clinical Associate Professor Courant Institute of Mathematical Sciences, New York University

JOSEPH A. CERNIGLIA

Visiting Researcher, Courant Institute of Mathematical Sciences, New York University

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: A factor is a common character among a group of assets. In the equities market, for example, it could be a particular financial ratio such as the price-earnings ratio or the book-price ratio. Factors fall into three categories—macroeconomic influences, cross-sectional characteristics, and statistical factors. Within asset management firms, factors and factor-forecasting models are used for a number of purposes. Those purposes could be central to managing portfolios. Within a trading strategy, for example, factors determine when to buy and sell securities. Factors are employed in other areas of financial theory, such as asset pricing, risk management, and performance attribution.

Common stock investment strategies can be broadly classified into the following categories: (1) *factor-based trading* strategies (also called *stock selection* or alpha models), (2) statistical arbitrage, (3) high-frequency strategies, and (4) event studies. Factors and factor-based models form the core of a major part of today's quantitative trading strategies. The focus of this entry is on developing trading strategies based on factors constructed from common (cross-sectional) characteristics of stocks. For this purpose, first we provide a definition of factors. We then examine the major sources of risk associated with trading strategies, and demonstrate how factors

are constructed from company characteristics and market data. The quality of the data used in this process is critical. We examine several data cleaning and adjustment techniques to account for problems occurring with backfilling and restatements of data, missing data, inconsistently reported data, as well as survivorship and look-ahead biases. In the last section of this entry, we discuss the analysis of the statistical properties of factors.

In a series of examples, we show the individual steps for developing a basic trading strategy. The purpose of these examples is not to provide yet another profitable trading strategy,

but rather to illustrate the process an analyst may follow when performing research. In fact, the factors that we use for this purpose are well known and have for years been exploited by industry practitioners. The value added of these examples is in the concrete illustration of the research and development process of a factor-based trading model.

FACTOR-BASED TRADING

Since the first version of the classic text on security analysis by Benjamin Graham and David Dodd¹—considered to be the Bible on the fundamental approach to security analysis—was first published in 1934, equity portfolio management and trading strategies have developed considerably. Graham and Dodd were early contributors to factor-based strategies because they extended traditional valuation approaches by using information throughout the financial

statements² and by presenting concrete rules of thumb to be used to determine the attractiveness of securities.³

Today's quantitative managers use factors as fundamental building blocks for trading strategies. Within a trading strategy, factors determine when to buy and sell securities. We define a *factor* as a common characteristic among a group of assets. In the equities market, it could be a particular financial ratio such as the price–earnings (P/E) or the book–price (B/P) ratios. Some of the most well-known factors and their underlying basic economic rationale references are provided in Table 1.

Most often this basic definition is expanded to include additional objectives. First, factors frequently are intended to capture some economic intuition. For instance, a factor may help understand the prices of assets by reference to their exposure to sources of macroeconomic risk, fundamental characteristics, or basic market behavior. Second, we should recognize that

Table 1 Summary of Well-Known Factors and Their Underlying Economic Rationale

Factor	Economic Rationale
Dividend yield	Investors prefer to immediately receive receipt of their investment returns.
Value	Investors prefer stocks with low valuations.
Size (market capitalization)	Smaller companies tend to outperform larger companies.
Asset turnover	This measure evaluates the productivity of assets employed by a firm. Investors believe higher turnover correlates with higher future return.
Earnings revisions	Positive analysts' revisions indicate stronger business prospects and earnings for a firm.
Growth of fiscal year 1 and fiscal year 2 earnings estimates	Investors are attracted to companies with growing earnings.
Momentum	Investors prefer stocks that have had good past performance.
Return reversal	Investors overreact to information, that is, stocks with the highest returns in the current month tend to earn lower returns the following month.
Idiosyncratic risk	Stocks with high idiosyncratic risk in the current month tend to have lower returns the following month.
Earnings surprises	Investors like positive earnings surprises and dislike negative earnings surprises.
Accounting accruals	Companies with earnings that have a large cash component tend to have higher future returns.
Corporate governance	Firms with better corporate governance tend to have higher firm value, higher profits, higher sales growth, lower capital expenditures, and fewer corporate acquisitions.
Executive compensation factors	Firms that align compensation with shareholders' interest tend to outperform.
Accounting risk factors	Companies with lower accounting risk tend to have higher future returns.

assets with similar factors (characteristics) tend to behave in similar ways. This attribute is critical to the success of a factor. Third, we would like our factor to be able to differentiate across different markets and samples. Fourth, we want our factor to be robust across different time periods.

Factors fall into three categories—macroeconomic influences, cross-sectional characteristics, and statistical factors. Macroeconomic influences are time series that measure observable economic activity. Examples include interest rate levels, gross domestic production, and industrial production. Cross-sectional characteristics are observable asset specifics or firm characteristics. Examples include dividend yield, book value, and volatility. Statistical factors are unobservable or latent factors common across a group of assets. These factors make no explicit assumptions about the asset characteristics that drive commonality in returns. Statistical factors are not derived using exogenous data but are extracted from other variables such as returns. These factors are calculated using various statistical techniques such as principal components analysis or factor analysis.

Within asset management firms, factors and forecasting models are used for a number of purposes. Those purposes could be central to managing portfolios. For example, a portfolio manager can directly send the model output to the trading desk to be executed. In other uses, models provide analytical support to analysts and portfolio management teams. For instance, models are used as a way to reduce the investable universe to a manageable number of securities so that a team of analysts can perform fundamental analysis on a smaller group of securities.

Factors are employed in other areas of financial theory, such as asset pricing, risk management, and performance attribution. In asset pricing, researchers use factors as proxies for common, undiversifiable sources of risk in the economy to understand the prices or values of

securities to uncertain payments. Examples include the dividend yield of the market or the yield spread between a long-term bond yield and a short-term bond yield.⁴ In risk management, risk managers use factors in risk models to explain and to decompose variability of returns from securities, while portfolio managers rely on risk models for covariance construction, portfolio construction, and risk measurement. In performance attribution, portfolio managers explain past portfolio returns based on the portfolio's exposure to various factors. Within these areas, the role of factors continues to expand. Recent research presents a methodology for attributing active return, tracking error, and the information ratio to a set of custom factors.⁵

The focus in this entry is on using factors to build equity forecasting models, also referred to as alpha or *stock selection models*. The models serve as mathematical representations of trading strategies. The mathematical representation uses future returns as dependent variables and factors as independent variables.

DEVELOPING FACTOR-BASED TRADING STRATEGIES

The development of a trading strategy has many similarities with an engineering project. We begin by designing a framework that is flexible enough so that the components can be easily modified, yet structured enough that we remain focused on our end goal of designing a profitable trading strategy.

Basic Framework and Building Blocks

The typical steps in the development of a trading strategy are:

- Defining a trading idea or investment strategy.
- Developing factors.

- Acquiring and processing data.
- Analyzing the factors.
- Building the strategy.
- Evaluating the strategy.
- Backtesting the strategy.
- Implementing the strategy.

In what follows, we take a closer look at each step.

Defining a Trading Idea or Investment Strategy

A successful trading strategy often starts as an idea based on sound economic intuition, market insight, or the discovery of an anomaly. Background research can be helpful in order to understand what others have tried or implemented in the past.

We distinguish between a trading idea and trading strategy based on the underlying economic motivation. A trading idea has a more short-term horizon often associated with an event or mispricing. A trading strategy has a longer horizon and is frequently based on the exploitation of a premium associated with an anomaly or a characteristic.

Developing Factors

Factors provide building blocks of the model used to build an investment strategy. We introduced a general definition of factors earlier in this entry. After having established the trading strategy, we move from the economic concepts to the construction of factors that may be able to capture our intuition. In this entry, we provide a number of examples of factors based on the cross-sectional characteristics of stocks.

Acquiring and Processing Data

A trading strategy relies on accurate and clean data to build factors. There are a number of third-party solutions and databases available for this purpose such as Thomson MarketQA,⁶

Factset Research Systems,⁷ and Compustat Xpressfeed.⁸

Analyzing the Factors

A variety of statistical and econometric techniques must be performed on the data to evaluate the empirical properties of factors. This empirical research is used to understand the risk and return potential of a factor. The analysis is the starting point for building a model of a trading strategy.

Building the Strategy

The model represents a mathematical specification of the trading strategy. There are two important considerations in this specification: the selection of which factors and how these factors are combined. Both considerations need to be motivated by the economic intuition behind the trading strategy. We advise against model specification being strictly data driven because that approach often results in overfitting the model and consequently overestimating forecasting quality of the model.

Evaluating, Backtesting, and Implementing the Strategy

The final step involves assessing the estimation, specification, and forecast quality of the model. This analysis includes examining the goodness of fit (often done in sample), forecasting ability (often done out of sample), and sensitivity and risk characteristics of the model.

RISK TO TRADING STRATEGIES

In investment management, risk is a primary concern. The majority of trading strategies are not risk free but rather subject to various risks. It is important to be familiar with the most

common risks in trading strategies. By understanding the risks in advance, we can structure our empirical research to identify how risks will affect our strategies. Also, we can develop techniques to avoid these risks in the model construction stage when building the strategy.

We describe the various risks that are common to factor trading strategies as well as other trading strategies such as risk arbitrage. Many of these risks have been categorized in the behavioral finance literature.⁹ The risks discussed include fundamental risk, noise trader risk, horizon risk, model risk, implementation risk, and liquidity risk.

Fundamental risk is the risk of suffering adverse fundamental news. For example, say our trading strategy focuses on purchasing stocks with high earnings-to-price ratios. Suppose that the model shows a pharmaceutical stock maintains a high score. After purchasing the stock, the company releases a news report that states it faces class-action litigation because one of its drugs has undocumented adverse side effects. While during this period other stocks with high earnings-to-price ratio may perform well, this particular pharmaceutical stock will perform poorly despite its attractive characteristic. We can minimize the exposure to fundamental risk within a trading strategy by diversifying across many companies. Fundamental risk may not always be company specific; sometimes this risk can be systemic. Some examples include the exogenous market shocks of the stock market crash in 1987, the Asian financial crisis in 1997, and the tech bubble in 2000. In these cases, diversification was not that helpful. Instead, portfolio managers that were sector or market neutral in general fared better.

Noise trader risk is the risk that a mispricing may worsen in the short run. The typical example includes companies that clearly are undervalued (and should therefore trade at a higher price). However, because noise traders may trade in the opposite direction, this mispricing can persist for a long time. Closely related to noise trader risk is *horizon risk*. The

idea here is that the premium or value takes too long to be realized, resulting in a realized return lower than a target rate of return.

Model risk, also referred to as *misspecification risk*, refers to the risk associated with making wrong modeling assumptions and decisions. This includes the choice of variables, methodology, and context the model operates in. There are different sources that may result in model misspecification and there are several remedies based on information theory, Bayesian methods, shrinkage, and random coefficient models.¹⁰

Implementation risk is another risk faced by investors implementing trading strategies. This risk category includes transaction costs and funding risk. Transaction costs such as commissions, bid-ask spreads, and market impact can adversely affect the results from a trading strategy. If the strategy involves shorting, other implementation costs arise such as the ability to locate securities to short and the costs to borrow the securities. *Funding risk* occurs when the portfolio manager is no longer able to get the funding necessary to implement a trading strategy. For example, many statistical arbitrage funds use leverage to increase the returns of their funds. If the amount of leverage is constrained, then the strategy will not earn attractive returns. Khandani and Lo (2007) confirm this example by showing that greater competition and reduced profitability of quantitative strategies today require more leverage to maintain the same level of expected return.

Liquidity risk is a concern for investors. Liquidity is defined as the ability to (1) trade quickly without significant price changes, and (2) trade large volumes without significant price changes. Cerniglia and Kolm (2009) discuss the effects of liquidity risk during the "quant crisis" in August 2007. They show how the rapid liquidation of quantitative funds affected the trading characteristics and price impact of trading individual securities as well as various factor-based trading strategies.

These risks can detract or contribute to the success of a trading strategy. It is obvious how these risks can detract from a strategy. What is not always clear is when any one of these unintentional risks contributes to a strategy. That is, sometimes when we build a trading strategy we take on a bias that is not obvious. If there is a premium associated with this unintended risk, then a strategy will earn additional return. Later the premium to this unintended risk may disappear. For example, a trading strategy that focuses on price momentum performed strongly in the calendar years of 1998 and 1999. What an investor might not notice is that during this period the portfolio became increasingly weighted toward technology stocks, particularly Internet-related stocks. During 2000, these stocks severely underperformed.

DESIRABLE PROPERTIES OF FACTORS

Factors should be founded on sound economic intuition, market insight, or an anomaly. In addition to the underlying economic reasoning, factors should have other properties that make them effective for forecasting.

It is an advantage if factors are intuitive to investors. Many investors will only invest in a particular fund if they understand and agree with the basic ideas behind the trading strategies. Factors give portfolio managers a tool in communicating to investors what themes they are investing in.

The search for the economic meaningful factors should avoid strictly relying on pure historical analysis. Factors used in a model should not emerge from a sequential process of evaluating successful factors while removing less favorable ones.

Most importantly, a group of factors should be parsimonious in its description of the trading strategy. This requires careful evaluation of the interaction between the different factors. For ex-

ample, highly correlated factors will cause the inferences made in a multivariate approach to be less reliable. Another possible problem when using multiple factors is the possibility of overfitting in the modeling process.

Any data set contains outliers, that is, observations that deviate from the average properties of the data. Outliers are not always trivial to handle and sometimes we may want to exclude them and other times not. For example, they could be erroneously reported or legitimate abnormal values. Later in this entry we discuss a few standard techniques to perform data cleaning. The success or failure of factors selected should not depend on a few outliers. In most cases, it is desirable to construct factors that are reasonably robust to outliers.

SOURCES FOR FACTORS

How do we find factors? The sources are widespread with no one source clearly dominating. Employing a variety of sources seems to provide the best opportunity to uncover factors that will be valuable for developing a new model.

There are a number of ways to develop factors based on economic foundations. It may start with thoughtful observation or study of how market participants act. For example, we may ask ourselves how other market participants will evaluate the prospects of the earnings or business of a firm. We may also want to consider what stock characteristics investors will reward in the future. Another common approach is to look for inefficiencies in the way that investors process information. For instance, research may discover that consensus expectations of earnings estimates are biased.

A good source for factors is the various reports released by the management of companies. Many reports contain valuable information and may provide additional context on how management interprets the company results and financial characteristics. For

example, quarterly earning reports (10-Qs) may highlight particular financial metrics relevant to the company and the competitive space they are operating within. Other company financial statements and SEC filings, such as the 10-K or 8-K, also provide a source of information to develop factors. It is often useful to look at the financial measures that management emphasize in their comments.

Factors can be found through discussions with market participants such as portfolio managers and traders. Factors are uncovered by understanding the heuristics experienced investors have used successfully. These heuristics can be translated into factors and models.

Wall Street analyst reports—also called sell-side reports or equity research reports—may contain valuable information. The reader is often not interested in the final conclusions, but rather in the methodology or metrics the analysts use to forecast the future performance of a company. It may also be useful to study the large quantity of books written by portfolio managers and traders that describe the process they use in stock selection.

Academic literature in finance, accounting, and economics provides evidence of numerous factors and trading strategies that earn abnormal returns. Not all strategies will earn abnormal profits when implemented by practitioners, for example, because of institutional constraints and transaction costs. Bushee and Raedy (2006) find that trading strategy returns are significantly decreased due to issues such as price pressure, restrictions against short sales, incentives to maintain an adequately diversified portfolio, and restrictions to hold no more than 5% ownership in a firm.

In uncovering factors, we should put economic intuition first and data analysis second. This avoids performing pure data mining or simply overfitting our models to past history. Research and innovation is the key to finding new factors. Today, analyzing and testing new factors and improving upon existing ones is itself a big industry.

BUILDING FACTORS FROM COMPANY CHARACTERISTICS

The following sections focus on the techniques for building factors from company characteristics. Often we desire our factors to relate the financial data provided by a company to metrics that investors use when making decisions about the attractiveness of a stock such as valuation ratios, operating efficiency ratios, profitability ratios, and solvency ratios. Factors should also relate to the market data such as analysts' forecasts, prices and returns, and trading volume.

WORKING WITH DATA

In this section, we discuss how to work with data and data quality issues, including some well-probed techniques used to improve the quality of the data. Though the role of getting and analyzing data can be mundane and tedious, we need not forget that high-quality data are critical to the success of a trading strategy. It is important to realize model output is only as good as the data used to calibrate it. As the saying goes: "Garbage in, garbage out."

Understanding the structure of financial data is important. We distinguish three different categories of financial data: time series, cross-sectional, and panel data. Time series data consist of information and variables collected over multiple time periods. Cross-sectional data consist of data collected at one point in time for many different companies (the cross-section of companies of interest). A panel data set consists of cross-sectional data collected at different points in time. We note that a panel data set may not be homogeneous. For instance, the cross-section of companies may change from one point in time to another.

Data Integrity

Quality data maintain several attributes such as providing a consistent view of history,

maintaining good data availability, containing no survivorship, and avoiding look-ahead bias. As all data sets have their limitations, it is important for the quantitative researcher to be able to recognize the limitations and adjust the data accordingly.

Data used in research should provide a consistent view of history. Two common problems that distort the consistency of financial data are backfilling and restatements of data. Backfilling of data happens when a company is first entered into a database at the current period and its historical data are also added. This process of backfilling data creates a selection bias because we now find historical data on this recently added company when previously it was not available. Restatements of data are prevalent in distorting consistency of data. For example, if a company revises its earnings per share numbers after the initial earnings release, then many database companies will overwrite the number originally recorded in the database with the newly released figure.

A frequent and common concern with financial databases is data availability. First, data items may only be available for a short period of time. For example, there were many years when stock options were granted to employees but the expense associated with the option grant was not required to be disclosed in financial statements. It was not until 2005 that accounting standards required companies to recognize directly stock options as an expense on the income statement. Second, data items may be available for only a subset of the cross-section of firms. Some firms, depending on the business they operate in, have research and development expenses while others do not. For example, many pharmaceutical companies have research and development expenses while utilities companies do not. A third issue is that a data item may simply not be available because it was not recorded at certain points in time. Sometimes this happens for just a few observations, other times it is the case for the whole time-series for a specific data item for a company. Fourth, dif-

ferent data items are sometimes combined. For example, sometimes depreciation and amortization expenses are not a separate line item on an income statement. Instead it is included in cost of goods sold. Fifth, certain data items are only available at certain periodicities. For instance, some companies provide more detailed financial reports quarterly while others report more details annually. Sixth, data items may be inconsistently reported across different companies, sectors, or industries. This may happen as the financial data provider translates financial measures from company reports to the specific database items (incomplete mapping), thereby ignoring or not correctly making the right adjustments.

For these issues some databases provide specific codes to identify the causes of missing data. It is important to have procedures in place that can distinguish among the different reasons for the missing data and be able to make adjustments and corrections.

Two other common problems with databases are survivorship and look-ahead bias. *Survivorship bias* occurs when companies are removed from the database when they no longer exist. For example, companies can be removed because of a merger or bankruptcy. This bias skews the results because only successful firms are included in the entire sample. Look-ahead bias occurs when data are used in a study that would not have been available during the actual period analyzed. For example, the use of year-end earnings data immediately at the end of the reporting period is incorrect because the data are not released by the firm until several days or weeks after the end of the reporting period.

Data alignment is another concern when working with multiple databases. Many databases have different identifiers used to identify a firm. Some databases have vendor specific identifiers, others have common identifiers such as CUSIPs or ticker symbols. Unfortunately, CUSIPs and ticker symbols change over time and are often reused. This practice makes

it difficult to link an individual security across multiple databases across time.

Example: The EBITDA/EV Factor

This example illustrates how the nuances of data handling can influence the results of a particular study. We use data from the Compustat Point-In-Time database and calculate the EBITDA/EV factor.¹¹ This factor is defined as earnings before interest, taxes, depreciation, and amortization divided by enterprise value (EBITDA/EV). Our universe of stocks is the Russell 1000 from December 1989 to December 2008, excluding financial companies. We calculate EBITDA /EV by two equivalent but different approaches. Each approach differs by the data items used in calculating the numerator (EBITDA):

1. EBITDA = Sales (Compustat data item 2) – Cost of goods sold (Compustat data item 30) – Selling and general administrative expenses (Compustat data item 1).
2. EBITDA = Operating income before depreciation (Compustat data item 21).

According to the Compustat manual, the following identity holds:

$$\begin{aligned} &\text{Operating income before depreciation} \\ &= \text{Sales} - \text{Cost of goods sold} - \text{Selling} \\ &\quad \text{and general administrative expenses} \end{aligned}$$

However, while this mathematical identity is true, this is not what we discover in the data. After we calculate the two factors, we form quintile portfolios of each factor and compare the individual holding rankings between the portfolio. Figure 1 displays the percentage differences in rankings for individual companies between the two portfolios. We observe that the results are not identical. As a matter of fact, there are large differences, particularly in the early period. In other words, the two mathematically equivalent approaches do not deliver the same empirical results.

Potential Biases from Data

There are numerous potential biases that may arise from data quality issues. It is important to recognize the direct effects of these data issues

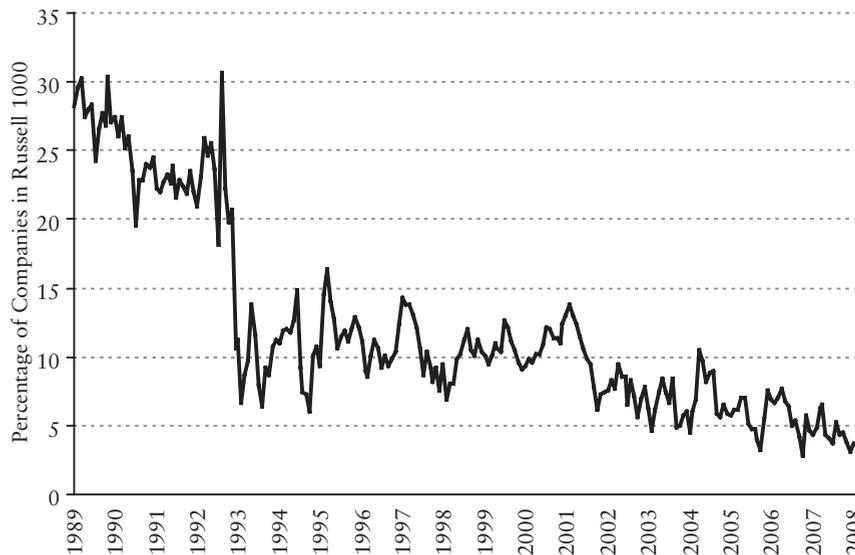


Figure 1 Percentage of Companies in Russell 1000 with Different Ranking According to the EBITDA/EV Factor

are not apparent a priori. We emphasize three important effects:¹²

1. *Effect on average stock characteristics.* When calculating cross-sectional averages of various metrics such as book-to-price or market capitalization, data issues can skew statistics and lead to incorrect inference about the population characteristics used in the study.
2. *Effect on portfolio returns.* The portfolio return implications of data issues are not always clear. For example, survivor bias results in firms being removed from the sample. Typically firms are removed from the sample for one of two reasons—mergers and acquisitions or failure. In most cases firms are acquired at a premium from the prevailing stock price. Leaving these firms out of the sample would have a downward bias on returns. In cases where companies fail, the stock price falls dramatically and removing these firms from the sample will have an upward bias on returns.
3. *Effects on estimated moments of returns.* A study by Kothari, Sabino, and Zach (2005) found that nonsurviving firms tend to be either extremely bad or extremely good performers. Survivor bias implies truncation of such extreme observations. The authors of the study show that even a small degree of such non-random truncation can have a strong impact on the sample moments of stock returns.

Dealing with Common Data Issues

Most data sets are subject to some quality issues. To work effectively, we need to be familiar with data definitions and database design. We also need to use processes to reduce the potential impact of data problems as they could cause incorrect conclusions.

The first step is to become familiar with the data standardization process vendors use to collect and process data. For example, many vendors use different templates to store data.

Specifically, the Compustat US database has one template for reporting income statement data, while the Worldscope Global database has four different templates depending on whether a firm is classified as a bank, insurance company, industrial company, or other financial company. Other questions related to standardization a user should be familiar with include:

- What are the sources of the data—publicly available financial statements, regulatory filings, newswire services, or other sources?
- Is there a uniform reporting template?
- What is the delay between publication of information and its availability in the database?
- Is the data adjusted for stock splits?
- Is history available for extinct or inactive companies?
- How is data handled for companies with multiple share classes?
- What is the process used to aggregate the data?

Understanding of the accounting principles underlying the data is critical. Here, two principles of importance are the valuation methodology and data disclosure or presentation. For the valuation, we should understand the type of cost basis used for the various accounting items. Specifically, are assets calculated using historical cost basis, fair value accounting, or another type? For accounting principles regarding disclosure and presentation, we need to know the definition of accounting terms, the format of the accounts, and the depth of detail provided.

Researchers creating factors that use financial statements should review the history of the underlying accounting principles. For example, the cash flow statement reported by companies has changed over the years. Effective for fiscal years ending July 15, 1988, Statement of Financial Accounting Standards No. 85 (SFAS No. 85) requires companies to report the Statement of Cash Flows. Prior to the adoption of that accounting standard, companies could report one

of three statements: Working Capital Statement, Cash Statement by Source and Use of Funds, or Cash Statement by Activity. Historical analysis of any factor that uses cash flow items will require adjustments to the definition of the factor to account for the different statements used by companies.

Preferably, automated processes should be used to reduce the potential impact of data problems. We start by checking the data for consistency and accuracy. We can perform time series analysis on individual factors looking at outliers and for missing data. We can use magnitude tests to compare current data items with the same items for prior periods, looking for data that are larger than a predetermined variance. When suspicious cases are identified, the cause of the error should be researched and any necessary changes made.

Methods to Adjust Factors

At first, factors consist of raw data from a database combined in an economically meaningful way. After the initial setup, a factor may be adjusted using analytical or statistical techniques to be more useful for modeling. The following three adjustments are common.

Standardization

Standardization rescales a variable while preserving its order. Typically, we choose the standardized variable to have a mean of zero and a standard deviation of one by using the transformation

$$x_i^{\text{new}} = \frac{x_i - \bar{x}}{\sigma_x}$$

where x_i is the stock's factor score, \bar{x} is the universe average, and σ_x is the universe standard deviation. There are several reasons to scale a variable in this way. First, it allows one to determine a stock's position relative to the universe average. Second, it allows better comparison across a set of factors since means and standard

deviations are the same. Third, it can be useful in combining multiple variables.

Orthogonalization

Sometimes the performance of our factor might be related to another factor. Orthogonalizing a factor for other specified factor(s) removes this relationship. We can orthogonalize by using averages or running regressions.

To orthogonalize the factor using averages according to industries or sectors, we can proceed as follows. First, for each industry we calculate the industry scores

$$s_k = \frac{\sum_{i=1}^n x_i \cdot \text{ind}_{i,k}}{\sum_{i=1}^n \text{ind}_{i,k}}$$

where x_i is a factor and $\text{ind}_{i,k}$ represent the weight of stock i in industry k . Next, we subtract the industry average of the industry scores, s_k , from each stock. We compute

$$x_i^{\text{new}} = x_i - \sum_{k \in \text{Industries}} \text{ind}_{i,k} \cdot s_k$$

where x_i^{new} is the new industry neutral factor.

We can use linear regression to orthogonalize a factor. We first determine the coefficients in the equation

$$x_i = a + b \cdot f_i + \varepsilon_i$$

where f_i is the factor to orthogonalize the factor x_i by, b is the contribution of f_i to x_i , and ε_i is the component of the factor x_i not related to f_i . ε_i is orthogonal to f_i (that is, ε_i is independent of f_i) and represents the neutralized factor

$$x_i^{\text{new}} = \varepsilon_i$$

In the same fashion, we can orthogonalize our variable relative to a set of factors by using the multivariate linear regression

$$x_i = a + \sum_j b_j \cdot f_j + \varepsilon_i$$

and then setting $x_i^{\text{new}} = \varepsilon_i$.

Often portfolio managers use a risk model to forecast risk and an alpha model to forecast returns. The interaction between factors in a risk model and an alpha model often concerns portfolio managers. One possible approach to address this concern is to orthogonalize the factors or final scores from the alpha model against the factors used in the risk model. Later in the entry, we discuss this issue in more detail.

Transformation

It is common practice to apply transformations to data used in statistical and econometric models. In particular, factors are often transformed such that the resulting series is symmetric or close to being normally distributed. Frequently used transformations include natural logarithms, exponentials, and square roots. For example, a factor such as market capitalization has a large skew because a sample of large-cap stocks typically includes mega-capitalization stocks. To reduce the influence of mega-capitalization companies, we may instead use the natural logarithm of market capitalization in a linear regression model.

Outlier Detection and Management

Outliers are observations that seem to be inconsistent with the other values in a data set. Financial data contain outliers for a number of reasons including data errors, measurement errors, or unusual events. Interpretation of data containing outliers may therefore be misleading. For example, our estimates could be biased or distorted, resulting in incorrect conclusions.

Outliers can be detected by several methods. Graphs such as boxplots, scatter plots, or histograms can be useful to visually identify them. Alternatively, there are a number of numerical techniques available. One common method is to compute the interquartile-range and then identify outliers as those values that are some multiple of the range. The interquartile-range is a

measure of dispersion and is calculated as the difference between the third and first quartiles of a sample. This measure represents the middle 50% of the data, removing the influence of outliers.

After outliers have been identified, we need to reduce their influence in our analysis. Trimming and winsorization are common procedures for this purpose. Trimming discards extreme values in the data set. This transformation requires the researcher to determine the direction (symmetric or asymmetric) and the amount of trimming to occur.

Winsorization is the process of transforming extreme values in the data. First, we calculate percentiles of the data. Next we define outliers by referencing a certain percentile ranking. For example, any data observation that is greater than the 97.5 percentile or less than the 2.5 percentile could be considered an outlier. Finally, we set all values greater or less than the reference percentile ranking to particular values. In our example, we may set all values greater than the 97.5 percentile to the 97.5 percentile value and all values less than 2.5 percentile set to the 2.5 percentile value. It is important to fully investigate the practical consequences of using either one of these procedures.

ANALYSIS OF FACTOR DATA

After constructing factors for all securities in the investable universe, each factor is analyzed individually. Presenting the time-series and cross-sectional averages of the mean, standard deviations, and key percentiles of the distribution provide useful information for understanding the behavior of the chosen factors.

Although we often rely on techniques that assume the underlying data generating process is normally distributed, or at least approximately, most financial data is not. The underlying data generating processes that embody aggregate investor behavior and characterize the financial markets are unknown and exhibit significant

uncertainty. Investor behavior is uncertain because not all investors make rational decisions or have the same goals. Analyzing the properties of data may help us better understand how uncertainty affects our choice and calibration of a model.

Below we provide some examples of the cross-sectional characteristics of various factors. For ease of exposition we use histograms to evaluate the data rather than formal statistical tests. We let particular patterns or properties of the histograms guide us in the choice of the appropriate technique to model the factor. We recommend that an intuitive exploration should be followed by a more formal statistical testing procedure. Our approach here is to analyze the entire sample, all positive values, all negative values, and zero values. Although omitted here, a thorough analysis should also include separate subsample analysis.

Example 1: EBITDA/EV

The first factor we discuss is the earnings before interest, taxes, and amortization to enterprise value (EBITDA/EV) factor. Enterprise value is calculated as the market value of the capital structure. This factor measures the price (enterprise value) investors pay to receive the cash flows (EBITDA) of a company. The economic intuition underlying this factor is that the valuation of a company's cash flow determines the attractiveness of companies to an investor.

Figure 2(A) presents a histogram of all cross-sectional values of the EBITDA/EV factor throughout the entire history of the study. The distribution is close to normal, showing there is a fairly symmetric dispersion among the valuations companies receive. Figure 2(B) shows that the distribution of all the positive values of the factor is also almost normally distributed. On the other hand, Figure 2(C) shows that the distribution of the negative values is skewed to the left. However, because there are only a small number of negative values, it is likely that they will not greatly influence our model.

Example 2: Revisions

We evaluate the cross-sectional distribution of the earnings revisions factor.¹³ The revisions factor we use is derived from sell-side analyst earnings forecasts from the IBES database. The factor is calculated as the number of analysts who revise their earnings forecast upward minus the number of downward forecasts, divided by the total number of forecasts. The economic intuition underlying this factor is that there should be a positive relation to changes in forecasts of earnings and subsequent returns.

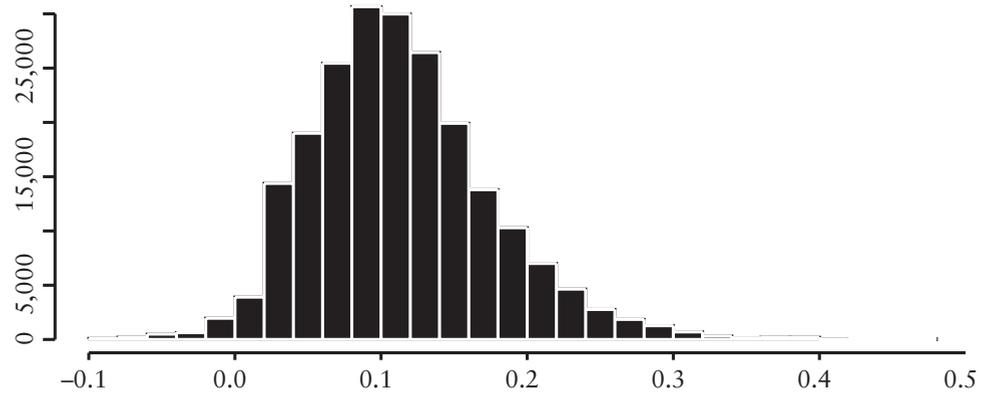
In Figure 3(A) we see that the distribution of revisions is symmetric and leptokurtic around a mean of about zero. This distribution ties with the economic intuition behind the revisions. Since business prospects of companies typically do not change from month-to-month, sell-side analysts will not revise their earnings forecast every month. Consequently, we expect and find the cross-sectional range to be peaked at zero. Figure 3(B) and (C), respectively, show there is a smaller number of both positive and negative earnings revisions and each one of these distributions are skewed.

Example 3: Share Repurchase

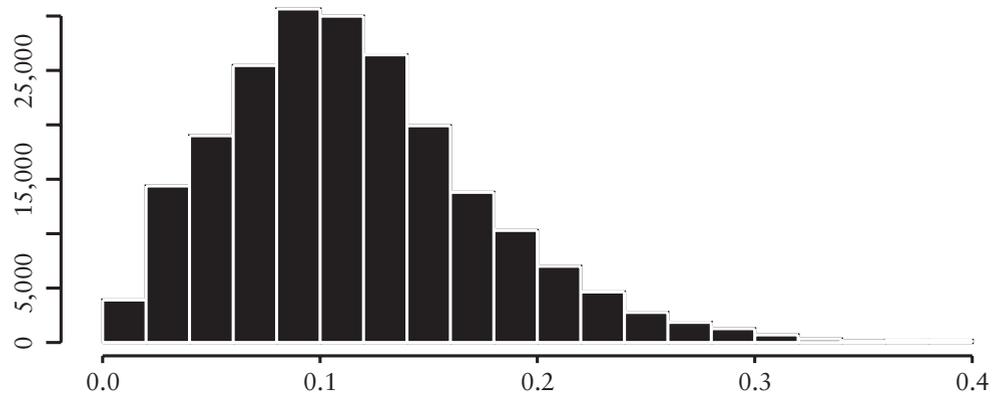
We evaluate the cross-sectional distribution of the shares repurchases factor. This factor is calculated as the difference of the current number of common shares outstanding and the number of shares outstanding 12 months ago, divided by the number of shares outstanding 12 months ago. The economic intuition underlying this factor is that share repurchase provides information to investors about future earnings and valuation of the company's stock.¹⁴ We expect there to be a positive relationship between a reduction in shares outstanding and subsequent returns.

We see in Figure 4(A) that the distribution is leptokurtic. The positive values (see Figure 4(B)) are skewed to the right and the

A. All Factor Values



B. Positive Factor Values



C. Negative Factor Values

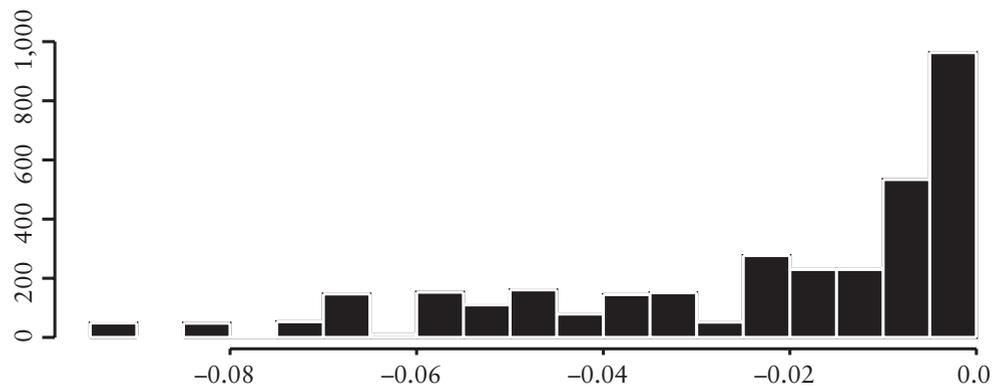
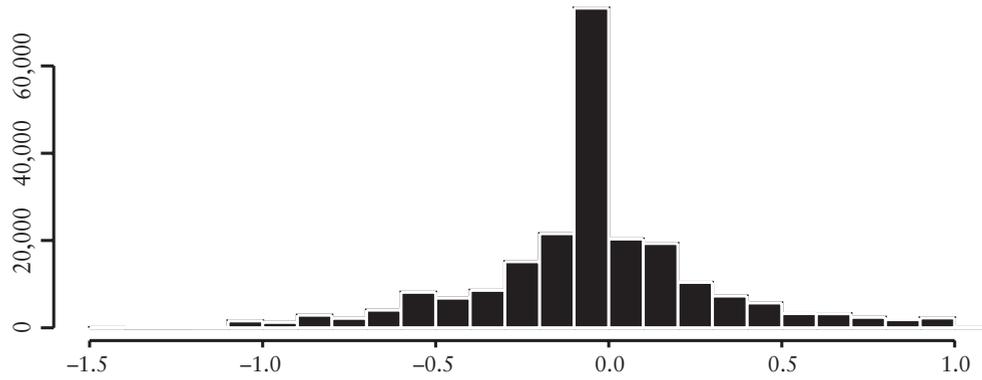
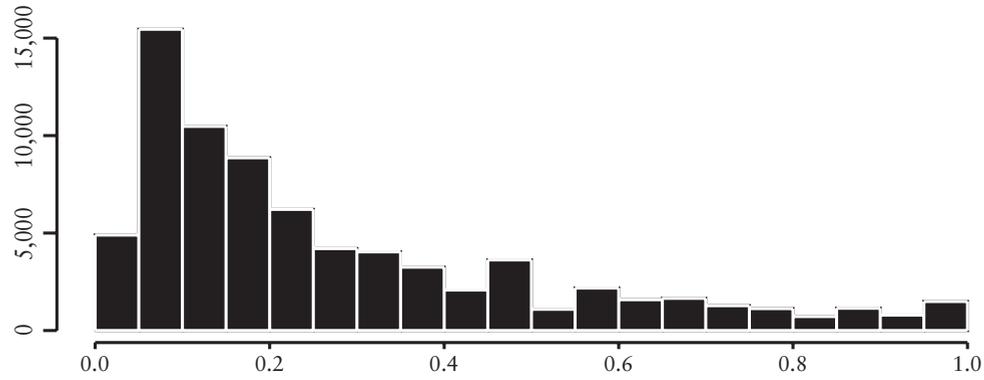


Figure 2 Histograms of the Cross-Sectional Values for the EBITDA/EV Factor

A. All Factor Values



B. Positive Factor Values



C. Negative Factor Values

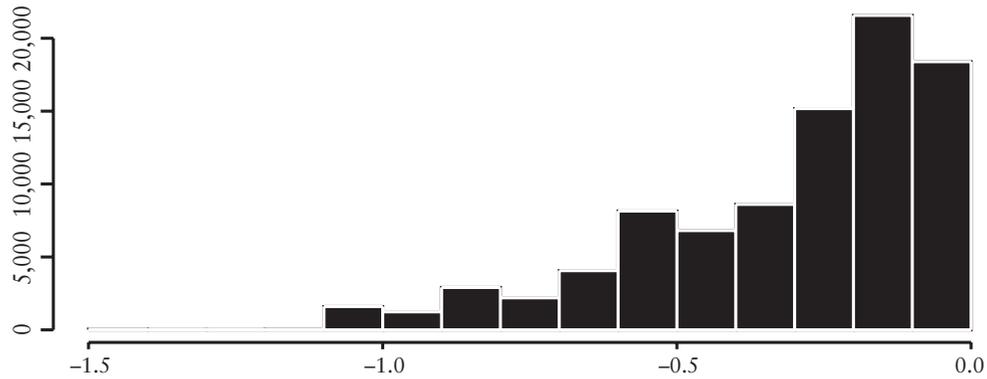


Figure 3 Histograms of the Cross-Sectional Values for the Revisions Factor

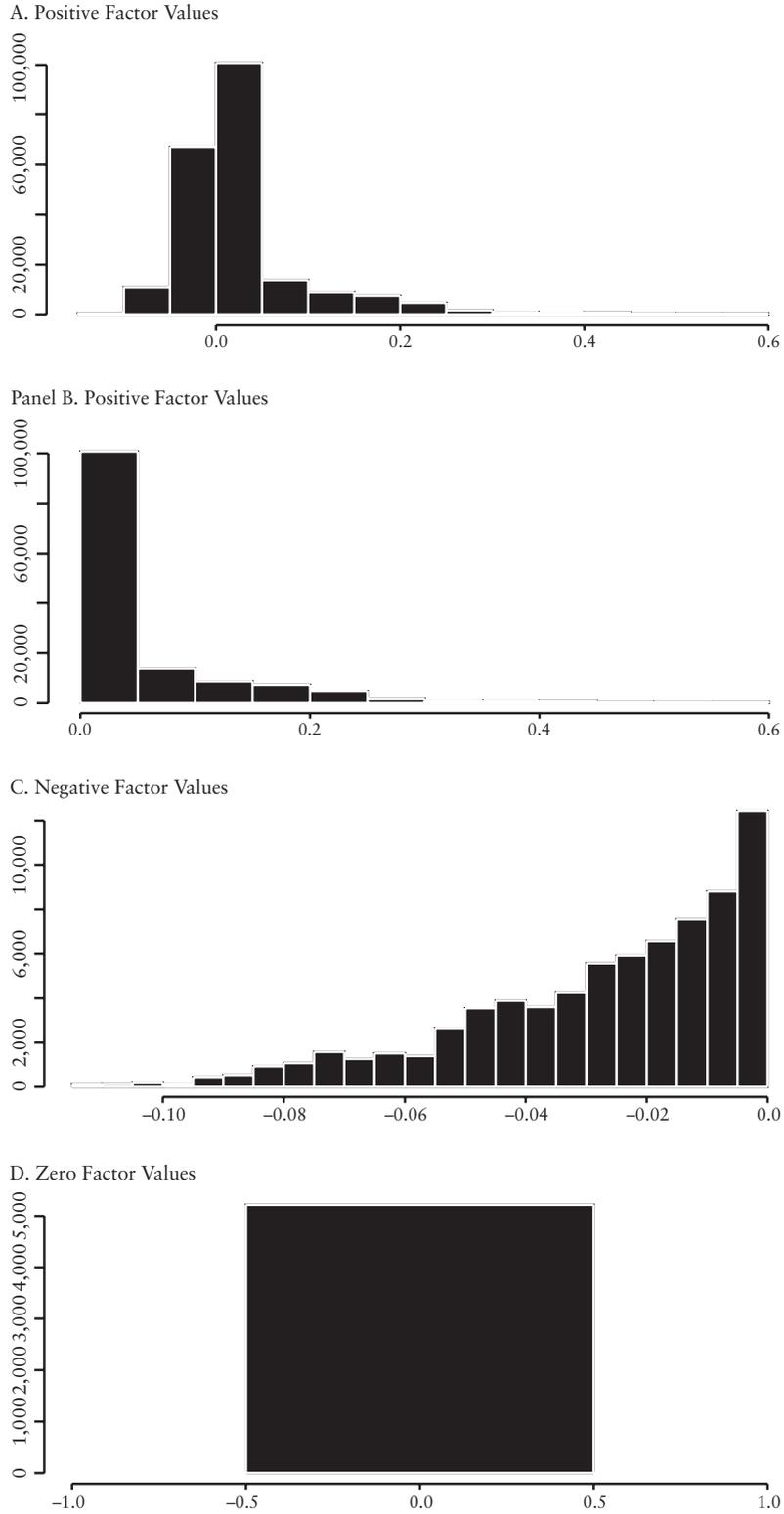


Figure 4 Histograms of the Cross-Sectional Values for the Share Repurchase Factor

negative values (see Figure 4(C)) are clustered in a small band. The economic intuition underlying share repurchases is the following. Firms with increasing share count indicate they require additional sources of cash. This need could be an early sign that the firm is experiencing higher operating risks or financial distress. We would expect these firms to have lower future returns. Firms with decreasing share count have excess cash and are returning value back to shareholders. Decreasing share count could result because management believes the shares are undervalued. As expected, we find the cross-sectional range to be peaked at zero (see Figure 4(D)) since not all firms issue or repurchase shares on a regular basis.

KEY POINTS

- A factor is a common characteristic among a group of assets. Factors should be founded on sound economic intuition, market insight, or an anomaly.
- Factors fall into three categories—macroeconomic, cross-sectional, and statistical factors.
- The main steps in the development of a factor-based trading strategy are (1) defining a trading idea or investment strategy, (2) developing factors, (3) acquiring and processing data, (4) analyzing the factors, (5) building the strategy, (6) evaluating the strategy, (7) back-testing the strategy, and (8) implementing the strategy.
- Most trading strategies are exposed to risk. The main sources of risk are fundamental risk, noise trader risk, horizon risk, model risk, implementation risk, and liquidity risk.
- Factors are often derived from company characteristics and metrics, and market data. Examples of company characteristics and metrics include valuation ratios, operating efficiency ratios, profitability ratios, and solvency ratios. Example of useful market data include analysts forecasts, prices and returns, and trading volume.
- High-quality data are critical to the success of a trading strategy. Model output is only as good as the data used to calibrate it.
- Some common data problems and biases are backfilling and restatements of data, missing data, inconsistently reported data, and survivorship and look-ahead biases.
- The ability to detect and adjust outliers is crucial to a quantitative investment process.
- Common methods used for adjusting data are standardization, orthogonalization, transformation, trimming, and winsorization.
- The statistical properties of factors need to be carefully analyzed. Basic statistical measures include the time-series and cross-sectional averages of the mean, standard deviations, and key percentiles.

NOTES

1. Graham and Dodd (1962).
2. Graham (1949).
3. See Bernstein (1992).
4. See Fama and French (1988).
5. See, for example, Menchero and Poduri (2008).
6. Thomson MarketQA, http://thomsonreuters.com/products_services/financial/financial_products/quantitative_analysis/quantitative_analytics.
7. Factset Research Systems, <http://www.factset.com>.
8. Compustat Xpressfeed, <http://www.compustat.com>.
9. See Barberis and Thaler (2003).
10. For a discussion of the sources of model misspecification and remedies, see Fabozzi, Focardi, and Kolm (2010).
11. The ability of EBITDA/EV to forecast future returns is discussed in, for example, Dechow, Kothari, and Watts (1988).
12. See Nagel (2001).
13. For a representative study see, for example, Bercel (1994).
14. See Grullon and Michaely (2004).

REFERENCES

- Barberis, N., and Thaler, R. (2003). A survey of behavioral finance. In G. M. Constantinides, M. Harris, and R. M. Stulz (Eds.), *Handbook of the Economics of Finance*. Amsterdam: Elsevier Science.
- Bercel, A. (1994). Consensus expectations and international equity returns. *Financial Analysts Journal* 50, 4: 76–80.
- Bernstein, P. L. (1992). *Capital Ideas: The Improbable Origins of Modern Wall Street*. New York: The Free Press.
- Bushee, B. J., and Raedy, J. S. (2006). Factors affecting the implementability of stock market trading strategies. Working paper, University of Pennsylvania and University of North Carolina.
- Cerniglia, J. A., and Kolm, P. N. (2009). The information content of order in a tick-by-tick analysis of the equity market in August 2007. Working paper, Courant Institute, New York University.
- Dechow, P. M., Kothari, S. P., and Watts, R. L. (1998). The relation between earnings and cash flows. *Journal of Accounting and Economics* 25, 2: 133–168.
- Fabozzi, F. J., Focardi, S. M., and Kolm, P. N. (2010). *Quantitative Equity Investing*. Hoboken, NJ: John Wiley & Sons.
- Fama, E. F., and French, K. R. (1988). Dividend yields and expected stock returns. *Journal of Financial Economics* 22, 1: 3–25.
- Graham, B. (1973). *The Intelligent Investor*. New York: Harper & Row.
- Graham, B., and Dodd, D. (1962). *Security Analysis*. New York: McGraw-Hill.
- Grullon, G., and Michaely, R. (2004). The information content of share repurchase programs. *Journal of Finance* 59, 2: 651–680.
- Khandani, A. E., and Lo, A. W. (2007). What happened to the quants in August 2007? *Journal of Investment Management* 5, 4: 5–54.
- Kothari, S. P., Sabino, J. S., and Zach, T. (2005). Implications of survival and data trimming for tests of market efficiency. *Journal of Accounting and Economics* 39, 1: 129–161.
- Mencherio, J., and Poduri, V. (2008). Custom factor attribution. *Financial Analysts Journal* 62, 2: 81–92.
- Nagel, S. (2001). Accounting information free of selection bias: A New UK database 1953–1999. Working paper, Stanford Graduate School of Business.

Cross-Sectional Factor-Based Models and Trading Strategies

JOSEPH A. CERNIGLIA

Visiting Researcher, Courant Institute of Mathematical Sciences, New York University

PETTER N. KOLM, PhD

Director of the Mathematics in Finance Masters Program and Clinical Associate Professor
Courant Institute of Mathematical Sciences, New York University

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Quantitative asset managers construct and apply models that can be used for dynamic multifactor trading strategies. These models incorporate a number of common institutional constraints such as turnover, transaction costs, sector, and tracking error. Approaches for the evaluation of return premiums and risk characteristics to factors include portfolio sorts, factor models, factor portfolios, and information coefficients. Several techniques are used to combine several factors into a single model—a trading strategy. These techniques include data driven, factor model, heuristic, and optimization approaches.

In the construction of factor models, factors are constructed from company characteristics and market data. In this entry, we explain and illustrate how to include multiple factors with the purpose of developing a dynamic multifactor trading strategy that incorporates a number of common institutional constraints such as turnover, transaction costs, sector, and *tracking error*. For this purpose, we use a combination of growth, value, quality, and momentum factors. For the purpose of our illustration, our universe of stocks is the Russell 1000 from December 1989 to December 2008, and we construct our factors by using the Compustat

Point-In-Time and IBES databases. A complete list of the factors and data sets used is provided in the appendix.

We begin by reviewing several approaches for the evaluation of return premiums and risk characteristics to factors, including portfolio sorts, factor models, factor portfolios, and information coefficients. We then turn to techniques that are used to combine several factors into a single model—a trading strategy. In particular, we discuss the data driven, factor model, heuristic, and optimization approaches. It is critical to perform out-of-sample backtests of a trading strategy to understand its performance

and risk characteristics. We cover the split-sample and recursive out-of-sample tests.

Throughout this entry, we provide a series of examples, including backtests of a multifactor trading strategy. The purpose of these examples is not to attempt to provide a profitable trading strategy, but rather to illustrate the process a financial modeler may follow when performing research. We emphasize that the factors that we use are well known and have for years been exploited by industry practitioners. We think that the value added of these examples is in the concrete illustration of the research and development process of a factor-based trading model.

CROSS-SECTIONAL METHODS FOR EVALUATION OF FACTOR PREMIUMS

There are several approaches used for the evaluation of return premiums and risk characteristics to factors. In this section, we discuss the four most commonly used approaches: portfolio sorts, factor models, factor portfolios, and information coefficients. We examine the methodology of each approach and summarize its advantages and disadvantages.

In practice, to determine the right approach for a given situation there are several issues to consider. One determinant is the structure of the financial data. A second determinant is the economic intuition underlying the factor. For example, sometimes we are looking for a monotonic relationship between returns and factors while at other times we care only about extreme values. A third determinant is whether the underlying assumptions of each approach are valid for the data-generating process at hand.

Portfolio Sorts

In the asset pricing literature, the use of *portfolio sorts* can be traced back to the earliest tests of the capital asset pricing model (CAPM). The goal of this particular test is to determine whether a

factor earns a systematic premium. The portfolios are constructed by grouping together securities with similar characteristics (factors). For example, we can group stocks by market capitalization into 10 portfolios—from smallest to largest—such that each portfolio contains stocks with similar market capitalization. The next step is to calculate and evaluate the returns of these portfolios.

The return for each portfolio is calculated by equally weighting the individual stock returns. The portfolios provide a representation of how returns vary across the different values of a factor. By studying the return behavior of the factor portfolios, we may assess the return and risk profile of the factor. In some cases, we may identify a monotonic relationship of the returns across the portfolios. In other cases, we may identify a large difference in returns between the extreme portfolios. In still other cases, there may be no relationship between the portfolio returns. Overall, the return behavior of the portfolios will help us conclude whether there is a premium associated with a factor and describe its properties.

One application of the portfolio sort is the construction of a *factor mimicking portfolio* (FMP). An FMP is a long-short portfolio that goes long stocks with high values of a factor and short stocks with low values of a factor, in equal dollar amounts. An FMP is a zero-cost factor trading strategy.

Portfolio sorts have become so widespread among practitioners and academics alike that they elicit few econometric queries, and often no econometric justification for the technique is offered. While a detailed discussion of these topics is beyond the scope of this book, we would like to point out that asset pricing tests used on sorted portfolios may exhibit a bias that favors rejecting the asset pricing model under consideration.¹

The construction of portfolios sorted on a factor is straightforward:

- Choose an appropriate sorting methodology.
- Sort the assets according to the factor.

- Group the sorted assets into N portfolios (usually $N = 5$, or $N = 10$).
- Compute average returns (and other statistics) of the assets in each portfolio over subsequent periods.

The standard statistical testing procedure for portfolio sorts is to use a Student's t -test to evaluate the significance of the mean return differential between the portfolios of stocks with the highest and lowest values of the factor.

Choosing the Sorting Methodology

The sorting methodology should be consistent with the characteristics of the distribution of the factor and the economic motivation underlying its premium. We list six ways to sort factors:

Method 1

- Sort stocks with factor values from the highest to lowest.

Method 2

- Sort stocks with factor values from the lowest to highest.

Method 3

- First allocate stocks with zero factor values into the bottom portfolio.
- Sort the remaining stocks with nonzero factor values into the remaining portfolios.

For example, the dividend yield factor would be suitable for this sorting approach. This approach aligns the factor's distributional characteristics of dividend and nondividend-paying stocks with the economic rationale. Typically, nondividend-paying stocks maintain characteristics that are different from dividend paying stocks. So we group nondividend-paying stocks into one portfolio. The remaining stocks are then grouped into portfolios depending on the size of their nonzero dividend yields. We differentiate among stocks with dividend yield because of two reasons: (1) the size of the dividend yield is related to the maturity of the company, and (2) some investors prefer to receive their investment return as dividends.

Method 4

- Allocate stocks with zero factor values into the middle portfolio.
- Sort stocks with positive factor values into the remaining higher portfolios (greater than the middle portfolio).
- Sort stocks with negative factor values into the remaining lower portfolios (less than the middle portfolio).

Method 5

- Sort stocks into partitions.
- Rank assets within each partition.
- Combine assets with the same ranking from the different partitions into portfolios.

An example will clarify this procedure. Suppose we want to rank stocks according to earnings growth on a sector-neutral basis. First, we separate stocks into groups corresponding to their sector. Within each sector, we rank the stocks according to their earnings growth. Lastly, we group all stocks with the same rankings of earnings growth into the final portfolio. This process ensures that each portfolio will contain an equal number of stocks from every sector, thereby the resulting portfolios are sector neutral.

Method 6

- Separate all the stocks with negative factor values. Split the group of stocks with negative values into two portfolios using the median value as the break point.
- Allocate stocks with zero factor values into one portfolio.
- Sort the remaining stocks with nonzero factor values into portfolios based on their factor values.

An example of method 6 is the share repurchase factor. We are interested in the extreme positive and negative values of this factor. As we see in Figure 5(A), the distribution of these factors is leptokurtic with the positive values skewed to the right and the negative values clustered in a small range. By choosing method 6 to sort this variable, we can distinguish

between those values we view as extreme. The negative values are clustered so we want to distinguish among the magnitudes of those values. We accomplish this because our sorting method separates the negative values by the median of the negative values. The largest negative values form the extreme negative portfolio. The positive values are skewed to the right, so we want to differentiate between the larger and smaller positive values. When implementing portfolio method 6, we would also separate the zero values from the positive values.

The portfolio sort methodology has several advantages. The approach is easy to implement and can easily handle stocks that drop out or enter into the sample. The resulting portfolios diversify away idiosyncratic risk of individual assets and provide a way of assessing how average returns differ across different magnitudes of a factor.

The portfolio sort methodology has several disadvantages. The resulting portfolios may be exposed to different risks beyond the factor the portfolio was sorted on. In those instances, it is difficult to know which risk characteristics have an impact on the portfolio returns. Because portfolio sorts are nonparametric, they do not give insight as to the functional form of the relation between the average portfolio returns and the factor.

Next we provide three examples to illustrate how the economic intuition of the factor and cross-sectional statistics can help determine the sorting methodology.

Example 1: Portfolio Sorts Based on the EBITDA/EV Factor

Panel A of Figure 1 contains the cross-sectional distribution of the EBITDA/EV factor. This distribution is approximately normally distributed

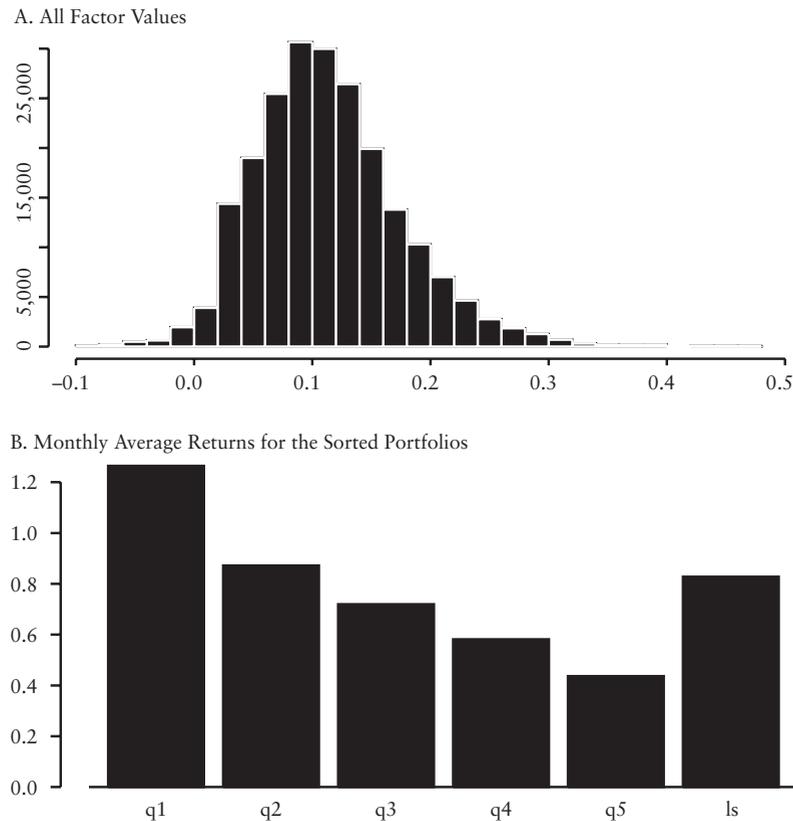


Figure 1 Portfolio Sorts Based on the EBITDA/EV Factor

around a mean of 0.1, with a slight right skew. We use method 1 to sort the variables into five portfolios (denoted by q_1, \dots, q_5) because this sorting method aligns the cross-sectional distribution of factor returns with our economic intuition that there is a linear relationship between the factor and subsequent return. In Figure 1(B), we see that there is a large difference between the equally weighted monthly returns of portfolio 1 (q_1) and portfolio 5 (q_5). Therefore, a trading strategy (denoted by ls in the graph) that goes long portfolio 1 and short portfolio 5 appears to produce abnormal returns.

Example 2: Portfolio Sorts Based on the Revisions Factor

In Figure 2(A), we see that the distribution of earnings revisions is leptokurtic around a mean of about zero, with the remaining val-

ues symmetrically distributed around the peak. The pattern in this cross-sectional distribution provides insight on how we should sort this factor. We use method 3 to sort the variables into five portfolios. The firms with no change in revisions we allocate to the middle portfolio (portfolio 3). The stocks with positive revisions we sort into portfolios 1 and 2, according to the size of the revisions—while we sort stocks with negative revisions into portfolios 4 and 5, according to the size of the revisions. In Figure 2(B), we see there is a relationship between the portfolios and subsequent monthly returns. The positive relationship between revisions and subsequent returns agrees with the factor’s underlying economic intuition: We expect that firms with improving earnings should outperform. The trading strategy that goes long portfolio 1 and short portfolio 5 (denoted by

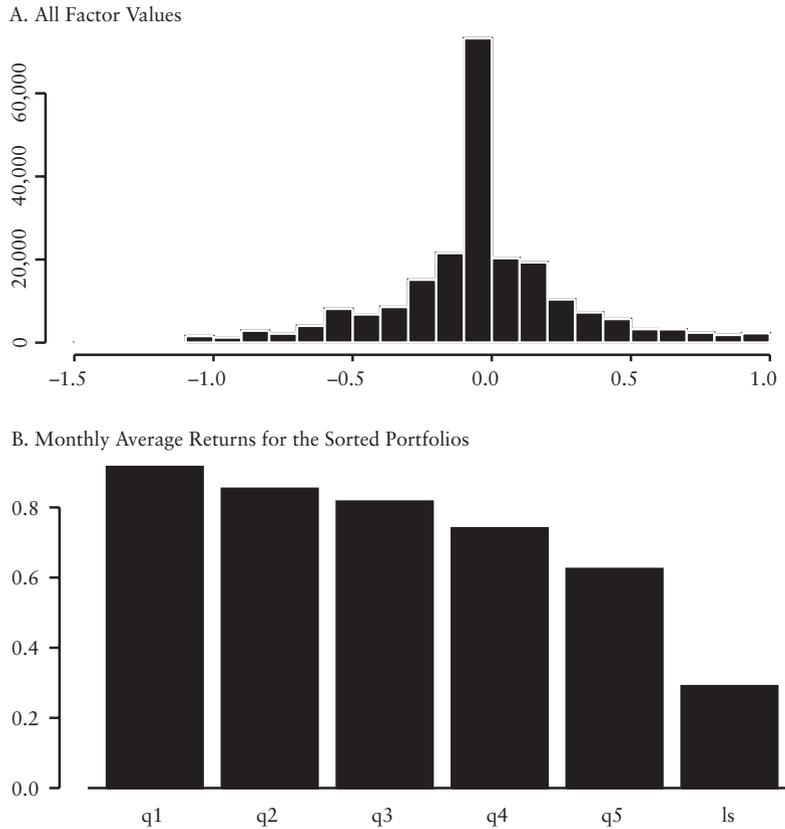


Figure 2 The Revisions Factor

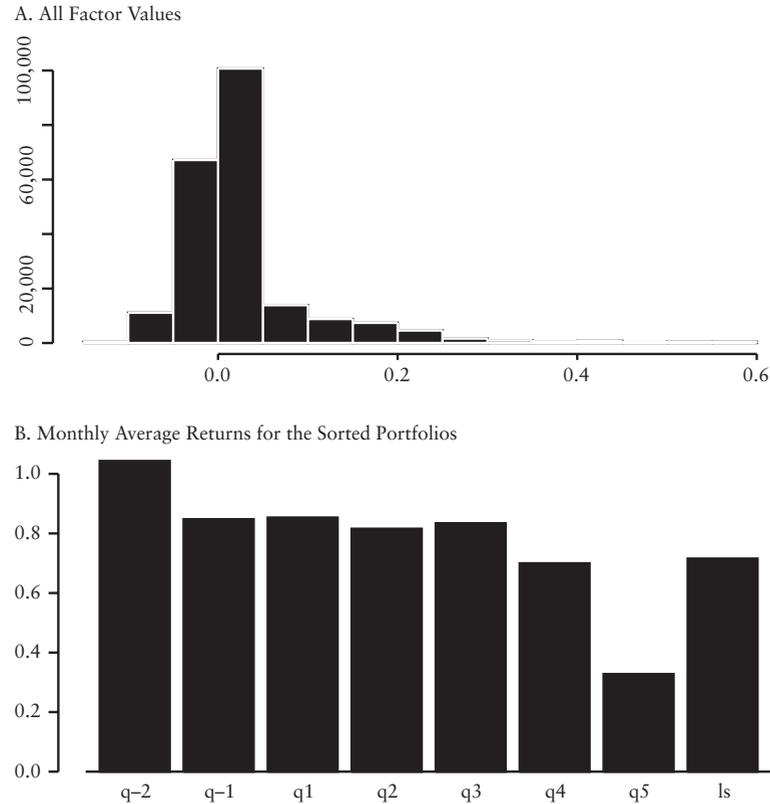


Figure 3 The Share Repurchase Factor

ls in the graph) appears to produce abnormal returns.

Example 3: Portfolio Sorts Based on the Share Repurchase Factor

In Figure 3(A), we see the distribution of share repurchase is asymmetric and leptokurtic around a mean of zero. The pattern in this cross-sectional distribution provides insight on how we should sort this factor. We use method 6 to sort the variables into seven portfolios. We group stocks with positive revisions into portfolios 1 through 5 (denoted by q_1, \dots, q_5 in the graph) according to the magnitude of the share repurchase factor. We allocate stocks with negative repurchases into portfolios q-2 and q-1 where the median of the negative values determines their membership. We split the negative numbers because we are interested in large changes in the shares outstanding. In Fig-

ure 3(B), unlike the other previous factors, we see that there is not a linear relationship between the portfolios. However, there is a large difference in return between the extreme portfolios (denoted by ls in the graph). This large difference agrees with the economic intuition of this factor. Changes in the number of shares outstanding are a potential signal for the future value and prospects of a firm. On the one hand, a large increase in shares outstanding may signal to investors (1) the need for additional cash because of financial distress, or (2) that the firm may be overvalued. On the other hand, a large decrease in the number of shares outstanding may indicate that management believes the shares are undervalued. Finally, small changes in shares outstanding, positive or negative, typically do not have an impact on stock price and therefore are not significant.

Information Ratios for Portfolio Sorts

The information ratio (IR) is a statistic for summarizing the risk-adjusted performance of an investment strategy. It is defined as the ratio of the average excess return to the standard deviation of return. For actively managed equity long portfolios, the IR measures the risk-adjusted value a portfolio manager is adding relative to a benchmark.² IR can also be used to capture the risk-adjusted performance of long-short portfolios from portfolio sorts. When comparing portfolios built using different factors, the IR is an effective measure for differentiating the performance between the strategies.

New Research on Portfolio Sorts

As we mentioned earlier in this section, the standard statistical testing procedure for portfolio sorts is to use a Student's t -test to evaluate the mean return differential between the two portfolios containing stocks with the highest and lowest values of the sorting factor. However, evaluating the return between these two portfolios ignores important information about the overall pattern of returns among the remaining portfolios.

Recent research by Patton and Timmermann (2009) provides new analytical techniques to increase the robustness of inference from portfolio sorts. The technique tests for the presence of a monotonic relationship between the portfolios and their expected returns. To find out if there is a systematic relationship between a factor and portfolio returns, they use the monotonic relation (MR) test to reveal whether the null hypothesis of no systematic relationship can be rejected in favor of a monotonic relationship predicted by economic theory. By MR it is meant that the expected returns of a factor should rise or decline monotonically in one direction as one goes from one portfolio to another. Moreover, Patton and Timmermann develop separate tests to determine the direction of deviations in support of or against the theory.

The authors emphasize several advantages in using this approach. The test is nonparametric and applicable to other cases of portfolios such as two-way and three-way sorts. This test is easy to implement via bootstrap methods. Furthermore, this test does not require specifying the functional form (e.g., linear) in relating the sorting variable to expected returns.

FACTOR MODELS

Classical financial theory states that the average return of a stock is the payoff to investors for taking on risk. One way of expressing this risk-reward relationship is through a factor model. A factor model can be used to decompose the returns of a security into factor-specific and asset-specific returns

$$r_{i,t} = \alpha_i + \beta_{i,1}f_{1,t} + \dots + \beta_{i,K}f_{K,t} + \varepsilon_{i,t}$$

where $\beta_{i,1}, \beta_{i,2}, \dots, \beta_{i,K}$ are the factor exposures of stock i , $f_{1,t}, f_{2,t}, \dots, f_{K,t}$ are the factor returns, α_i is the average abnormal return of stock i , and $\varepsilon_{i,t}$ is the residual.

This factor model specification is contemporaneous, that is, both left- and right-hand side variables (returns and factors) have the same time subscript, t . For trading strategies one generally applies a forecasting specification where the time subscript of the return and the factors are $t+h$ ($h \geq 1$) and t , respectively. In this case, the econometric specification becomes

$$r_{i,t+h} = \alpha_i + \beta_{i,1}f_{1,t} + \dots + \beta_{i,K}f_{K,t} + \varepsilon_{i,t+h}$$

How do we interpret a trading strategy based on a factor model? The explanatory variables represent different factors that forecast security returns, and each factor has an associated factor premium. Therefore, future security returns are proportional to the stock's exposure to the factor premium

$$E(r_{i,t+h} | f_{1,t}, \dots, f_{K,t}) = \alpha_i + \beta_i' \mathbf{f}_t$$

and the variance of future stock return is given by

$$\text{Var}(r_{i,t+b} | f_{1,t}, \dots, f_{K,t}) = \beta_i' E(\mathbf{f}_t \mathbf{f}_t') \beta_i$$

where and $\beta_i = (\beta_{i,1}, \beta_{i,2}, \dots, \beta_{i,k})'$ and $\mathbf{f}_t = (f_{1,t}, f_{2,t}, \dots, f_{K,t})'$.

In the next section we discuss some specific econometric issues regarding cross-sectional regressions and factor models.

Econometric Considerations for Cross-Sectional Factor Models

In cross-sectional regressions, where the dependent variable³ is a stock's return and the independent variables are factors, inference problems may arise that are the result of violations of classical linear regression theory. The three most common problems are measurement problems, common variations in residuals, and multicollinearity.

Measurement Problems

Some factors are not explicitly given, but need to be estimated. These factors are estimated with an error. This error can have an impact on the inference from a factor model. This problem is commonly referred to as the "errors in variables problem." For example, a factor that is comprised of a stock's beta is estimated with an error because beta is determined from a regression of stock excess returns on the excess returns of a market index. While beyond the scope of this entry, several approaches have been suggested to deal with this problem:⁴

Common Variation in Residuals

The residuals from a regression often contain a source of common variation. Sources of common variation in the residuals are heteroskedasticity and serial correlation.⁵ We note that when the form of heteroskedasticity and serial correlation is known, we can apply generalized least squares (GLS). If the form is not known, it has

to be estimated, for example as part of feasible generalized least squares (FGLS). We summarize some additional possibilities next.

Heteroskedasticity occurs when the variance of the residual differs across observations and affects the statistical inference in a linear regression. In particular, the estimated standard errors will be underestimated and the t -statistics will therefore be inflated. Ignoring heteroskedasticity may lead the researcher to find significant relationships where none actually exist. Several procedures have been developed to calculate standard errors that are robust to heteroskedasticity, also known as heteroskedasticity-consistent standard errors.

Serial correlation occurs when residuals terms in a linear regression are correlated, violating the assumptions of regression theory. If the serial correlation is positive, then the standard errors are underestimated and the t -statistics will be inflated. Cochrane (2005) suggests that the errors in cross-sectional regressions using financial data are often off by a factor of 10. Procedures are available to correct for serial correlation when calculating standard errors.

When the residuals from a regression are both heteroskedastic and serially correlated, procedures are available to correct them. One commonly used procedure is the one proposed by Newey and West (1987) referred to as the "Newey-West corrections," and its extension by Andrews (1991).

Petersen (2009) provides guidance on choosing the appropriate method to use for correctly calculating standard errors in panel data regressions when the residuals are correlated. He shows the relative accuracy of the different methods depends on the structure of the data. In the presence of firm effects, where the residuals of a given firm may be correlated across years, ordinary least squares (OLS), Newey-West (modified for panel data sets), or Fama-MacBeth,⁶ corrected for first-order autocorrelation, all produce biased standard errors. To correct for this, Petersen recommends using standard errors clustered by firms. If the firm

effect is permanent, the fixed effects and random effects models produce unbiased standard errors. In the presence of time effects, where the residuals of a given period may be correlated across different firms (cross-sectional dependence), Fama-MacBeth produces unbiased standard errors. Furthermore, standard errors clustered by time are unbiased when there are a sufficient number of clusters. To select the correct approach he recommends determining the form of dependence in the data and comparing the results from several methods.

Gow, Ormazabal, and Taylor (2009) evaluate empirical methods used in accounting research to correct for cross-sectional and time-series dependence. They review each of the methods, including several methods from the accounting literature that have not previously been formally evaluated, and discuss when each method produces valid inferences.

Multicollinearity

Multicollinearity occurs when two or more independent variables are highly correlated. We may encounter several problems when this happens. First, it is difficult to determine which factors influence the dependent variable. Second, the individual p values can be misleading—a p value can be high even if the variable is important. Third, the confidence intervals for the regression coefficients will be wide. They may even include zero. This implies that we cannot determine whether an increase in the independent variable is associated with an increase—or a decrease—in the dependent variable. There is no formal solution based on theory to correct for multicollinearity. The best way to correct for multicollinearity is by removing one or more of the correlated independent variables. It can also be reduced by increasing the sample size.

Fama-MacBeth Regression

To address the inference problem caused by the correlation of the residuals, Fama and MacBeth (1973) proposed the following methodol-

ogy for estimating cross-sectional regressions of returns on factors. For notational simplicity, we describe the procedure for one factor. The multifactor generalization is straightforward.

First, for each point in time t we perform a cross-sectional regression:

$$r_{i,t} = \beta_{i,t} f_t + \varepsilon_{i,t}, \quad i = 1, 2, \dots, N$$

In the academic literature, the regressions are typically performed using monthly or quarterly data, but the procedure could be used at any frequency.

The mean and standard errors of the time series of slopes and residuals are evaluated to determine the significance of the cross-sectional regression. We estimate f and ε_i as the average of their cross-sectional estimates, therefore,

$$\hat{f} = \frac{1}{T} \sum_{t=1}^T \hat{f}_t, \quad \hat{\varepsilon}_i = \frac{1}{T} \sum_{t=1}^T \hat{\varepsilon}_{i,t}$$

The variations in the estimates determine the standard error and capture the effects of residual correlation without actually estimating the correlations.⁷ We use the standard deviations of the cross-sectional regression estimates to calculate the sampling errors for these estimates

$$\sigma_{\hat{f}}^2 = \frac{1}{T^2} \sum_{t=1}^T (\hat{f}_t - \hat{f})^2, \quad \sigma_{\hat{\varepsilon}_i}^2 = \frac{1}{T^2} \sum_{t=1}^T (\hat{\varepsilon}_{i,t} - \hat{\varepsilon}_i)^2$$

Cochrane (2005) provides a detailed analysis of this procedure and compares it to cross-sectional OLS and pooled time-series cross-sectional OLS. He shows that when the factors do not vary over time and the residuals are cross-sectionally correlated, but not correlated over time, then these procedures are all equivalent.

Information Coefficients

To determine the forecast ability of a model, practitioners commonly use a statistic called the *information coefficient* (IC). The IC is a linear statistic that measures the cross-sectional

correlation between a factor and its subsequent realized return:⁸

$$IC_{t,t+k} = \text{corr}(\mathbf{f}_t, \mathbf{r}_{t,t+k})$$

where \mathbf{f}_t is a vector of cross sectional factor values at time t and $\mathbf{r}_{t,t+k}$ is a vector of returns over the time period t to $t+k$.

Just like the standard correlation coefficient, the values of the IC range from -1 to $+1$. A positive IC indicates a positive relation between the factor and return. A negative IC indicates a negative relation between the factor and return. ICs are usually calculated over an interval, for example, daily or monthly. We can evaluate how a factor has performed by examining the time series behavior of the ICs. Looking at the mean IC tells how predictive the factor has been over time.

An alternate specification of this measure is to make \mathbf{f}_t the rank of a cross-sectional factor. This calculation is similar to the Spearman rank coefficient. By using the rank of the factor, we focus on the ordering of the factor instead of its value. Ranking the factor value reduces the undue influence of outliers and reduces the influence of variables with unequal variances. For the same reasons, we may also choose to rank the returns instead of using their numerical value.

Sorensen, Qian, and Hua (2007) present a framework for factor analysis based on ICs. Their measure of IC is the correlation between the factor ranks, where the ranks are the normalized z-score of the factor,⁹ and subsequent return. Intuitively, this IC calculation measures the return associated with a one standard deviation exposure to the factor. Their IC calculation is further refined by risk adjusting the value. To risk adjust, the authors remove systematic risks from the IC and accommodate the IC for specific risk. By removing these risks, Qian and Hua (2004) show that the resulting ICs provide a more accurate measure of the return forecasting ability of the factor.

The subsequent realized returns to a factor typically vary over different time horizons. For example, the return to a factor based on

price reversal is realized over short horizons, while valuation metrics such as EBITDA/EV are realized over longer periods. It therefore makes sense to calculate multiple ICs for a set of factor forecasts whereby each calculation varies the horizon over which the returns are measured.

The IC methodology has many of the same advantages as regression models. The procedure is easy to implement. The functional relationship between factor and subsequent returns is known (linear).

ICs can also be used to assess the risk of factors and trading strategies. The standard deviation of the time series (with respect to t) of ICs for a particular factor ($\text{std}(IC_{t,t+k})$) can be interpreted as the strategy risk of a factor. Examining the time series behavior of $\text{std}(IC_{t,t+k})$ over different time periods may give a better understanding of how often a particular factor may fail. Qian and Hua show that $\text{std}(IC_{t,t+k})$ can be used to more effectively understand the active risk of investment portfolios. Their research demonstrates that ex post tracking error often exceeds the ex ante tracking provided by risk models. The difference in tracking error occurs because tracking error is a function of both ex ante tracking error from a risk model and the variability of information coefficients, $\text{std}(IC_{t,t+k})$. They define the expected tracking error as

$$\sigma_{TE} = \text{std}(IC_{t,t+k})\sqrt{N}\sigma_{\text{model}}\text{dis}(\mathbf{R}_t)$$

where N is the number of stocks in the universe (breadth), σ_{model} is the risk model tracking error, and $\text{dis}(\mathbf{R}_t)$ is dispersion of returns¹⁰ defined by

$$\text{dis}(\mathbf{R}_t) = \text{std}(r_{1,t}, r_{2,t}, \dots, r_{N,t})$$

Example: Information Coefficients

Figure 4 displays the time-varying behavior of ICs for each one of the factors EBITDA/EV, growth of fiscal year 1 and fiscal year 2 earnings estimates, revisions, and momentum. The

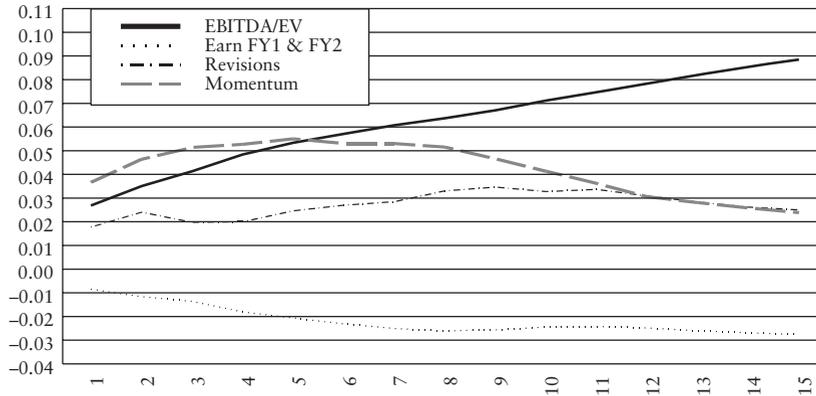


Figure 4 Information Coefficients over Various Horizons for EBITDA/EV, Growth of Fiscal Year 1 and Fiscal Year 2 Earnings Estimates, Revisions, and Momentum Factors

graph shows the time series average of information coefficients:

$$\overline{IC}_k = \text{mean}(IC_{t,t+k})$$

The graph depicts the information horizons for each factor, showing how subsequent return is realized over time. The vertical axis shows the size of the average information coefficient \overline{IC}_k for $k = 1, 2, \dots, 15$.

Specifically, the EBITDA/EV factor starts at almost 0.03 and monotonically increases as the investment horizon lengthens from one month to 15 months. At 15 months, the EBITDA/EV factor has an IC of 0.09, the highest value among all the factors presented in the graph. This relationship suggests that the EBITDA/EV factor earns higher returns as the holding period lengthens.

The other ICs of the factors in the graph are also interesting. The growth of fiscal year 1 and fiscal year 2 earnings estimates factor is defined as the growth in current fiscal year (fy1) earnings estimates to the next fiscal year (fy2) earnings estimates provided by sell-side analysts.¹¹ We call the growth of fiscal year 1 and fiscal year 2 earnings estimates factor the *earnings growth factor* throughout the remainder of the entry. The IC is negative and decreases as the investment horizon lengthens. The momentum factor starts with a positive IC of 0.02

and increases to approximately 0.055 in the fifth month. After the fifth month, the IC decreases. The revisions factor starts with a positive IC and increases slightly until approximately the eleventh month at which time the factor begins to decay.

Looking at the overall patterns in the graph, we see that the return realization pattern to different factors varies. One notable observation is that the returns to factors don't necessarily decay but sometimes grow with the holding period. Understanding the multiperiod effects of each factor is important when we want to combine several factors. This information may influence how one builds a model. For example, we can explicitly incorporate this information about information horizons into our model by using a function that describes the decay or growth of a factor as a parameter to be calibrated. Implicitly, we could incorporate this information by changing the holding period for a security traded for our trading strategy. Specifically, Sneddon (2008) discusses an example that combines one signal that has short-range predictive power with another that has long-range power. Incorporating this information about the information horizon often improves the return potential of a model. Kolm (2010) describes a general multiperiod model that combines information decay, market impact costs, and real world constraints.

Factor Portfolios

Factor portfolios are constructed to measure the information content of a factor. The objective is to mimic the return behavior of a factor and minimize the residual risk. Similar to portfolio sorts, we evaluate the behavior of these factor portfolios to determine whether a factor earns a systematic premium.

Typically, a factor portfolio has a unit exposure to a factor and zero exposure to other factors. Construction of factor portfolios requires holding both long and short positions. We can also build a factor portfolio that has exposure to multiple attributes, such as beta, sectors, or other characteristics. For example, we could build a portfolio that has a unit exposure to book-to-price and small size stocks. Portfolios with exposures to multiple factors provide the opportunity to analyze the interaction of different factors.

A Factor Model Approach

By using a multifactor model, we can build factor portfolios that control for different risks.¹² We decompose return and risk at a point in time into a systematic and specific component using the regression:

$$\mathbf{r} = \mathbf{X}\mathbf{b} + \mathbf{u}$$

where \mathbf{r} is an N vector of excess returns of the stocks considered, \mathbf{X} is an N by K matrix of factor loadings, \mathbf{b} is a K vector of factor returns, and \mathbf{u} is an N vector of firm specific returns (residual returns). Here, we assume that factor returns are uncorrelated with the firm specific return. Further assuming that firm specific returns of different companies are uncorrelated, the N by N covariance matrix of stock returns \mathbf{V} is given by

$$\mathbf{V} = \mathbf{X}\mathbf{F}\mathbf{X}' + \mathbf{\Delta}$$

where \mathbf{F} is the K by K factor return covariance matrix and $\mathbf{\Delta}$ is the N by N diagonal matrix of variances of the specific returns.

We can use the Fama-MacBeth procedure discussed earlier to estimate the factor returns over time. Each month, we perform a GLS regression to obtain

$$\mathbf{b} = (\mathbf{X}'\mathbf{\Delta}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Delta}^{-1}\mathbf{r}$$

OLS would give us an unbiased estimate, but since the residuals are heteroskedastic the GLS methodology is preferred and will deliver a more efficient estimate. The resulting holdings for each factor portfolio are given by the rows of $(\mathbf{X}'\mathbf{\Delta}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Delta}^{-1}$.

An Optimization-Based Approach

A second approach to build factor portfolios uses mean-variance optimization. Using optimization techniques provides a flexible approach for implementing additional objectives and constraints.¹³

Using the notation from the previous subsection, we denote by \mathbf{X} the set of factors. We would like to construct a portfolio that has maximum exposure to one target factor from \mathbf{X} (the alpha factor), zero exposure to all other factors, and minimum portfolio risk. Let us denote the alpha factor by \mathbf{X}_α and all the remaining ones by \mathbf{X}_σ . Then the resulting optimization problem takes the form

$$\begin{aligned} \max_{\mathbf{w}} \quad & \left\{ \mathbf{w}'\mathbf{X}_\alpha - \frac{1}{2}\lambda\mathbf{w}'\mathbf{V}\mathbf{w} \right\} \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{X}_\sigma = 0 \end{aligned}$$

The analytical solution to this optimization problem is given by

$$h^* = \frac{1}{\lambda}\mathbf{V}^{-1}[\mathbf{I} - \mathbf{X}_\sigma(\mathbf{X}'_\sigma\mathbf{V}^{-1}\mathbf{X}_\sigma)^{-1}\mathbf{X}'_\sigma\mathbf{V}^{-1}]\mathbf{X}_\alpha$$

We may want to add additional constraints to the problem. Constraints are added to make factor portfolios easier to implement and meet additional objectives. Some common constraints include limitations on turnover, transaction costs, the number of assets, and liquidity preferences. These constraints¹⁴ are typically implemented as linear inequality constraints. When no analytical solution is available to solve the

optimization with linear inequality constraints, we have to resort to quadratic programming (QP).¹⁵

PERFORMANCE EVALUATION OF FACTORS

Analyzing the performance of different factors is an important part of the development of a *factor-based trading strategy*. A researcher may construct and analyze over a hundred different factors, so a process to evaluate and compare these factors is needed. Most often this process starts by trying to understand the time-series properties of each factor in isolation and then study how they interact with each other.

To give a basic idea of how this process may be performed, we use the five factors introduced earlier in this entry: EBITDA/EV, revisions, share repurchase, momentum, and earnings growth. These are a subset of the factors that we use in the factor trading strategy model discussed later in the entry. We choose a limited number of factors for ease of exposition. In particular, we emphasize those factors that possess more interesting empirical characteristics.

Figure 5(A) presents summary statistics of monthly returns of long-short portfolios con-

structed from these factors. We observe that the average monthly return ranges from -0.05% for the earnings growth to 0.90% for the momentum factor. The t -statistics for the mean return are significant at the 95% level for the EBITDA/EV, share repurchase, and momentum factors. The monthly volatility ranges from 3.77% for the revisions factor to 7.13% for the momentum factor. In other words, the return and risk characteristics among factors vary significantly. We note that the greatest monthly drawdown has been large to very large for all of the factors, implying significant downside risk. Overall, the results suggest that there is a systematic premium associated with the EBITDA/EV, share repurchase, and momentum factors.

Let pctPos and pctNeg denote the fraction of positive and negative returns over time, respectively. These measures offer another way of interpreting the strength and consistency of the returns to a factor. For example, EBITDA/EV and momentum have t -statistics of 2.16 and 1.90, respectively, indicating that the former is stronger. However, pctPos (pctNeg) are 0.55 versus 0.61 (0.45 versus 0.39) showing that positive returns to momentum occur more frequently. This may provide reassurance of the

A. Summary Statistics of Monthly Returns of Long-Short Portfolios

	Mean	Stdev	Median	t -stat	Max	Min	pctPos	pctNeg
Revisions	0.29	3.77	0.77	1.17	10.43	-19.49	0.55	0.45
EBITDA/EV	0.83	5.80	0.72	2.16	31.61	-30.72	0.55	0.45
Share repurchase	0.72	3.89	0.43	2.78	22.01	-14.06	0.61	0.39
Momentum	0.90	7.13	0.97	1.90	25.43	-42.71	0.61	0.39
Earnings growth	-0.05	4.34	0.25	-0.18	14.03	-23.10	0.53	0.47

B. Correlations between Long-Short Portfolios

	Revisions	EBITDA/EV	Share Repurchase	Momentum	Earnings Growth
Revisions	1.00	-0.28	0.01	0.79	0.25
EBITDA/EV	-0.28	1.00	0.78	-0.12	0.01
Share repurchase	0.01	0.78	1.00	0.20	0.12
Momentum	0.79	-0.12	0.20	1.00	0.28
Earnings growth	0.25	0.01	0.12	0.28	1.00

Figure 5 Results from Portfolio Sorts

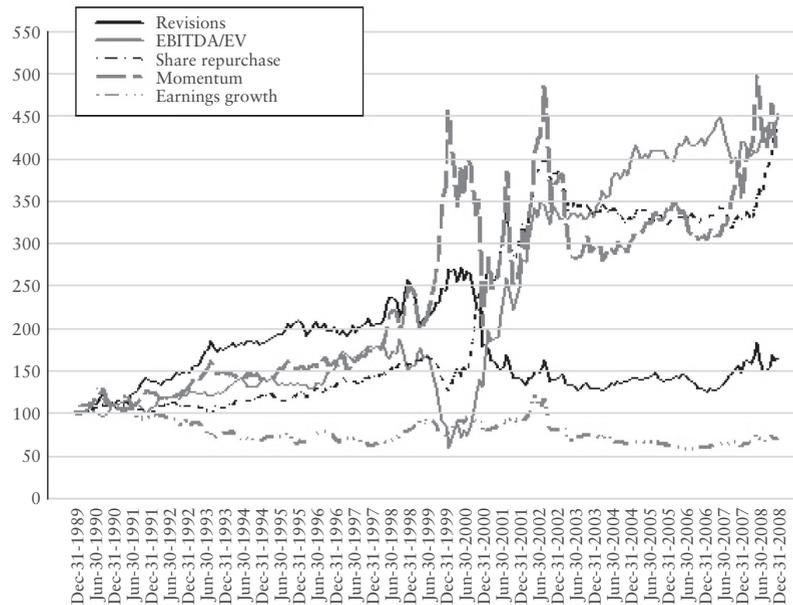


Figure 6 Cumulative Returns of Long-Short Portfolios

usefulness of the momentum factor, despite the fact that its t -statistic is below the 95% level.

Figure 5(B) presents unconditional correlation coefficients of monthly returns for long-short portfolios. The comovement of factor returns varies among the factors. The lowest correlation is -0.28 between EBITDA/EV and revisions. The highest correlation is 0.79 between momentum and revisions. In addition, we observe that the correlation between revisions and share repurchase, and between EBITDA/EV and earnings growth are close to zero. The broad range of correlations provides evidence that combining uncorrelated factors could produce a successful strategy.

Figure 6 presents the cumulative returns for the long-short portfolios. The returns of the long-short factor portfolios experience substantial volatility. We highlight the following patterns of cumulative returns for the different factors:

- The cumulative return of the revisions factor is positive in the early periods (12/1989 to 6/1998). While it is volatile, its cumulative return is higher in the next period (7/1998 to 7/2000). It deteriorates sharply in the following period (8/2000 to 6/2003), and levels out in the later periods (7/2003 to 12/2008).
- The performance of the EBITDA/EV factor is consistently positive in the early periods (12/1989 to 9/1998), deteriorates in the next period (10/1998 to 1/2000) and rebounds sharply (2/2000 to 7/2002), grows at a slower but more historically consistent rate in the later periods (8/2002 to 4/2007), deteriorates in the next period (5/2007 to 9/2007), and returns to more historically consistent returns in last period (10/2007 to 12/2008).
- The cumulative return of the share repurchase factor grows at a slower pace in the early years (12/1989 to 5/1999), falls slightly in the middle periods (6/1999 to 1/2000), rebounds sharply (2/2000 to 7/2002), falls then flattens out in the next period (8/2002 to 4/2008), and increases at a large rate late in the graph (5/2008 to 12/2008).
- The momentum factor experiences the largest volatility. This factor performs consistently well in the early period (12/1989 to 12/1998),

A. Basic Statistics for Monthly Information Coefficients								
	Mean	Stdev	Median	<i>t</i> -stat	Max	Min	pctPos	pctNeg
Revisions	0.02	0.10	0.02	2.51	0.31	-0.29	0.58	0.42
EBITDA/EV	0.03	0.13	0.02	3.13	0.48	-0.41	0.59	0.41
Share repurchase	-0.01	0.10	-0.00	-2.13	0.20	-0.45	0.48	0.52
Momentum	0.03	0.18	0.05	2.86	0.50	-0.57	0.59	0.41
Earnings growth	-0.00	0.13	0.00	-0.56	0.26	-0.28	0.51	0.49

B. Correlations for Monthly Average Information Coefficients					
	Revisions	EBITDA/EV	Share Repurchase	Momentum	Earnings Growth
Revisions	1.00	-0.31	0.13	0.79	-0.14
EBITDA/EV	-0.31	1.00	-0.66	-0.26	-0.49
Share repurchase	0.13	-0.66	1.00	0.02	0.58
Momentum	0.79	-0.26	0.02	1.00	-0.05
Earnings growth	-0.14	-0.49	0.58	-0.05	1.00

Figure 7 Summary of Monthly Factor Information Coefficients

experiences sharp volatility in the middle period (1/1999 to 5/2003), flattens out (6/2003 to 6/2007), and grows at an accelerating rate from (7/2007 to 12/2008).

- The performance of the earnings growth factor is flat or negative throughout the entire period.

The overall pattern of the cumulative returns among the factors clearly illustrates that factor returns and correlations are time varying.

In Figure 7(A), we present summary statistics of the monthly information coefficients of the factors. The average monthly information coefficients range from 0.03 for EBITDA/EV and momentum, to 0.01 for the share repurchase factor. The *t*-statistics for the mean ICs are significant at the 95% level for all factors except earnings growth. With the exception of share repurchase and earnings growth, the fraction of positive returns of the factors are significantly greater than that of the negative returns.

The share repurchase factor requires some comments. The information coefficient is negative, in contrast to the positive return in the long-short portfolio sorts, because negative share repurchases are correlated with subsequent return. The information coefficient is lower than we would expect because there is

not a strong linear relation between the return and the measures. As the results from the portfolio sorts indicate, the extreme values of this factor provide the highest returns.

Figure 7(B) displays unconditional correlation coefficients of the monthly information coefficients. The comovement of the ICs factor returns varies among the factors. The lowest correlation is -0.66 between EBITDA/EV and share repurchases. But again this should be interpreted with caution because it is negative repurchases that we view as attractive. The highest correlation reported in the exhibit is 0.79 between momentum and revisions. Similar to the correlation of long-short factor portfolio returns, the diverse set of correlations provides evidence that combining uncorrelated factors may produce a successful strategy.

In Figure 8(A), we present summary statistics of the time series average of the monthly coefficients from the Fama-MacBeth (FM) regressions of the factors. The information provided by the FM coefficients differs from the information provided by portfolio sorts. The FM coefficients show the linear relationship between the factor and subsequent returns, while the results from the portfolio sorts provide information on the extreme values of the factors and subsequent returns. The difference in the

A. Basic Statistics for Fama-MacBeth Regression Coefficients								
	Mean	Stdev	Median	<i>t</i> -stat	Max	Min	pctPos	pctNeg
Revisions	0.09	1.11	0.22	1.22	3.36	-5.26	0.59	0.41
EBITDA/EV	0.27	1.61	0.14	2.50	8.69	-7.81	0.59	0.41
Share repurchase	-0.18	0.96	-0.06	-2.90	3.21	-5.91	0.44	0.56
Momentum	0.31	2.42	0.29	1.94	9.97	-12.37	0.60	0.40
Earnings growth	-0.08	0.99	-0.04	-1.20	2.83	-4.13	0.48	0.52

B. Correlations for Fama-MacBeth Regression Coefficients					
	Revisions	EBITDA/EV	Share Repurchase	Momentum	Earnings Growth
Revisions	1.00	-0.27	0.05	0.77	-0.26
EBITDA/EV	-0.27	1.00	-0.75	-0.18	-0.58
Share repurchase	0.05	-0.75	1.00	-0.04	0.64
Momentum	0.77	-0.18	-0.04	1.00	-0.18
Earnings growth	-0.26	-0.58	0.64	-0.18	1.00

Figure 8 Summary of Monthly Fama-MacBeth Regression Coefficients

size of the mean returns between the FM coefficients and portfolio sorts exits partially because the intercept terms from the FM regressions are not reported in the exhibit.

The average monthly FM coefficient ranges from -0.18 for share repurchase to 0.31 for the momentum factor. Again the share repurchase results should be interpreted with caution because it is negative repurchases that we view as attractive. The *t*-statistics are significant at the 95% level for the EBITDA/EV and share repurchase factors.

Also, we compare the results of portfolio sorts in Figure 7(A) with the FM coefficients in Figure 8(A). The rank ordering of the magnitude of factor returns is similar between the two panels. The *t*-statistics are slightly higher in the FM regressions than the portfolio sorts. The correlation coefficients for the portfolio sorts in Figure 7(B) are consistent with the FM coefficients in Figure 8(B) for all the factors except for shares repurchases. The results for share repurchases need to be interpreted with caution because it is negative repurchases that we view as attractive. The portfolio sorts take that into account while FM regressions do not.

To better understand the time variation of the performance of these factors, we calculate rolling 24-month mean returns and correlations

of the factors. The results are presented in Figure 9. We see that the returns and correlations to all factors are time varying. A few of the time series experience large volatility in the rolling 24-month returns. The EBITDA/EV factor shows the largest variation followed by the momentum and share repurchase factors. All factors experience periods where the rolling average returns are both positive and negative.

Figure 10 presents the rolling correlation between pairs of the factors. There is substantial variability in many of the pairs. In most cases the correlation moves in a wave-like pattern. This pattern highlights the time-varying property of the correlations among the factors. This property will be important to incorporate in a factor trading model. The most consistent correlation is between momentum and revisions factors and this correlation is, in general, fairly high.

MODEL CONSTRUCTION METHODOLOGIES FOR A FACTOR-BASED TRADING STRATEGY

In the previous section, we analyzed the performance of each factor. The next step in

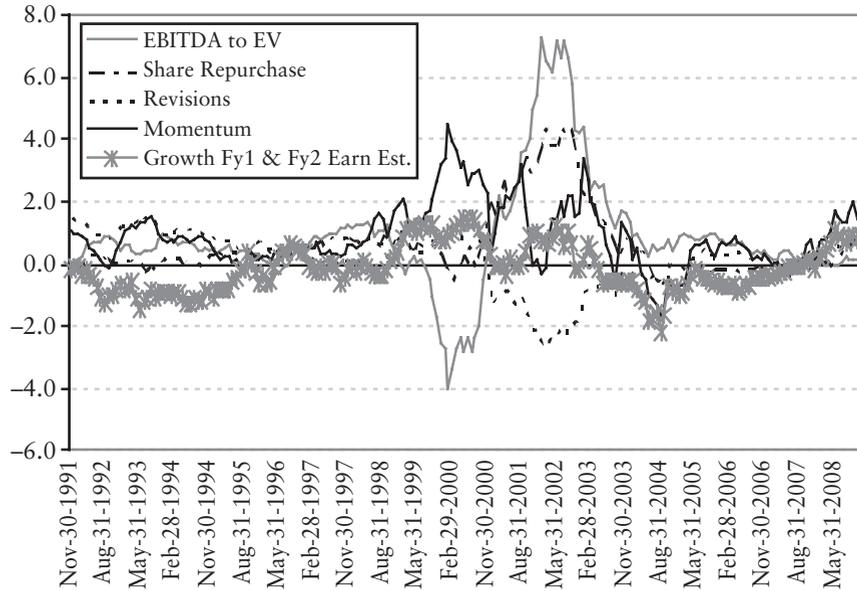


Figure 9 Rolling 24-Month Mean Returns for the Factors

building our trading strategy is to determine how to combine the factors into one model. The key aspect of building this model is to (1) determine what factors to use out of the universe of factors that we have, and (2) how to weight them.

We describe four methodologies to combine and weight factors to build a model for a trading strategy. These methodologies are used to translate the empirical work on factors into a working model. Most of the methodologies are flexible in their specification and there is

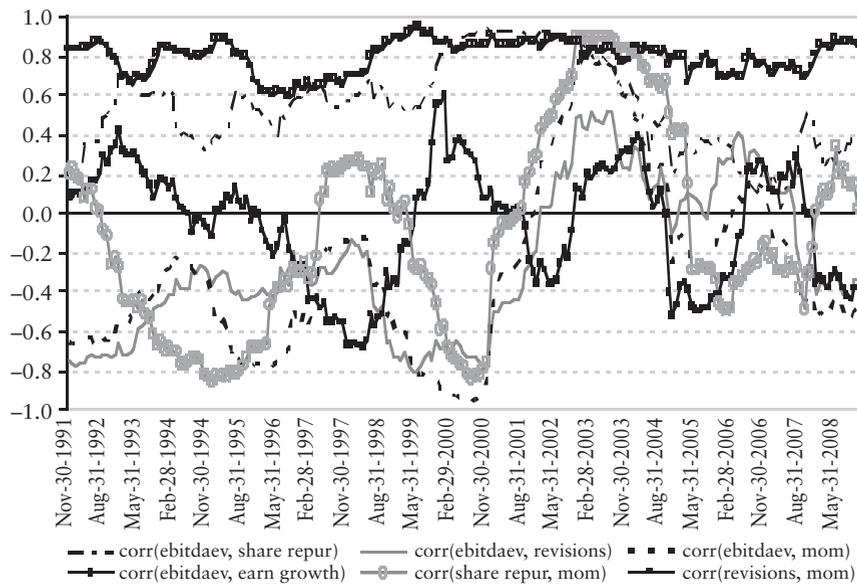


Figure 10 Rolling 24-Month Correlations of Monthly Returns for the Factors

some overlap between them. Though the list is not exhaustive, we highlight those processes frequently used by quantitative portfolio managers and researchers today. The four methodologies are the data driven, the factor model, the heuristic, and the optimization approaches.

It is important to be careful how each methodology is implemented. In particular, it is critical to balance the iterative process of finding a robust model with good forecasting ability versus finding a model that is a result of data mining.

The Data Driven Approach

A *data driven approach* uses statistical methods to select and weight factors in a forecasting model. This approach uses returns as the independent variables and factors as the dependent variables. There are a variety of estimation procedures, such as neural nets, classification trees, and principal components, that can be used to estimate these models. Usually a statistic is established to determine the criteria for a successful model. The algorithm of the statistical method evaluates the data and compares the results against the criteria.

Many data driven approaches have no structural assumptions on potential relationships the statistical method finds. Therefore, it is sometimes difficult to understand or even explain the relationship among the dependent variables used in the model.

Deistler and Hamann (2005) provide an example of a data driven approach to model development. The model they develop is used for forecasting the returns to financial stocks. To start, they split their data sample into two parts—an in-sample part for building the model and an out-of-sample part to validate the model. They use three different types of factor models for forecasting stock returns: quasistatic principal components, quasistatic factor models with idiosyncratic noise, and reduced rank regression. For model selection Deistler

and Hamann use an iterative approach where they find the optimal mix of factors based on the Akaike's information criterion and the Bayesian information criterion. A large number of different models are compared using the out-of-sample data. They find that the reduced rank model provides the best performance. This model produced the highest out-of-sample R^2 s, hit rates,¹⁶ and Diebold-Mariano test statistic¹⁷ among the different models evaluated.

The Factor Model Approach

In this section, we briefly address the use of factor models for forecasting. The goal of the factor model is to develop a parsimonious model that forecasts returns accurately. One approach is for the researcher to predetermine the variables to be used in the factor model based on economic intuition. The model is estimated and then the estimated coefficients are used to produce the forecasts.

A second approach is to use statistical tools for model selection. In this approach we construct several models—often by varying the factors and the number of factors used—and have them compete against each other, just like a horse race. We then choose the best performing model.

Factor model performance can be evaluated in three ways. We can evaluate the fit, forecast ability, and economic significance of the model. The measure to evaluate the fit of a model is based on statistical measures including the model's R^2 and adjusted R^2 , and F - and t -statistics of the model coefficients.

There are several methods to evaluate how well a model will forecast. West (2004) discusses the theory and conventions of several measures of relative model quality. These methods use the resulting time series of predictions and prediction errors from a model. In the case where we want to compare models, West suggests ratios or differences of mean; mean-square or mean-absolute prediction errors; correlation between

one model's prediction and another model's realization (also known as forecast encompassing); or comparison of utility or profit-based measures of predictive ability. In other cases where we want to assess a single model, he suggests measuring the correlation between prediction and realization, the serial correlation in one step ahead prediction errors, the ability to predict direction of change, and the model prediction bias.

We can evaluate economic significance by using the model to predict values and using the predicted values to build portfolios. The profitability of the portfolios is evaluated by examining statistics such as mean returns, information ratios, dollar profits, and drawdown.

The Heuristic Approach

The *heuristic approach* is another technique used to build trading models. Heuristics are based on common sense, intuition, and market insight and are not formal statistical or mathematical techniques designed to meet a given set of requirements. Heuristic-based models result from the judgment of the researcher. The researcher decides the factors to use, creates rules in order to evaluate the factors, and chooses how to combine the factors and implement the model.

Piotroski (2000) applies a heuristic approach in developing an investment strategy for high-value stocks (high book-to-market firms). He selects nine fundamental factors¹⁸ to measure three areas of the firm's financial condition: profitability, financial leverage and liquidity, and operating efficiency. Depending on the factor's implication for future prices and profitability, each factor is classified as either "good" or "bad." An indicator variable for the factor is equal to one (zero) if the factor's realization is good (bad). The sum of the nine binary factors is the F_SCORE. This aggregate score measures the overall quality, or strength, of the firm's financial position. According to the historical re-

sults provided by Piotroski, this trading strategy is very profitable. Specifically, a trading strategy that buys expected winners and shorts expected losers would have generated a 23% annual return between 1976 and 1996.

There are different approaches to evaluate a heuristic approach. Statistical analysis can be used to estimate the probability of incorrect outcomes. Another approach is to evaluate economic significance. For example, Piotroski determines economic significance by forming portfolios based on the firm's aggregate score (F_SCORE) and then evaluates the size of the subsequent portfolio returns.

There is no theory that can provide guidance when making modeling choices in the heuristic approach. Consequently, the researcher has to be careful not to fall into the data-mining trap.

The Optimization Approach

In this approach, we use optimization to select and weight factors in a forecasting model. An *optimization approach* allows us flexibility in calibrating the model and simultaneously optimizing an objective function specifying a desirable investment criteria.

There is substantial overlap between optimization use in forecast modeling and portfolio construction. There is frequently an advantage in working with the factors directly, as opposed to all individual stocks. The factors provide a lower dimensional representation of the complete universe of the stocks considered. Besides the dimensionality reduction, which reduces computational time, the resulting optimization problem is typically more robust to changes in the inputs.

Sorensen, Hua, Qian, and Schoen (2004) present a process that uses an optimization framework to combine a diverse set of factors (alpha sources) into a multifactor model. Their procedure assigns optimal weights across the factors to achieve the highest information ratio. They show that the optimal weights are a

function of average ICs and IC covariances. Specifically,

$$\mathbf{w} \propto \text{cov}(\mathbf{IC})^{-1} \times \overline{\mathbf{IC}}$$

where \mathbf{w} is the vector of factor weights, $\overline{\mathbf{IC}}$ is the vector of the average of the risk-adjusted ICs, and $\text{cov}(\mathbf{IC})^{-1}$ is the inverse of the covariance matrix of the ICs.

In a subsequent paper, Sorensen, Hua, and Qian (2005) apply this optimization technique to capture the idiosyncratic return behavior of different security contexts. The contexts are determined as a function of stock risk characteristics (value, growth, or earnings variability). They build a multifactor model using the historical risk-adjusted IC of the factors, determining the weights of the multifactor model by maximizing the IR of the combined factors. Their research demonstrates that the weights to factors of an alpha model (trading strategy) differ depending on the security contexts (risk dimensions). The approach improves the ex post information ratio compared to a model that uses a one-size-fits-all approach.

Importance of Model Construction and Factor Choice

Empirical research shows that the factors and the weighting scheme of the factors are important in determining the efficacy of a trading strategy model. Using data from the stock selection models of 21 major quantitative funds, the quantitative research group at Sanford Bernstein analyzed the degree of overlap in rankings and factors.¹⁹ They found that the models maintained similar exposures to many of the same factors. Most models showed high exposure to cash flow-based valuations (e.g., EV/EBITDA) and price momentum, and less exposure to capital use, revisions, and normalized valuation factors. Although they found commonality in factor exposures, the stock rankings and performance of the models were substantially different. This surprising finding indicates that model construction differs among

the various stock selection models and provides evidence that the efficacy of common signals has not been completely arbitrated away.

A second study by the same group showed commonality across models among cash flow and price momentum factors, while stock rankings and realized performance were vastly different.²⁰ They hypothesize that the difference between good and poor performing models may be related to a few unique factors identified by portfolio managers, better methodologies for model construction (e.g., static, dynamic, or contextual models), or good old-fashioned luck.

Example: A Factor-Based Trading Strategy

In building this model, we hope to accomplish the following objectives: identify stocks that will outperform and underperform in the future, maintain good diversification with regard to alpha sources, and be robust to changing market conditions such as time varying returns, volatilities, and correlations.

We have identified 10 factors that have an ability to forecast stock returns.²¹ Of the four model construction methodologies discussed previously, we use the optimization framework to build the model as it offers the greatest flexibility.

We determine the allocation to specific factors by solving the following optimization problem:

$$\begin{aligned} \min_w \quad & \mathbf{w}'\Sigma\mathbf{w}, \quad \mathbf{w} \geq 0 \\ & \sum_{v \in \text{Value}} \mathbf{w}_v \geq 0.35 \\ & \sum_{g \in \text{Growth}} \mathbf{w}_g \geq 0.20 \\ & 3 \leq \sum_{i=1}^{10} \delta_i \leq 7 \end{aligned}$$

with the budget constraint

$$\mathbf{w}'\mathbf{e} = 1, \quad \mathbf{e} = (1, \dots, 1)'$$

where Σ is the covariance matrix of factor returns, Value and Growth are the sets of value and growth factors, and δ_i is equal to one if $w_i > 0$ or zero otherwise.

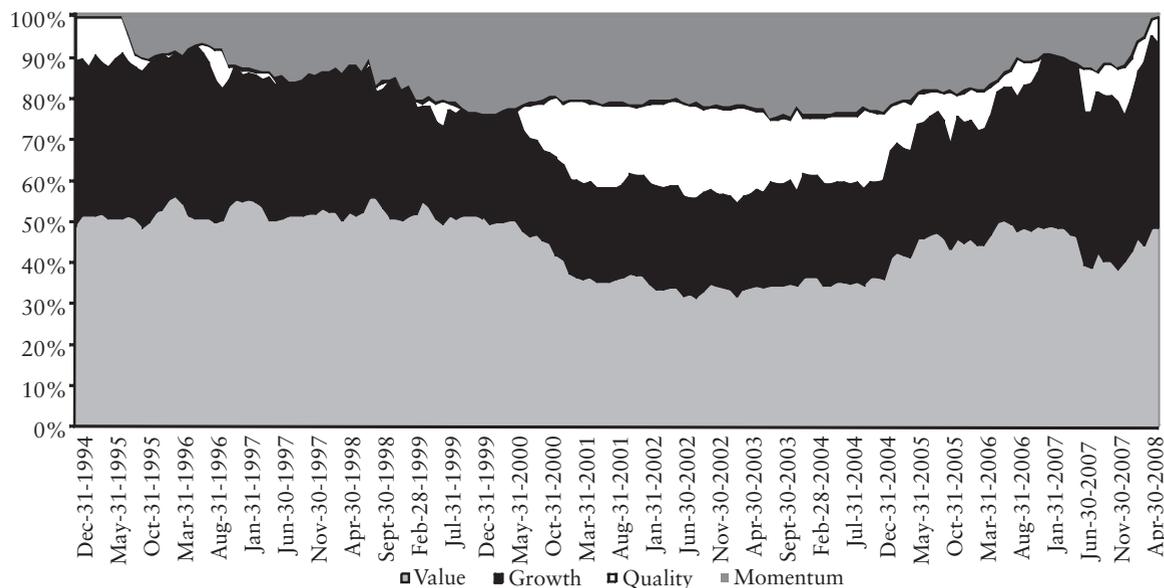


Figure 11 Factor Weights of the Trading Strategy

We constrain the minimum exposure to values factors to be greater than or equal to 35% of the weight in the model based on the belief that there is a systematic long-term premium to value.

Using the returns of our factors, we perform this optimization monthly to determine which factors to hold and in what proportions. Figure 11 displays how the factor weights change over time.

In the next step, we use the factor weights to determine the attractiveness of the stocks in our universe. We score each stock in the universe by multiplying the standardized values of the factors by the weights provided by the optimization of our factors. Stocks with high scores are deemed attractive and stocks with low scores are deemed unattractive.

To evaluate how the model performs, we sort the scores of stocks into five equally weighted portfolios and evaluate the returns of these portfolios. Table 1(A) provides summary statistics of the returns for each portfolio. Note that there is a monotonic increasing relationship among the portfolios with portfolio 1 (q1) earning the highest return and portfolio 5 (q5) earn-

ing the lowest return. Over the entire period, the long-short portfolio (LS) that is long portfolio 1 and short portfolio 5 averages about 1% per month with a monthly Sharpe ratio of 0.33. Its return is statistically significant at the 97.5% level.

Table 1 Summary of Model Results
A. Summary Statistics of the Model Returns

	q1	q2	q3	q4	q5	LS
Mean	1.06	0.98	0.83	0.65	0.12	0.94
Stdev	5.64	5.18	4.98	5.31	5.88	2.82
Median	1.61	1.61	1.58	1.55	1.11	0.71
Max	15.79	11.18	10.92	13.26	13.01	12.84
Min	-23.59	-23.32	-19.45	-21.25	-24.51	-6.87
Num	169	169	169	169	169	169
t-statistic	2.44	2.45	2.17	1.59	0.27	4.33
IR	0.19	0.19	0.17	0.12	0.02	0.33

B. Summary Statistics of Turnover for Portfolio 1 (q1) and Portfolio 5 (q5)

	q1	q5
Mean	0.20	0.17
Stdev	0.07	0.06
Median	0.19	0.16
Max	0.53	0.39
Min	0.07	0.05
Num	169	169
t-statistic	36.74	39.17

Table 1(B) shows the monthly average stock turnover of portfolio 1 (q1) and portfolio 5 (q5). Understanding how turnover varies from month to month for a trading strategy is important. If turnover is too high then it might be prohibitive to implement because of execution costs. While beyond the scope of this entry, we could explicitly incorporate transaction costs in this trading strategy using a market impact model.²² Due to the dynamic nature of our trading strategy—where active factors may change from month to month—our turnover of 20% is a bit higher than what would be expected using a static approach.

We evaluate the monthly information coefficient between the model scores and subsequent return. This analysis provides information on how well the model forecasts return. The monthly mean information coefficient of the model score is 0.03 and is statistically sig-

nificant at the 99% level. The monthly standard deviation is 0.08. We note that both the information coefficients and returns were stronger and more consistent in the earlier periods.

Figure 12 displays the cumulative return to portfolio 1 through portfolio 5. Throughout the entire period there is a monotonic relationship between the portfolios. To evaluate the overall performance of the model, we analyze the performance of the long-short portfolio returns. We observe that the model performs well in December 1994 to May 2007 and April 2008 to June 2008. This is due to the fact that our model correctly picked the factors that performed well in those periods. We note that the model performs poorly in the period July 2007–April 2008, losing an average of 1.09% a month. The model appears to suffer from the same problems many quantitative equity funds and hedge funds faced during this period.²³

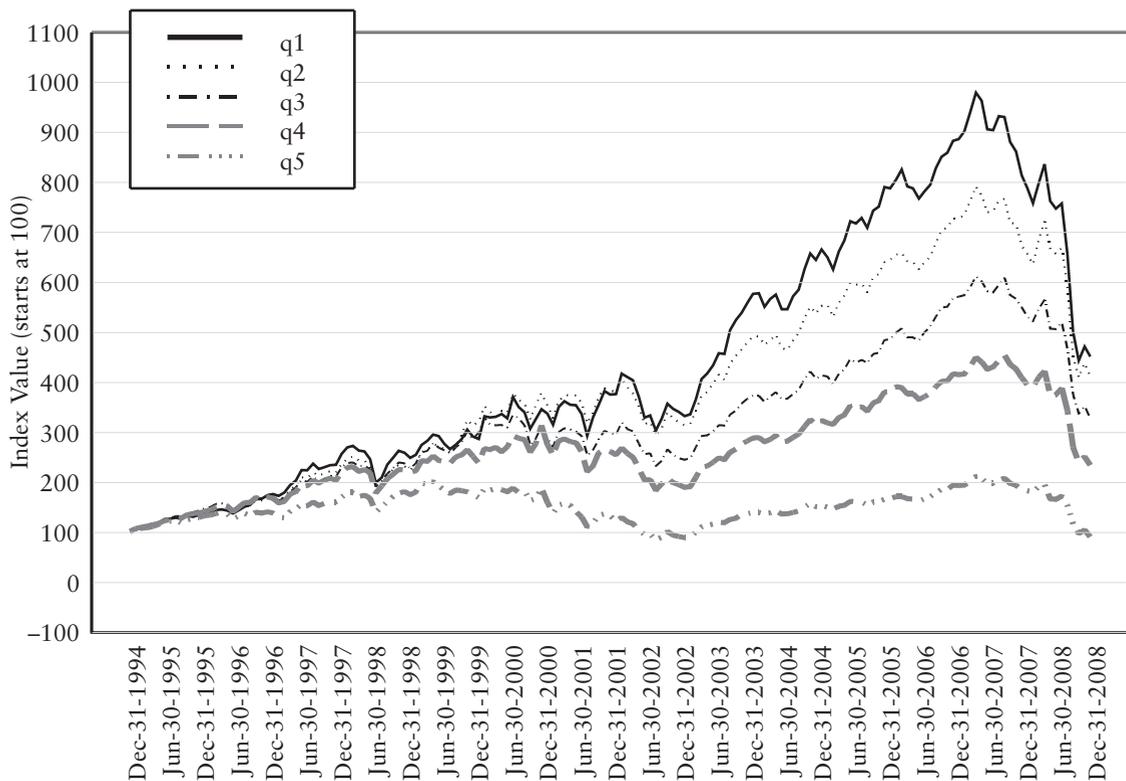


Figure 12 Cumulative Return of the Model

The worst performance in a single month was -6.87 , occurring in January 2001, and the maximum drawdown of the model was -13.7% , occurring during the period from May 2006 (peak) to June 2008 (trough).²⁴

To more completely understand the return and risk characteristic of the strategy, we would have to perform a more detailed analysis, including risk and performance attribution, and model sensitivity analysis over the full period as well as over subperiods. As the turnover is on the higher side, we may also want to introduce turnover constraints or use a market impact model.

Periods of poor performance of a strategy should be disconcerting to any analyst. The poor performance of the model during the period June 2007–March 2008 indicates that many of the factors we use were not working. We need to go back to each individual factor and analyze them in isolation over this time frame. In addition, this highlights the importance of research to improve existing factors and develop new ones using unique data sources.

BACKTESTING

In the research phase of the trading strategy, model scores are converted into portfolios and then examined to assess how these portfolios perform over time. This process is referred to as *backtesting a strategy*. The backtest should mirror as closely as possible the actual investing environment incorporating both the investment's objectives and the trading environment.

When it comes to mimicking the trading environment in backtests, special attention needs to be given to transaction costs and liquidity considerations. The inclusion of transaction costs is important because they may have a major impact on the total return. Realistic market impact and trading costs estimates affect what securities are chosen during portfolio construction. Liquidity is another attribute that needs to be evaluated. The investable universe of

stocks should be limited to stocks where there is enough liquidity to be able to get in and out of positions.

Portfolio managers may use a number of constraints during portfolio construction. Frequently these constraints are derived from the portfolio policy of the firm, risk management policy, or investor objectives. Common constraints include upper and lower bounds for each stock, industry, or risk factor—as well as holding size limits, trading size limits, turnover, and the number of assets long or short.

To ensure the portfolio construction process is robust we use sensitivity analysis to evaluate our results. In sensitivity analysis we vary the different input parameters and study their impact on the output parameters. If small changes in inputs give rise to large changes in outputs, our process may not be robust enough. For example, we may eliminate the five best and worst performing stocks from the model, rerun the optimization, and evaluate the performance. The results should be similar as the success of a trading strategy should not depend on a handful of stocks.

We may want to determine the effect of small changes in one or more parameters used in the optimization. The performance of the optimal portfolio should in general not differ significantly after we have made these small changes.

Another useful test is to evaluate a model by varying the investment objective. For example, we may evaluate a model by building a low-tracking-error portfolio, a high-tracking-error portfolio, and a market-neutral portfolio. If the returns from each of these portfolios are decent, the underlying trading strategy is more likely to be robust.

Understanding In-Sample and Out-of-Sample Methodologies

There are two basic backtesting methodologies: in-sample and out-of-sample. It is important to understand the nuances of each.

We refer to a backtesting methodology as an in-sample methodology when the researcher uses the same data sample to specify, calibrate, and evaluate a model.

An out-of-sample methodology is a backtesting methodology where the researcher uses a subset of the sample to specify and calibrate a model, and then evaluates the forecasting ability of the model on a different subset of data. There are two approaches for implementing an out-of-sample methodology. One approach is the split-sample method. This method splits the data into two subsets of data where one subset is used to build the model while the remaining subset is used to evaluate the model.

A second method is the recursive out-of-sample test. This approach uses a sequence of recursive or rolling windows of past history to forecast a future value and then evaluates that value against the realized value. For example, in a rolling regression-based model we will use data up to time t to calculate the coefficients in the regression model. The regression model forecasts the $t + h$ dependent values, where $h > 0$. The prediction error is the difference between the realized value at $t + h$ and the predicted value from the regression model. At $t + 1$ we recalculate the regression model and evaluate the predicted value of $t + 1 + h$ against realized value. We continue this process throughout the sample.

The conventional thinking among econometricians is that in-sample tests tend to reject the null hypotheses of no predictability more often than out-of-sample tests. This view is supported by many researchers because they reason that in-sample tests are unreliable, often finding spurious predictability. Two reasons given to support this view are the presence of unmodeled structural changes in the data and the use of techniques that result in data mining and model overfitting.

Inoue and Kilian (2002) question this conventional thinking. They use asymptotic theory to evaluate the “trade-offs between in-sample tests and out-of-sample tests of predictability

in terms of their size and power.” They argue strong in-sample results and weak out-of-sample results are not necessarily evidence that in-sample tests are not reliable. Out-of-sample tests using sample-splitting result in a loss of information and lower power for small samples. As a result, an out-of-sample test may fail to detect predictability while the in-sample test will correctly identify predictability. They also show that out-of-sample tests are not more robust to parameter instability that results from unmodeled structural changes.

A Comment on the Interaction between Factor-Based Strategies and Risk Models

Frequently, different factor models are used to calculate the risk inputs and the expected return forecasts in a portfolio optimization. A common concern is the interaction between factors in the models for risk and expected returns. Lee and Stefek (2008) evaluate the consequences of using different factor models, and conclude that (1) using different models for risk and alpha can lead to unintended portfolio exposures that may worsen performance; (2) aligning risk factors with alpha factors may improve information ratios; and (3) modifying the risk model by including some of the alpha factors may mitigate the problem.

BACKTESTING OUR FACTOR TRADING STRATEGY

Using the model scores from the trading strategy example, we build two optimized portfolios and evaluate their performance. Unlike the five equally weighted portfolios built only from model scores, the models we now discuss were built to mirror as close as possible tradable portfolios a portfolio manager would build in real time. Our investable universe is the Russell 1000. We assign alphas for all stock in the Russell 1000 with our dynamic factor model.

Table 2 Total Return Report (annualized)

From 01/1995 to 06/2008	QTD	YTD	1 Year	2 Year	3 Year	5 Year	10 Year	Since Inception
Portfolio: Low-tracking error	-0.86	-10.46	-11.86	4.64	7.73	11.47	6.22	13.30
Portfolio: High-tracking error	-1.43	-10.47	-11.78	4.15	8.29	13.24	7.16	14.35
S&P 500: Total return	-2.73	-11.91	-13.12	2.36	4.41	7.58	2.88	9.79

The portfolios are long only and benchmarked to the S&P 500. The difference between the portfolios is in their benchmark tracking error. For the low-tracking error portfolio the risk aversion in the optimizer is set to a high value, sectors are constrained to plus or minus 10% of the sector weightings in the benchmark, and portfolio beta is constrained to 1.00. For the high-tracking error portfolio, the risk aversion is set to a low value, the sectors are constrained to plus or minus 25% of the sector weightings in the benchmark, and portfolio beta is constrained to 1.00. Rebalancing is performed once a month. Monthly turnover is limited to 10% of the portfolio value for the low-tracking error portfolio and 15% of the portfolio value for the high-tracking error portfolio.

Table 2 presents the results of our backtest. The performance numbers are gross of fees and transaction costs. Performance over the entire period is good and consistent throughout. The portfolios outperform the benchmark over the various time periods. The resulting annualized Sharpe ratios over the full period are 0.66 for the low-tracking error portfolio, 0.72 for the high-tracking error portfolio, and 0.45 for the S&P 500.²⁵

KEY POINTS

- The four most commonly used approaches for the evaluation of return premiums and risk characteristics to factors are portfolio sorts, factor models, factor portfolios, and information coefficients.
- The portfolio sorts approach ranks stocks by a particular factor into a number of portfolios. The sorting methodology should be consis-

tent with the characteristics of the distribution of the factor and the economic motivation underlying its premium.

- The information ratio (IR) is a statistic for summarizing the risk-adjusted performance of an investment strategy and is defined as the ratio of average excess return to the standard deviation of return.
- We distinguish between contemporaneous and forecasting factor models, dependent on whether both left- and right-hand side variables (returns and factors) have the same time subscript, or the time subscript of the left-hand side variable is greater.
- The three most common violations of classical regression theory that occur in *cross-sectional factor models* are (1) the errors in variables problem, (2) common variation in residuals such as heteroskedasticity and serial correlation, and (3) multicollinearity. There are statistical techniques that address the first two. The third issue is best dealt with by removing collinear variables from the regression, or by increasing the sample size.
- The Fama-MacBeth regression addresses the inference problem caused by the correlation of the residuals in cross-sectional regressions.
- The information coefficient (IC) is used to evaluate the return forecast ability of a factor. It measures the cross-sectional correlation between a factor and its subsequent realized return.
- Factor portfolios are used to measure the information content of a factor. The objective is to mimic the return behavior of a factor and minimize the residual risk. We can build factor portfolios using a factor model or an optimization. An optimization is more flexible as it is able to incorporate constraints.

- Analyzing the performance of different factors is an important part of the development of a factor-based trading strategy. This process begins with understanding the time-series properties of each factor in isolation and then studying how they interact with each other.
- Techniques used to combine and weight factors to build a trading strategy model include the data driven, the factor model, the heuristic, and the optimization approaches.
- An out-of-sample methodology is a backtesting methodology where the researcher uses a subset of the sample to specify a model and then evaluates the forecasting ability of the model on a different subset of data. There are two approaches for implementing an out-of-sample methodology: the split-sample approach and the recursive out-of-sample test.
- Caution should be exercised if different factor models are used to calculate the risk inputs and the expected return forecasts in a portfolio optimization.

APPENDIX: THE COMPUSTAT POINT-IN-TIME, IBES CONSENSUS DATABASES AND FACTOR DEFINITIONS

The factors used in this entry were constructed on a monthly basis with data from the Compustat Point-In-Time and IBES Consensus databases. Our sample includes the largest 1,000 stocks by market capitalization over the period December 31, 1989, to December 31, 2008.

The Compustat Point-In-Time database (Capital IQ, Compustat, <http://www.compustat.com>) contains quarterly financial data from the income, balance sheet, and cash flow statements for active and inactive companies. This database provides a consistent view of historical financial data, both reported data and subsequent restatements, the way it appeared at the

end of any month. Using these data allows the researcher to avoid common data issues such as survivorship and look-ahead bias. The data are available from March 1987.

The Institutional Brokers Estimate System (IBES) database (Thomson Reuters, <http://www.thomsonreuters.com>) provides actual earnings from companies and estimates of various financial measures from sell-side analysts. The estimated financial measures include estimates of earnings, revenue and sales, operating profit, analyst recommendations, and other measures. The data are offered on a summary (consensus) level or detailed (analyst-by-analyst) basis. The U.S. data cover reported earnings estimates and results since January 1976.

The factors used in this entry are defined as follows. (LTM refers to the last four reported quarters.)

Value Factors

Operating income before depreciation to enterprise value = EBITDA/EV
where

$$\begin{aligned} \text{EBITDA} = & \text{Sales LTM (Compustat Item 2)} \\ & - \text{Cost of goods Sales LTM} \\ & \quad \text{(Compustat Item 30)} \\ & - \text{SG\&A Exp (Compustat Item 1)} \end{aligned}$$

and

$$\begin{aligned} \text{EV} = & [\text{Long-term debt (Compustat} \\ & \quad \text{Item 51)} \\ & + \text{Common shares outstanding} \\ & \quad \text{(Compustat Item 61)} \\ & \times \text{Price (PRCCM)} - \text{Cash} \\ & \quad \text{(Compustat Item 36)}] \end{aligned}$$

$$\begin{aligned} \text{Book to price} = & \text{Stockholders' equity total} \\ & \quad \text{(Compustat Item 60)} \\ & \div [\text{Common shares outstanding} \\ & \quad \text{(Compustat Item 59)} \\ & \times \text{Price (PRCCM)}] \end{aligned}$$

$$\begin{aligned} \text{Sales to price} &= \text{Sales LTM (Computstat Item 2)} \\ &\div [\text{Common shares outstanding} \\ &\quad \text{(Computstat Item 61)} \\ &\quad \times \text{Price(PRCCM)}] \end{aligned}$$

Quality Factors

$$\begin{aligned} \text{Share repurchase} &= [\text{Common shares} \\ &\quad \text{outstanding (Computstat Item 61)} - \text{Common} \\ &\quad \text{shares outstanding (Computstat Item 61)} \\ &\quad \text{from 12 months ago}] \div \text{Common shares} \\ &\quad \text{outstanding (Computstat Item 61) from} \\ &\quad \text{12 months ago} \end{aligned}$$

$$\begin{aligned} \text{Asset turnover} &= \text{Sales LTM (Computstat Item 2)} / \\ &\quad [(\text{Assets (Computstat Item 44)} \\ &\quad - \text{Assets (Computstat Item 44)} \\ &\quad \text{from 12 months ago}) / 2] \end{aligned}$$

$$\begin{aligned} \text{Return on invested capital} &= \text{Income} / \\ &\quad \text{Invested capital} \end{aligned}$$

where

$$\begin{aligned} \text{Income} &= \text{Income before extra items LTM} \\ &\quad \text{(Computstat Item 8)} \\ &\quad + \text{Interest expense LTM} \\ &\quad \text{(Computstat Item 22)} \\ &\quad + \text{Minority interest expense LTM} \\ &\quad \text{(Computstat Item 3)} \end{aligned}$$

and

$$\begin{aligned} \text{Invested capital} \\ &= \text{Common equity (Computstat Item 59)} \\ &\quad + \text{Long-term debt (Computstat Item 51)} \\ &\quad + \text{Minority interest (Computstat Item 53)} \\ &\quad + \text{Preferred stock (Computstat Item 55)} \\ \text{Debt to equity} &= \text{Total debt} / \text{Stockholders' equity} \end{aligned}$$

where

$$\begin{aligned} \text{Total debt} &= [\text{Debt in current liabilities} \\ &\quad \text{(Computstat Item 45)} + \text{Long-term debt} \\ &\quad - \text{Total (Computstat Item 51)}] \end{aligned}$$

$$\begin{aligned} \text{and} \\ \text{Stockholders' equity} &= \text{Stockholders' equity} \\ &\quad \text{(Computstat Item 60)} \end{aligned}$$

$$\begin{aligned} \text{Chg. debt to equity} &= (\text{Total debt} - \text{Total debt} \\ &\quad \text{from 12 months ago}) \\ &\quad \div [(\text{Stockholders' equity} \\ &\quad + \text{Stockholders' equity} \\ &\quad \text{from 12 months ago}) / 2] \end{aligned}$$

Growth

$$\begin{aligned} \text{Revisions} &= [\text{Number of up revisions} \\ &\quad \text{(IBES item NUMUP)} \\ &\quad - \text{Number of down revisions (IBES} \\ &\quad \quad \text{item NUMDOWN)}] \\ &\quad \div \text{Number of estimates revisions} \\ &\quad \quad \text{(IBES item NUMEST)} \end{aligned}$$

$$\begin{aligned} \text{Growth of fiscal Year 1 and fiscal Year 2} \\ \text{earnings estimates} &= \text{Consensus mean of FY2} \\ &\quad \text{(IBES item MEANFY2)} \div \text{Consensus mean of} \\ &\quad \text{FY 1 (IBES item MEAN FY1)} - 1 \end{aligned}$$

Momentum

$$\begin{aligned} \text{Momentum} &= \text{Total return of last 11 months} \\ &\quad \text{excluding the most returns from} \\ &\quad \text{the most recent month} \end{aligned}$$

Summary Statistics

The following table contains monthly summary statistics of the factors defined previously. Factor values greater than the 97.5 percentile or less than the 2.5 percentile are considered outliers. We set factor values greater than the 97.5 percentile value to the 97.5 percentile value, and factor values less than the 2.5 percentile value to the 2.5 percentile value, respectively.

	Mean	Standard Deviation	Median	25 Percentile	75 Percentile
EBITDA/EV	0.11	0.06	0.11	0.07	0.15
Book to price	0.46	0.30	0.40	0.24	0.62
Sales to price	0.98	0.91	0.69	0.36	1.25
Share repurchase	0.03	0.09	0.00	-0.01	0.03
Asset turnover	1.83	1.89	1.46	0.64	2.56
Return on invested capital	0.13	0.11	0.11	0.07	0.17
Debt to equity	0.97	1.08	0.62	0.22	1.26
Change in debt to equity	0.10	0.31	0.01	-0.04	0.17
Revisions	-0.02	0.33	0.00	-0.17	0.11
Growth of fiscal year 1 and fiscal year 2 earnings estimates	0.37	3.46	0.15	0.09	0.24
Momentum	13.86	36.03	11.00	-7.96	31.25

NOTES

- For a good overview of the most common issues, see Berk (2000) and references therein.
- Grinold and Kahn (1999) discuss the differences between the t -statistic and the information ratio. Both measures are closely related in their calculation. The t -statistic is the ratio of mean return of a strategy to its standard error. Grinold and Kahn state the related calculations should not obscure the distinction between the two ratios. The t -statistic measures the statistical significance of returns while the IR measures the risk-reward trade-off and the value added by an investment strategy.
- See, for example, Fama and French (2004).
- One approach is to use the Bayesian or model averaging techniques. For more details on the Bayesian approach, see, for example, Rachev, Hsu, Bagasheva, and Fabozzi (2008).
- For a discussion of dealing with these econometric problems, see Chapter 2 in Fabozzi, Focardi, and Kolm (2010).
- We cover Fama-MacBeth regression in this section.
- Fama and French (2004).
- See, for example, Grinold and Kahn (1999) and Qian, Hua, and Sorensen (2007).
- A factor normalized z -score is given by the formula $z\text{-score} = (\mathbf{f} - \bar{\mathbf{f}})/\text{std}(\mathbf{f})$ where \mathbf{f} is the factor, $\bar{\mathbf{f}}$ is the mean, and $\text{std}(\mathbf{f})$ is the standard deviation of the factor.
- We are conforming to the notation used in Qian and Hua (2004). To avoid confusion, Qian and Hua use $\text{dis}()$ to describe the cross-sectional standard deviation and $\text{std}()$ to describe the time series standard deviation.
- The earnings estimates come from the IBES database. See the appendix for a more detailed description of the data.
- This derivation of factor portfolios is presented in Grinold and Kahn (1999).
- See Melas, Suryanarayanan, and Cavaglia (2009).
- An exception is the constraint on the number of assets that results in integer constraints.
- For a more detailed discussion on portfolio optimization problems and optimization software see, for example, Fabozzi, Kolm, Pachamanova, and Focardi (2007).
- The hit rate is calculated as

$$h = \frac{1}{T_2 - T_1} \sum_{t=T_1+1}^{T_2} \text{sign}(y_t^i \hat{y}_{t-1}^i)$$
 where y_t^i is *one-step ahead* realized value and \hat{y}_{t-1}^i is the *one-step ahead* predicted value.
- For calculation of this measure, see Diebold and Mariano (2005).
- The nine factors are return on assets, change in return on assets, cash flow from

operations scaled by total assets, cash compared to net income scaled by total assets, change in long-term debt/assets, change in current ratio, change in shares outstanding, change in gross margin, and change in asset turnover.

19. Zlotnikov, Larson, Cheung, Kalaycioglu, Lao, and Apoian (2007).
20. Zlotnikov, Larson, Wally, Kalaycioglu, Lao, and Apoian (2007).
21. We use a combination of growth, value, quality, and momentum factors. The appendix to this entry contains definitions of all of them.
22. Cerniglia and Kolm (2010).
23. See Rothman (2007) and Daniel (2007).
24. We ran additional analysis on the model by extending the holding period of the model from 1 to 3 months. The results were much stronger as returns increased to 1.6% per month for a two-month holding period and 1.9% per month for a three-month holding period. The risk as measured by drawdown was higher at -17.4% for a two-month holding period and -29.5% for the three-month holding period.
25. Here we calculate the Sharpe ratio as portfolio excess return (over the risk-free rate) divided by the standard deviation of the portfolio excess return.

REFERENCES

- Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59, 3: 817–858.
- Berk, J. B. (2000). Sorting out sorts. *Journal of Finance* 55, 1: 407–427.
- Cerniglia, J. A., and Kolm, P. N. (2010). Factor-based trading strategies and market impact costs. Working paper, Courant Institute of Mathematical Sciences, New York University.
- Cochrane, J. H. (2005). *Asset Pricing*. Princeton, NJ: Princeton University Press.
- Daniel, K. (2001). The liquidity crunch in quant equities: Analysis and implications. Goldman Sachs Asset Management, December 13, 2007, presentation from The Second New York Fed–Princeton Liquidity Conference.
- Deistler, M., and Hamann, E. (2005). Identification of factor models for forecasting returns. *Journal of Financial Econometrics* 3, 2: 256–281.
- Diebold, F. X., and Mariano, R. S. (2005). Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13, 3: 253–263.
- Fabozzi, F. J., Focardi, S. M., and Kolm, P. N. (2010). *Quantitative Equity Investing*. Hoboken, NJ: John Wiley & Sons.
- Fabozzi, F. J., Kolm, P. N., Pachamanova, D., and Focardi, S. M. (2007). *Robust Portfolio Optimization and Management*. Hoboken, NJ: John Wiley & Sons.
- Fama, E. F., and French, K. R. (2004). The capital asset pricing model: Theory and evidence. *Journal of Economic Perspectives* 18, 3: 25–46.
- Fama, E. F., and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy* 81, 3: 607–636.
- Gow, I. D., Ormazabal, G., and Taylor, D. J. (2009). Correcting for cross-sectional and time-series dependence in accounting research. Working paper, Kellogg School of Business and Stanford Graduate School.
- Grinold, R. C., and Kahn, R. N. (1999). *Active Portfolio Management: A Quantitative Approach for Providing Superior Returns and Controlling Risk*. New York: McGraw-Hill.
- Inoune, A., and Kilian, L. (2002). In-sample or out-of-sample tests of predictability: Which one should we use? Working paper, North Carolina State University and University of Michigan.
- Hua, R., and Qian, E. (2004). Active risk and information ratio. *Journal of Investment Management* 2, 3: 1–15.
- Kolm, P. N. (2010). Multi-period portfolio optimization with transaction costs, alpha decay, and constraints. Working paper, Courant Institute of Mathematical Sciences, New York University.
- Lee, J.-H., and Stefek, D. (2008). Do risk factors eat alphas? *Journal of Portfolio Management* 34, 4: 12–24.
- Melas, D., Suryanarayanan, R., and Cavaglia, S. (2009). Efficient replication of factor returns. *MSCI Barra Research Insight*, June.
- Newey, W. K., and West, K. D. (1987). A simple, positive semidefinite heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 56, 3: 703–708.
- Patton, A. J., and Timmermann, A. (2009). Monotonicity in asset returns: New tests with

- applications to the term structure, the CAPM and portfolio sorts. Working paper, University of California–San Diego.
- Petersen, M. A. (2009). Estimating standard errors in finance panel sets: Comparing approaches. *Review of Financial Studies* 22, 1: 435–480.
- Piotroski, J. D. (2000). Value investing: The use of historical financial statement information to separate winners from losers. *Journal of Accounting Research* 38 (3 supplement): 1–41.
- Qian, E. E., Hua, R. H., and Sorensen, E. H. (2007). *Quantitative Portfolio Management: Modern Techniques and Applications*. New York: Chapman & Hall/CRC.
- Rachev, S. T., Hsu, J. S. J., Bagasheva, B. S., and Fabozzi, F. J. (2008). *Bayesian Methods in Finance*. Hoboken, NJ: John Wiley & Sons.
- Rothman, M. S. (2007). Turbulent times in quant land. Lehman Brothers Equity Research, August 9.
- Sneddon, L. (2008). The tortoise and the hare: Portfolio dynamics for active managers. *Journal of Investing* 17, 4: 106–111.
- Sorensen, E. H., Hua, R., and Qian, E. (2005). Contextual fundamentals, models, and active management. *Journal of Portfolio Management* 32, 1: 23–36.
- Sorensen, E. H., Hua, R., Qian, E., and Schoen, R. (2004). Multiple alpha sources and active management. *Journal of Portfolio Management* 30, 2: 39–45.
- West, K. D. (2006). Forecast evaluation. In G. Elliot, C. W. J. Granger, and A. G. Timmermann (Eds.), *Handbook of Economic Forecasting*, Vol.1. Amsterdam: Elsevier.
- Zlotnikov, V., Larson, A. M., Cheung, W., Kalaycioglu, S., Lao, R. D., and Apoian, Z. A. (2007). Quantitative research: Survey of quantitative models—Vastly different rankings and performance, despite similarity in factor exposures. Quantitative Research, Bernstein Research, January 16.
- Zlotnikov, V., Larson, A. M., Kalaycioglu, S., Lao, R. D., and Apoian, Z. A. (2007). Quantitative research: Survey of quantitative models—Continued emphasis on EV/EBIT, momentum, increased focus on capital use; some evidence on non-linear factor implementation; low return consistency. Quantitative Research, Bernstein Research, November 21.

The Fundamentals of Fundamental Factor Models

JENNIFER BENDER, PhD
Vice President, MSCI

FRANK NIELSEN, CFA
Managing Director of Quantitative Research, Fidelity Investments - Global Asset Allocation

Abstract: Fundamental factor risk models have been used in equity portfolio management and risk management for decades now. There persists, however, the notion that fundamental factor models are “quantitative” models that are divorced from fundamental analysis, the realm of traditional equity analysts. This perception is inaccurate in that the basic building blocks of analysts and factor modelers are in fact similar; both try to identify microeconomic traits that drive the risk and returns of individual securities. The differences between fundamental factor models and fundamental analysis lie not in their ideology but in their objectives. The objective of the fundamental analyst is to forecast return (or future stock values) for a particular stock. The objective of the fundamental factor model is to forecast the fluctuation of a portfolio around its expected return. Most importantly, the factor model captures the interaction of the firm’s microeconomic characteristics at the portfolio level. This is important because as names are added to the portfolio, company-specific returns are diversified away, and the common factor (systematic) portion becomes an increasingly larger part of the portfolio risk and return. Fundamental factor models are in fact complementary as opposed to antithetical to traditional security analysis.

Fundamental analysis is the process of determining a security’s future value by analyzing a combination of macro- and microeconomic events and company-specific characteristics. Though fundamental analysis focuses on the valuation of individual companies, most institutional investors recognize that there are common factors affecting all stocks. (Common factors are shared characteristics between firms that affect their

returns.) For example, macroeconomic events, like sudden changes in interest rate, inflation, or exchange rate expectations, can affect all stocks to varying degrees, depending on the stock’s characteristics.

Barr Rosenberg and Vinay Marathe (1976) developed the theory that the effects of macroeconomic events on individual securities could be captured through microeconomic

characteristics—essentially common factors, such as industry membership, financial structure, or growth orientation.

Rosenberg and Marathe (1976, p. 3) discuss possible effects of a money market crisis. They say a money market crisis would:

result in possible bankruptcy for some firms, dislocation of the commercial paper market, and a dearth of new bank lending beyond existing commitments. Firms with high financial risk (shown in extreme leverage, poor coverage of fixed charges, and inadequate liquid assets) might be driven to bankruptcy. Almost all firms would suffer to some degree from higher borrowing costs and worsened economic expectations: Firms with high financial risk would be impacted most; the market portfolio, which is a weighted average of all firms, would be somewhat less exposed; and firms with abnormally low financial risk would suffer the least. Moreover, some industries such as construction would suffer greatly because of their special exposure to interest rates. Others such as liquor might be unaffected.

This early insight into the linkage between macroeconomic events and microeconomic characteristics has had a profound impact on the asset management industry ever since. In this entry, we discuss the intuition behind a fundamental factor model based on microeconomic traits, showing how it is linked to traditional fundamental analysis. When building a fundamental factor model, we look for variables that explain return, just as fundamental analysts do. We highlight the complementary role of the fundamental factor model to traditional security analysis and point out the insights these models can provide.

FUNDAMENTAL ANALYSIS AND THE BARRA FUNDAMENTAL FACTOR MODEL

Fundamental analysts use many criteria when researching companies; they may investigate a firm's financial statements, talk to senior management, visit facilities and plants, or analyze a product pipeline. Most are seeking under-

Table 1 Main Areas of Stock Research

Qualitative	Quantitative
Business Model	Capital Structure
Competitive Advantage	Revenue, Expenses, and Earnings Growth
Management Quality	Cash Flows
Corporate Governance	

Note: Balance sheet and income statement data are readily available from 10K filings while access to company management and information about the business model and competitor landscape will vary on a case-by-case basis.

valued companies with good fundamentals or companies with strong growth potential. They typically review a range of quantitative and qualitative information to help predict future stock values. Table 1 summarizes key areas.

Similarly, the goal of a fundamental factor model is to identify traits that are important in forecasting security risk. These models may analyze microeconomic characteristics, such as industry membership, earnings growth, cash flow, debt-to-assets, and company specific traits. Figure 1 shows the cumulative returns to Merck, GlaxoSmithKline, and Bristol-Myers, three of the largest pharmaceutical companies in the United States. The chart illustrates the similarities in the return behavior of these stocks, primarily because they are U.S. large-cap equities within the same industry. We also see that Bristol-Myers underperformed the other two companies in recent years, indicating that other firm-specific factors also impacted its performance.

The first task when building a fundamental factor model is to identify microeconomic traits. These include characteristics from industry membership and financial ratios to technical indicators like price momentum and recent volatility that explain return variation across a relevant security universe. The next step is to determine the impact certain events may have on individual stocks, such as the sensitivity or weight of an individual security to a change in a given fundamental factor.¹ Finally, the remaining part of the returns needs to be

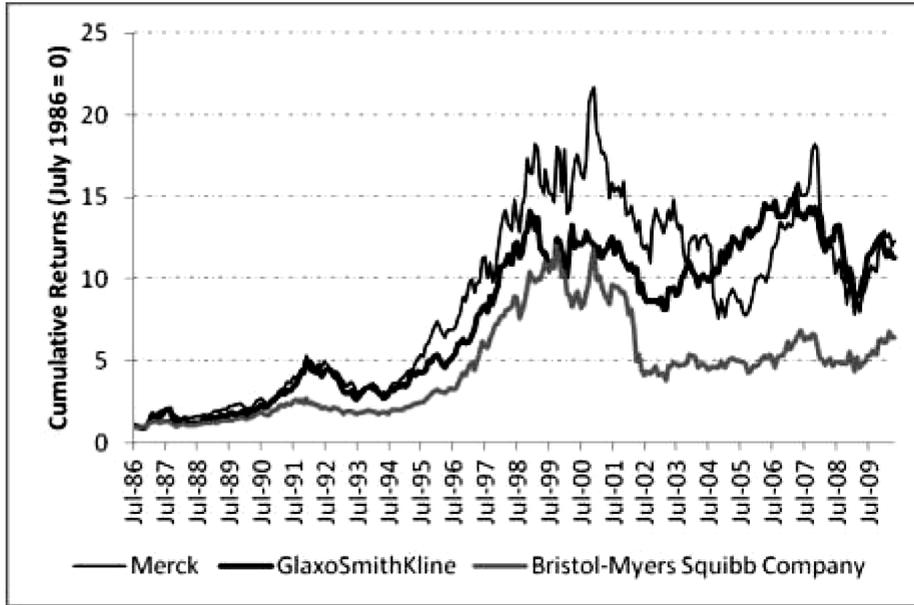


Figure 1 Industry Membership Drives Similarities between Stocks

modeled, which is the company-specific behavior of stocks.

How does the model we have described compare with the way a fundamental analyst or portfolio manager analyzes stocks? The basic building blocks of analysts and factor modelers are in fact similar; both try to identify microeconomic traits that drive the risk and returns of individual securities. Figure 2 compares the two perspectives. In both views, there are clearly firm-specific traits driving risk and return. There are also sources of risk and return from a stock’s exposure, or beta, to the overall market, its industry, and certain financial and technical ratios. But the objective of the fundamental analyst is to forecast return (or future stock values), whereas the fundamental factor model forecasts the fluctuation of a security or a portfolio of securities around its expected return.

Both the analyst and the factor model researcher look at similar macro- and microeconomic data and events. After all, both are seeking traits that explain the risk and returns of stocks. Table 2 shows examples used

in the Barra equity models (specifically the U.S. and Japan Equity Models). Variables like profitability and debt loads are incorporated in our models through factors like Earnings Yield, Growth, and Leverage. Expectations of

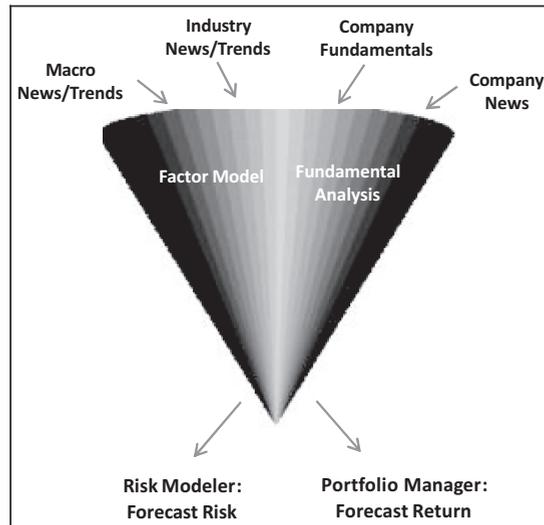


Figure 2 Overview of Stock Determinants: Fundamental Analysis versus Factor Model Analysis

Table 2 Sample Fundamental Data Used in Barra Models

Value	Growth	Earnings Variation	Leverage	Foreign Sensitivity
Book value	Five-year payout	Variability in earnings	Market leverage	Exchange rate sensitivity
Analyst predicted earnings	Variability in capital structure	Standard deviation of analyst-predicted earnings	Book leverage	Oil price sensitivity
Trailing earnings	Growth in total assets	Variability in cash flows	Debt to assets	Sensitivity to other market indices
Forecast operating income	Growth in revenues	Extraordinary items in earnings	Senior debt rating	Export revenue as percentage of total
Sales	Pension liabilities			
Forecast sales	Historical earnings growth			
	Analyst-predicted earnings growth			
	Recent earnings changes			

future revenue growth and cost savings are incorporated through variables like analyst consensus views. What about popular metrics that aren't included? Some factors may not be good risk factors despite being good return factors. (A good return factor has persistent direction though not a lot of volatility; the ability of a company to beat earnings estimates is one of these factors). Other factors may be relevant only for a particular sector or industry. (These risk factors would only be included in industry or sector risk models.)

Note that adjustments of financial statements are incorporated in several ways. A key task for the fundamental analyst is to adjust financial statements—each analyst wants to get at the “real” number rather than what is reported in financial statements. Even under generally accepted accounting principles, management can be aggressive with basic principles, such as revenue/expense recognition; usage of unusual, infrequent, or extraordinary items; and timing issues that may lead to violations of the matching principle. In a factor model, these types of adjustments are accounted for through the inclusion of forward-looking, analyst-derived descriptors.

In addition to fundamental data, technical variables such as price momentum, beta, option-implied volatility, and so on may also be used. For instance, price momentum has been shown to significantly explain returns (see Carhart, 1997).

How are the fundamental data used in a factor model? Certain factors are found to explain stock returns over time, for example, industries and certain financial and technical ratios. If such factors explain returns across a broad universe of stocks, they are deemed important. In financial theory, these factors are “priced” across assets, for example, Fama-French value (BTM) size (small-cap) factors (Fama and French, 1992).

Once we have identified the factors, we need to link each stock to each factor. For this, we use microeconomic characteristics. We start by identifying a set of characteristics we call descriptors. For instance, if the factor is growth, a few descriptors might include earnings growth, revenue growth, and asset growth (see Table 2). These include both historical and forward-looking descriptors, such as forecast earnings growth. After we identify the important descriptors, we standardize them

Table 3 Calculating Exposures from Raw Data (April 1, 2010)

Barra Factor	Size	Value	Yield
Descriptor for Factor	Capitalization (USD Bn)*	Book to Price	Predicted Dividend Yield
Microsoft	256.7	0.15	0.02
Estimation Universe Average	69.8	0.39	0.02
Estimation Universe Std Dev	21.1	0.37	0.02
Exposure	1.64	-0.62	0.06

Note: The actual descriptor for the USE3 Size factor uses the log of market capitalization. The log of market cap for Microsoft is 12.46. The estimation universe average is 10.22 and the standard deviation 1.36. The resulting exposure for Microsoft is 1.64.

across a universe of stocks, typically the constituents of a broad market index.² Table 3 illustrates how Microsoft's exposures for the Barra U.S. factors—size, value, and yield—are calculated. We subtract the estimation universe average³ descriptor for each factor and divide it by the standard deviation of the universe of stocks.

In some cases, factors reflect several characteristics. This occurs when multiple descriptors help explain the same factor. The Barra U.S. Growth factor, for instance, reflects five-year payouts, variability in capital structure, growth in total assets, recent large earnings changes, and forecast and historical earnings growth. Table 4 shows how we calculate Microsoft's exposure to the Growth factor. Each descriptor is first standardized and then the descriptors are combined to form the exposure.

In addition to factors like Value, Size, Yield, and Growth, which we call *style* factors, a stock's returns are also a function of its industry. Industry exposures are calculated in a different way. A company like Google, for instance, is engaged solely in Internet-related activities. It has an exposure of 100% (1.0) to the Internet industry factor in the Barra U.S. Equity Model. Its exposure to all other industry factors is zero. Some companies, like General Electric, have business activities that span multiple industries. In the U.S. model, industry exposures are based on sales, assets, and operating income in each industry.⁴

What does a factor exposure mean? In the same way the classic capital asset pricing model beta measures how much a stock price moves with every percentage change in the market, a factor exposure measures how much a stock

Table 4 Calculating Exposures When There Are Multiple Characteristics (April 1, 2010)

Factor Descriptor	Growth					
	Growth Rate of Total Assets	Recent Earnings Change	Analyst-Predicted Earnings Growth	Variability in Capital Structure	Earnings Growth Rate Over Last 5 Years	5-Year Payout
Microsoft	-0.01%	-0.14	-0.31	25%	0%	0.69
Estimation Universe Average	0.03%	-2.76	1.44	15%	-1%	0.39
Estimation Universe Std Dev	0.04%	47.08	4.36	39%	18%	3.28
Standardized descriptor	-0.95	0.06	-0.40	0.24	0.03	0.09
Weight of each descriptor	-0.34	0.20	0.15	0.13	0.10	0.08
Exposure						-0.33

price moves with every percentage change in a factor. Thus, if the value factor rises by 10%, a stock or portfolio with an exposure of 0.5 to the value factor will see a return of 5%, all else equal.⁵

Once we have predetermined the *factor exposures* for all stocks based on their underlying characteristics, we estimate the factor returns using a regression-based method.⁶

A stock's return can then be described by the returns of its subcomponents: its Size exposure times the return of the Size factor plus its Value exposure times the pure return of the Value factor, and so on. This process can account for a substantial proportion of a stock's return. The remainder of the stock's return is deemed company specific and unique to each security. For example, the return to Microsoft over the last month can be viewed as:

$$r_{MSFT} = x_{Industry_1}r_{Industry_1} + x_{Industry_2}r_{Industry_2} + \dots + x_{Size}r_{Size} + x_{Value}r_{Value} + \dots r_{Firm-Specific}$$

where x is the exposure of Microsoft to the various factors and r_{Factor} denotes returns to the factors.

The returns to the factors are important. They are returns to the particular style or characteristic net of all other factors. For instance, the Value factor is the return to stocks with low price to book ratio with all the industry effects and other style effects removed. Industry returns have a similar interpretation and differ from a Global Industry Classification Standard (GICS®) industry-based return. They are estimated returns that reflect the returns to that industry net of all other style characteristics. They offer insight into the pure returns to the industry.

The final building block of our fundamental factor model is the modeling of company-specific returns. Predicting specific returns and risk is a difficult task that has been approached in a number of ways. The simplest approach is to assume that specific returns and/or risk will be the same as they have been historically. Another approach is to use a structural model

where the predicted specific risk of a company depends on its industry, size, and other fundamental characteristics. Both approaches—simple historical and modeled—are used in the Barra models, depending on the market. The modeled approach has the advantage of using fundamental data.

CRITICAL INSIGHTS FROM THE BARRA FUNDAMENTAL FACTOR MODEL

Fundamental analysis and fundamental *factor models* may begin with the same ideology but they offer different insights. Fundamental analysis ultimately focuses on in-depth company research, while factor models tie the information together at the portfolio level. The critical value of the factor model is that it shows the interaction of the firm's microeconomic characteristics. The value of the factor model at the company level is magnified at the portfolio level as the company-specific component becomes less important. Figure 3 illustrates this principle of diversification. As names are added to the portfolio, company-specific returns are diversified away. Because the common factor (systematic) portion stays roughly the same, it becomes an increasingly larger part of the portfolio risk and return.

This means that at the portfolio level common factors are more important than company-specific drivers in determining a portfolio's return and risk. Understanding and managing the common factor component becomes critical to the portfolio's performance.

The complementary character of fundamental factor models and individual security analysis allows managers to use factor models to analyze portfolio characteristics. Next, we discuss the benefits of using fundamental factor models, including:

- Monitoring and managing portfolio exposures over time

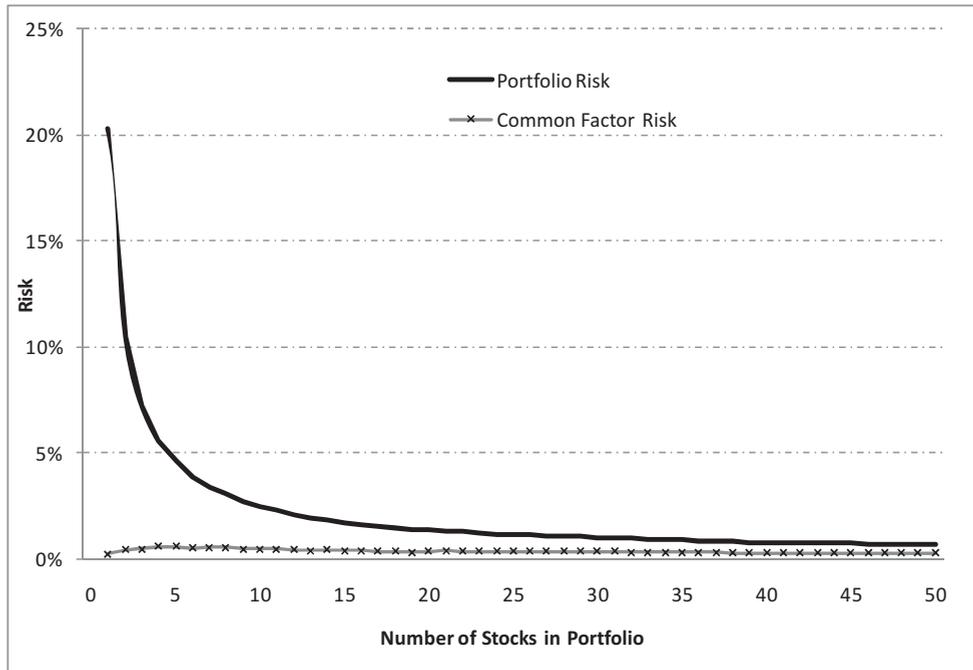


Figure 3 The Number of Stocks and the Impact on the Risk Composition

Note: This chart shows a stylized example of adding stocks to the portfolio where all the stocks have the same specific risk of 20%, there are two factors with risk of 10% and 5%, and correlation between them is 0.25. Factor exposures are drawn from a random distribution. Stocks are weighted equally in the portfolio.

- Understanding the contribution of factors and individual stocks to portfolio risk and tracking error relative to the relevant benchmark (risk decomposition)
- Attributing portfolio performance to factors and individual stocks to understand the return contribution of intended and accidental bets

Monitoring Portfolio Exposures

To illustrate, we use a portfolio of U.S. airline stocks. The concepts can be applied to any sector, multisector, or multicountry portfolio.

Since the middle of 2009, airline stocks have performed well. UAL (United), Delta, and Southwest saw big gains in December 2009 and February 2010. Table 5 lists the largest U.S. airline stocks as of April 30, 2010, with at least USD

1 billion market capitalization and their recent performance.

For the remainder of this section, we look at an equal-weighted portfolio of the stocks in Table 5. Figure 4 shows how the exposures of the airline portfolio to Barra factors evolved over time. The figure shows the top three exposures that changed the most in absolute terms between January 1995 and May 2010. The portfolio had an exposure to the Value factor of 1.8 in January 1995, and by May 2010 the exposure had declined to -0.9 . Essentially, the portfolio went from being relatively cheap to relatively expensive during this time. Airlines have also seen a long-term decrease in exposure to currency sensitivity, most likely due to changes in oil exposure management and global air traffic patterns.

There can also be important differences in the distribution of the stocks' exposures to a

Table 5 Largest Stocks in U.S. Airline Industry and Recent Performance

Company	Ticker	Market Cap (USD Bn)	1 year (3/31/09–3/31/10)	2009 Return	2008 Return
DELTA AIR LINES INC DE	DAL	10.4	111%	–1%	–23%
SOUTHWEST AIRLS CO	LUV	10.2	101%	33%	–29%
UAL CORP	UAUA	3.6	367%	17%	–67%
CONTINENTAL AIRLS [B]	CAL	3.1	109%	–1%	–19%
AMR CORP	AMR	2.8	63%	–28%	–24%
JETBLUE AIRWAYS CORP	JBLU	1.7	32%	–23%	20%
ALASKA AIR GROUP INC	ALK	1.5	161%	18%	17%
ALLEGiant TRAVEL CO	ALGT	1.1	3%	97%	68%
U S AIRWAYS GROUP INC	LCC	1.1	75%	–37%	–47%

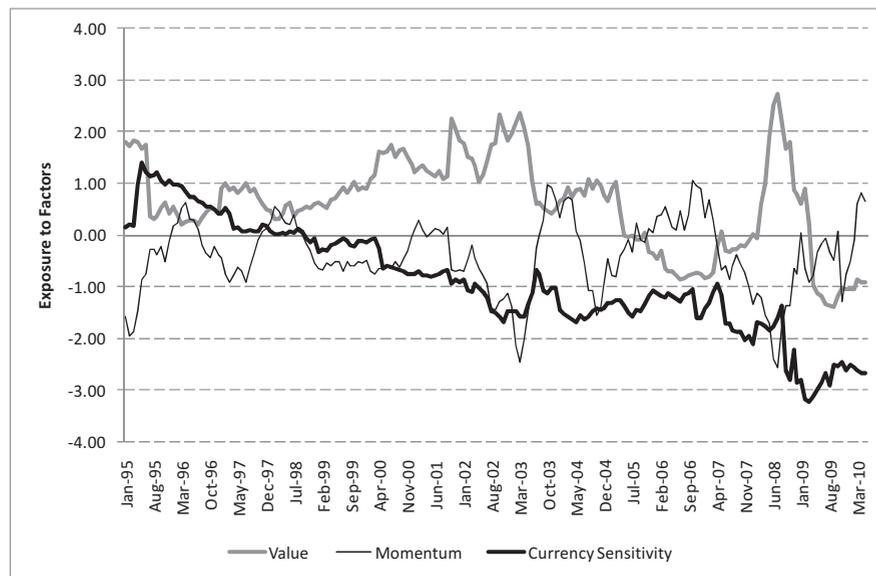
factor. Figure 5 shows the distribution of individual stock exposures to two of the U.S. factors—Value, which has the largest distribution, and Growth, which has among the most narrow distributions—as of May 2010. Two portfolios can have the same overall exposure to a factor but very different distributions.

Monitoring unintentional risk exposures that may not be visible on the surface can be critical. At the portfolio level, these exposures can be unintended bets that can impact overall performance. In addition, the distribution of exposures may be important. For example, a

portfolio of companies with a leverage exposure of zero has a very different economic profile than a portfolio with a barbell distribution where half the companies are overleveraged and potentially vulnerable to a collapse in credit conditions.

RISK DECOMPOSITION

Factor exposures highlight how sensitive a portfolio is to different sources of risk. However, to truly understand how risky these exposures are, we can use the factor model to attribute risk

**Figure 4** Airline Portfolio Exposures over Time

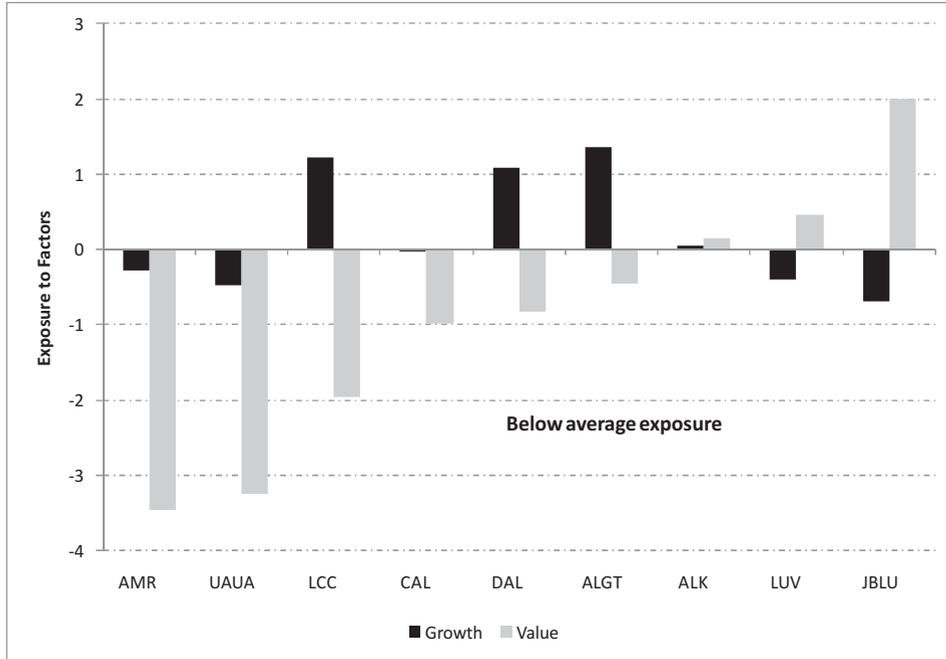


Figure 5 Monitoring the Distribution of Exposures (May 3, 2010)

fully. The combination of exposures and factor volatilities determines the riskiness of each position. For example, a portfolio can have a large exposure to a factor but if the factor isn't particularly risky, it won't be a major contributor to portfolio risk. Furthermore, the relationship between factors also matters. A large exposure to two factors that are highly correlated will also increase portfolio risk.

Continuing with the airline portfolio, we decompose risk as of April 30, 2010 across the four major sources (see Figure 6A). Since the stocks are within a single industry, industry risk contributes the most risk. Most importantly, we see that even with just 9 names in the portfolio, style risk far outweighs company-specific risk. The former contributes nearly three times that of the latter (16% versus 5.5%).

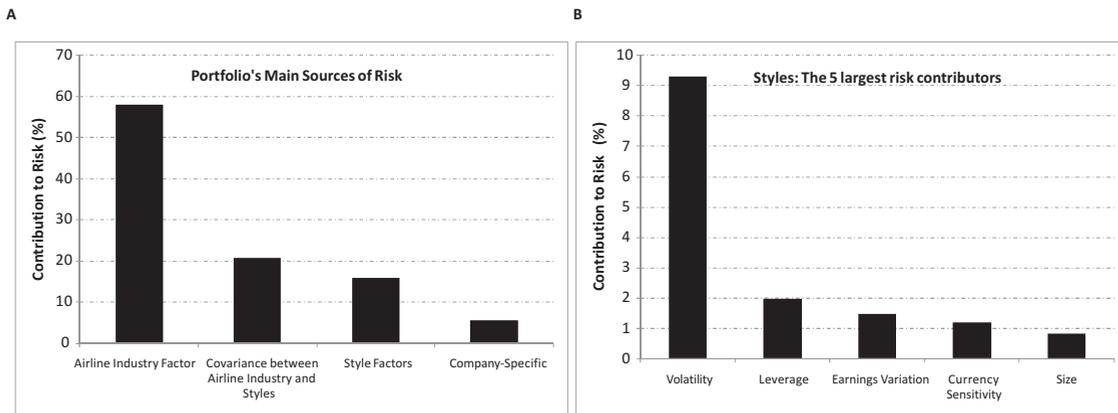


Figure 6 Sources of Risk in an Airline Portfolio, April 30, 2010, Using the Barra U.S. Equity Long-Term Model (USE3L)

Table 6 Exposure to Volatility of Stocks in an Airline Portfolio, April 30, 2010, Using the Barra U.S. Long-Term Equity Model (USE3L)

Portfolio	1.82		
USAir	3.28	JetBlue	1.52
UAL Corp	3.19	Alaska	1.01
AMR	2.70	Southwest	0.49
Continental	1.95	Allegiant	0.39
Delta	1.81		

Which specific style factors drive the style risk? As seen in Figure 6B, the Volatility factor is the biggest contributor by far to risk. This risk stems mostly from USAir and United's high exposure to the factor (see Table 6).

To summarize, *risk decomposition* provides two critical insights. First, as we move from the stock level to the portfolio level, style and industry risk become more important, overtaking company-specific risk. Second, we see that certain styles contribute more risk than others at the stock and portfolio levels. For example, the risk of United (UAL Corp) and USAir will be heavily impacted by the Volatility factor.

Performance Attribution

The fundamental factor model also provides insight into *performance attribution*. Managers can use the model to analyze past performance, attributing realized portfolio return to its various sources. This can include allocations to certain countries or sectors, or allocations to certain segments—small-cap names, emerging markets, or high beta names.

Table 7 shows the decomposition of realized returns for the airline portfolio for April 2010. The first column displays the portfolio return

attribution. The subsequent columns show the return attribution for each individual airline stock in isolation. The portfolio of airline stocks returned -4.3% for the month despite a positive contribution of 4.3% coming from style factors. JetBlue, for instance, was flat for the month, as its gain from style factors largely offset losses from the industry component. Similarly, Continental and UAL were helped by both strong contributions from style exposures. In contrast, positive gains from style factors were not enough to offset the company-specific losses suffered by USAir, Delta, AMR, and Allegiant. In fact, only about half the stocks realized positive company-specific returns.

Table 8 takes the last row in Table 7 and breaks it down into the individual styles in the model. The main source of positive return was the Size factor followed by the Currency Sensitivity, Leverage, and Volatility factors. In other words, airlines benefited from being smaller in cap size relative to the market (exposure of -1.7 to Size). They also benefited from the appreciation of the U.S. dollar (exposure of -2.7 to Currency Sensitivity). In addition, they were helped by being relatively levered (exposure of 2.6 to Leverage) and from having relatively higher overall and higher beta to the market (exposure of 1.7 to Volatility)

At the stock level, most of the airlines benefited from being relatively small. UAL and USAir benefited the most from the appreciation of the U.S. dollar. UAL, USAir, and AMR benefited the most from being relatively more levered than the other airlines. These three stocks also benefited the most from having relatively higher beta to the market and higher volatility.

Table 7 Return Attribution for Airline Portfolio and Stocks, %, March 31, 2010–April 30, 2010, Barra U.S. Equity Long-Term Model (USE3L)

	Portfolio	Alaska	Allegiant	AMR	Continental	Delta	JetBlue	South west	UAL	USAir
Total	-4.3	0.4	-11.1	-19.0	1.7	-17.2	0.2	-0.3	10.4	-3.8
Company-Specific	-4.4	2.6	-9.9	-22.6	1.5	-17.6	0.4	2.3	10.5	-6.6
Airline Industry	-4.2	-4.2	-4.2	-4.2	-4.2	-4.2	-4.2	-4.2	-4.2	-4.2
Styles	4.3	2.1	3.0	7.9	4.5	4.6	4.0	1.6	4.1	7.0

Table 8 Return Attribution for Styles Only in Percent, March 31, 2010–April 30, 2010, Barra U.S. Equity Long-Term Model (USE3L)

	Portfolio	Alaska	Allegiant	AMR	Continental	Delta	JetBlue	Southwest	UAL	USAir
Size	2.3	3.0	3.2	2.2	2.2	0.9	3.1	1.1	2.2	3.2
Currency Sensitivity	1.1	0.6	1.3	1.3	1.1	1.3	0.8	-0.1	2.0	2.0
Leverage	1.0	0.6	0.2	1.6	1.3	1.2	0.9	0.0	1.4	1.8
Volatility	0.9	0.7	0.1	1.3	1.0	0.9	0.5	0.3	1.6	1.5
Earnings Yield	0.8	-0.5	-0.1	4.0	0.7	1.9	0.1	0.4	-0.7	1.3
Trading Activity	0.1	0.2	0.0	0.2	0.2	0.2	0.1	0.1	0.2	0.2
Momentum	0.0	0.0	0.0	-0.1	-0.1	0.0	0.0	0.0	-0.1	-0.1
Growth	-0.1	0.0	-0.5	0.1	0.0	-0.4	0.2	0.1	-0.1	-0.1
Value	-0.2	0.0	-0.1	-0.8	-0.2	-0.2	0.4	0.1	-0.7	-0.4
Yield	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3
Size Nonlinearity	-0.3	-0.6	-0.7	-0.2	-0.2	0.1	-0.6	0.1	-0.2	-0.7
Earnings Variation	-1.0	-1.6	0.0	-1.5	-1.2	-0.8	-1.2	-0.1	-1.1	-1.4

Styles can contribute significantly to a manager's performance. In our example, the U.S. Volatility factor was the main driver. Looking at individual factors and stocks, we can also see that certain factors and stocks made a significant contribution to performance due to stock-specific performance or style contribution.

In summary, portfolio performance can be strongly impacted by unintended bets. The manager may be taking major risks without adequate compensation. The factor model helps uncover these issues.

KEY POINTS

- Fundamental analysis is the process of determining a security's future value by analyzing a combination of macro- and microeconomic events and company-specific characteristics.
- Though fundamental analysis focuses on the valuation of individual companies, most institutional investors recognize that there are common factors affecting all stocks. Common factors are shared characteristics between firms that affect their returns.
- Fundamental factor models are in fact complementary as opposed to antithetical to traditional security analysis. The basic building blocks of analysts and factor modelers are in fact similar: Both try to identify microeconomic traits that drive the risk and returns of individual securities.
- The objective of the fundamental analyst is to forecast return (or future stock values), whereas the fundamental factor model forecasts the fluctuation of a security or a portfolio of securities around its expected return. Some factors may help managers forecast return but not be good risk factors. A good return factor has persistent direction though not a lot of volatility—the ability of a company to beat earnings estimates is one of these factors. A good risk factor may be persistent or not but must be adequately volatile.
- Fundamental analysis and fundamental factor models may begin with the same ideology but they offer different insights. The most critical difference is that the factor model captures the interaction of the firm's microeconomic characteristics at the portfolio level. This is important because as names are added to the portfolio, company-specific returns are diversified away, and the common factor (systematic) portion becomes an increasingly larger part of the portfolio risk and return.
- There are three major benefits of using fundamental factor models: (1) monitoring and managing portfolio exposures over time; (2) understanding the contribution of factors and individual stocks to portfolio risk and tracking error relative to the relevant benchmark (risk decomposition); and (3) attributing portfolio performance to factors and individual

stocks to understand the return contribution of intended and accidental bets.

- Managers can use the model to analyze past performance, attributing realized portfolio return to its various sources. Portfolio performance can be strongly impacted by unintended bets. The manager may be taking major risks without adequate compensation. The factor model helps uncover these issues.
- The distribution of exposures may be important. For example, a portfolio of companies with a leverage exposure of zero has a very different economic profile than a portfolio with a barbell distribution where half the companies are overleveraged and potentially vulnerable to a collapse in credit conditions.

NOTES

1. In the Barra U.S. equity model, for example, we allow companies to be split up into five different industries, depending on their business structure.
2. All existing Barra models focus on a particular market, using an equity universe that includes all sectors and large to mid-caps with some small-caps.
3. The estimation universe average is a market-cap weighted average.
4. In effect, we build three separate valuation models. The results of each valuation model determine a set of weights, based on fundamental information. The final industry weights are a weighted average of the three weighting schemes. Further details are available in the *Barra U.S. Equity Model Handbook*.
5. Specifically, the effects of other factors as well as specific returns remain the same, and the risk-free rate is unchanged.
6. Details of the model construction are available in *The Barra Risk Model Handbook* or *Barra U.S. Equity Model Handbook*.

REFERENCES

- Carhart, M.M. (1997). On persistence in mutual fund performance. *Journal of Finance* 52, 1: 57–82.
- Fama, E. F., and French, K. R. (1992). The cross-section of expected stock returns. *Journal of Finance* 47, 2: 427–465.
- Rosenberg, B., and Marathe, V. (1976). Common factors in security returns: Microeconomic determinants and macroeconomic correlates. Working paper no. 44. University of California Institute of Business and Economic Research, Research Program in Finance.

Multifactor Equity Risk Models and Their Applications

FRANK J. FABOZZI, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

RAMAN VARDHARAJ, CFA
Vice President, OppenheimerFunds

FRANK J. JONES, PhD
Professor, Accounting and Finance Department, San Jose State University and
Chairman, Investment Committee, Private Ocean Wealth Management

Abstract: Multifactor equity risk models are classified as statistical models, macroeconomic models, and fundamental models. The most popular types of models used in practice are fundamental models. Many of the inputs used in a multifactor risk model are those used in traditional fundamental analysis. There are several commercially available fundamental multifactor risk models. There are asset management companies that develop proprietary models. Brokerage firms have developed models that they make available to institutional clients.

Quantitative-oriented common stock portfolio managers typically employ a *multifactor equity risk model* in constructing and rebalancing a portfolio and then for evaluating performance. The most popular type of multifactor equity risk model used is a fundamental factor model.¹ While some asset management firms develop their own model, most use commercially available models. In this entry we use one commer-

cially available model to illustrate the general features of *fundamental models* and how they are used to construct portfolios. In our illustration, we will use an old version of a model developed by Barra (now MSCI Barra). Although that model has been updated, the discussion and illustrations provide the essential points for appreciating the value of using multifactor equity models.

MODEL DESCRIPTION AND ESTIMATION

The basic relationship to be estimated in a multifactor risk model is

$$R_i - R_f = \beta_{i,F1} R_{F1} + \beta_{i,F2} R_{F2} + \dots + \beta_{i,FH} R_{FH} + e_i$$

where

R_i = rate of return on stock i

R_f = risk-free rate of return

$\beta_{i,Fj}$ = sensitivity of stock i to risk factor j

R_{Fj} = rate of return on risk factor j

e_i = nonfactor (specific) return on security i

The above function is referred to as a *return generating function*.

Fundamental factor models use company and industry attributes and market data as *descriptors*. Examples are price/earnings ratios, book/price ratios, estimated earnings growth, and trading activity. The estimation of a fundamental factor model begins with an analysis of historical stock returns and descriptors about a company. In the Barra model, for example, the process of identifying the *risk factors* begins with monthly returns for 1,900 companies that the descriptors must explain. Descriptors are not the “risk factors” but instead they are the candidates for risk factors. The descriptors are selected in terms of their ability to explain stock returns. That is, all of the descriptors are potential risk factors but only those that appear to be important in explaining stock returns are used in constructing risk factors.

Once the descriptors that are statistically significant in explaining stock returns are identified, they are grouped into *risk indexes* to capture related company attributes. For example, descriptors such as market leverage, book leverage, debt-to-equity ratio, and company’s debt rating are combined to obtain a risk index referred to as “leverage.” Thus, a risk index is a combination of descriptors that captures a particular attribute of a company.

The Barra fundamental multifactor risk model, the “E3 model” being the latest version, has 13 risk indexes and 55 industry groups. (The descriptors are the same variables that have been consistently found to be important in many well-known academic studies on risk factors.) Table 1 lists the 13 risk indexes in the Barra model.² Also shown in the table are the descriptors used to construct each risk index. The 55 industry classifications are grouped into 13 sectors. For example, the following three industries comprise the energy sector: energy reserves and production, oil refining, and oil services. The consumer noncyclicals sector consists of the following five industries: food and beverages, alcohol, tobacco, home products, and grocery stores. The 13 sectors in the Barra model are basic materials, energy, consumer noncyclicals, consumer cyclicals, consumer services, industrials, utility, transport, health care, technology, telecommunications, commercial services, and financial.

Given the risk factors, information about the exposure of every stock to each risk factor ($\beta_{i,Fj}$) is estimated using statistical analysis. For a given time period, the rate of return for each risk factor (R_{Fj}) also can be estimated using statistical analysis. The prediction for the expected return can be obtained from equation (1) for any stock. The nonfactor return (e_i) is found by subtracting the actual return for the period for a stock from the return as predicted by the risk factors.

Moving from individual stocks to portfolios, the predicted return for a portfolio can be computed. The exposure to a given risk factor of a portfolio is simply the weighted average of the exposure of each stock in the portfolio to that risk factor. For example, suppose a portfolio has 42 stocks. Suppose further that stocks 1 through 40 are equally weighted in the portfolio at 2.2%, stock 41 is 5% of the portfolio, and stock 42 is 7% of the portfolio. Then the exposure of the portfolio to risk factor j is

$$0.022 \beta_{1,Fj} + 0.022 \beta_{2,Fj} + \dots + 0.022 \beta_{40,Fj} + 0.050 \beta_{41,Fj} + 0.07 \beta_{42,Fj}$$

Table 1 Barra E3 Model Risk Definitions

Descriptors in Risk Index	Risk Index
Beta times sigma Daily standard deviation High-low price Log of stock price Cumulative range Volume beta Serial dependence Option-implied standard deviation	Volatility
Relative strength Historical alpha	Momentum
Log of market capitalization	Size
Cube of log of market capitalization	Size Nonlinearity
Share turnover rate (annual) Share turnover rate (quarterly) Share turnover rate (monthly) Share turnover rate (five years) Indicator for forward split Volume to variance	Trading Activity
Payout ratio over five years Variability in capital structure Growth rate in total assets Earnings growth rate over the last five years Analyst-predicted earnings growth Recent earnings change	Growth
Analyst-predicted earnings-to-price Trailing annual earnings-to-price Historical earnings-to-price	Earnings Yield
Book-to-price ratio	Value
Variability in earnings Variability in cash flows Extraordinary items in earnings Standard deviation of analyst-predicted earnings-to-price	Earnings Variability
Market leverage Book leverage Debt to total assets Senior debt rating	Leverage
Exposure to foreign currencies	Currency Sensitivity
Predicted dividend yield	Dividend Yield
Indicator for firms outside US-E3 estimation universe	Non-Estimation Universe Indicator

Adapted from Table 8-1 in Barra (1998, pp. 71–73). Adapted with permission.

The nonfactor error term is measured in the same way as in the case of an individual stock. However, in a well-diversified portfolio, the nonfactor error term will be considerably less for the portfolio than for the individual stocks in the portfolio.

The same analysis can be applied to a stock market index because an index is nothing more than a portfolio of stocks.

RISK DECOMPOSITION

The real usefulness of a linear multifactor model lies in the ease with which the risk of a portfolio with several assets can be estimated. Consider a portfolio with 100 assets. Risk is commonly defined as the variance of the portfolio’s returns. So, in this case, we need to find the variance–covariance matrix of the 100 assets. That would require us to estimate 100 variances (one for each of the 100 assets) and 4,950 covariances among the 100 assets. That is, in all we need to estimate 5,050 values, a very difficult undertaking. Suppose, instead, that we use a three-factor model to estimate risk. Then, we need to estimate (1) the three factor loadings for each of the 100 assets (i.e., 300 values), (2) the six values of the factor variance–covariance matrix, and (3) the 100 residual variances (one for each asset). That is, we need to estimate only 406 values in all. This represents a nearly 90% reduction from having to estimate 5,050 values, a huge improvement. Thus, with well-chosen factors, we can substantially reduce the work involved in estimating a portfolio’s risk.

Multifactor risk models allow a manager and a client to decompose risk in order to assess the exposure of a portfolio to the risk factors and to assess the potential performance of a portfolio relative to a benchmark. This is the portfolio construction and risk control application of the model. Also, the actual performance of a portfolio relative to a benchmark can be assessed. This is the performance attribution analysis application of the model.

Barra suggests that there are various ways that a portfolio’s total risk can be decomposed

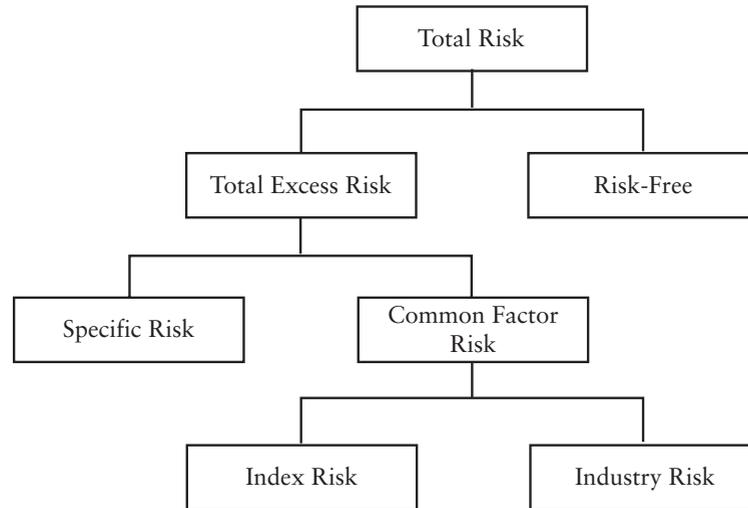


Figure 1 Total Risk Decomposition

Source: Figure 4.2 in Barra (1998, p. 34). Reprinted with permission.

when employing a multifactor risk model.³ Each decomposition approach can be useful to managers depending on the equity portfolio management that they pursue. The four approaches are (1) *total risk decomposition*, (2) *systematic-residual risk decomposition*, (3) *active risk decomposition*, and (4) *active systematic-active residual risk decomposition*.

In all of these approaches to risk decomposition, the total return is first divided into the risk-free return and the total excess return. The *total excess return* is the difference between the actual return realized by the portfolio and the risk-free return. The risk associated with the total excess return, called *total excess risk*, is what is further partitioned in the four approaches.

Total Risk Decomposition

There are managers who seek to minimize total risk. For example, a manager pursuing a long-short or market neutral strategy seeks to construct a portfolio that minimizes total risk. For such managers, total risk decomposition that breaks down the total excess risk into two components—*common factor risks* (e.g., capitalization and industry exposures) and *specific risk*—is useful. This decomposition is shown

in Figure 1. There is no provision for market risk, only risk attributed to the common factor risks and company-specific influences (i.e., risk unique to a particular company and therefore uncorrelated with the specific risk of other companies). Thus, the market portfolio is not a risk factor considered in this decomposition.

Systematic-Residual Risk Decomposition

There are managers who seek to time the market or who intentionally make bets to create a different exposure from that of a market portfolio. Such managers would find it useful to decompose total excess risk into systematic risk and residual risk as shown in Figure 2. Unlike in the total risk decomposition approach just described, this view brings market risk into the analysis. It is the type of decomposition where systematic risk is the risk related to a portfolio's beta.

Residual risk in the systematic-residual risk decomposition is defined in a different way from residual risk in the total risk decomposition. In the systematic-residual risk decomposition, residual risk is risk that is uncorrelated with the market portfolio. In turn, residual risk

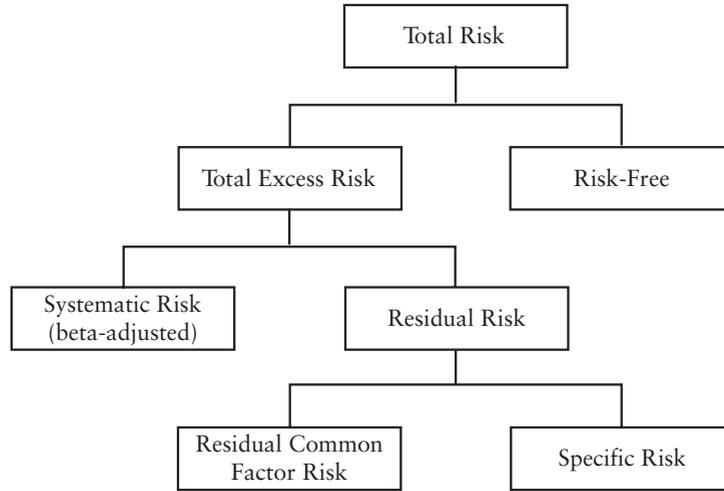


Figure 2 Systematic-Residual Risk Decomposition
 Source: Figure 4.3 in Barra (1998, p. 34). Reprinted with permission.

is partitioned into specific risk and common factor risk. Notice that the partitioning of risk described here is different from that in the arbitrage pricing theory model where all risk factors that could not be diversified away were referred to as “systematic risks.” In the discussion here, risk factors that cannot be diversified away are classified as market risk and common factor risk. Systematic risk can be diversified to a negligible level.

Active Risk Decomposition

The active risk decomposition approach is useful for assessing a portfolio’s risk exposure and actual performance relative to a benchmark index. In this type of decomposition, shown in Figure 3, the total excess return is divided into benchmark risk and active risk. Benchmark risk is defined as the risk associated with the benchmark portfolio.

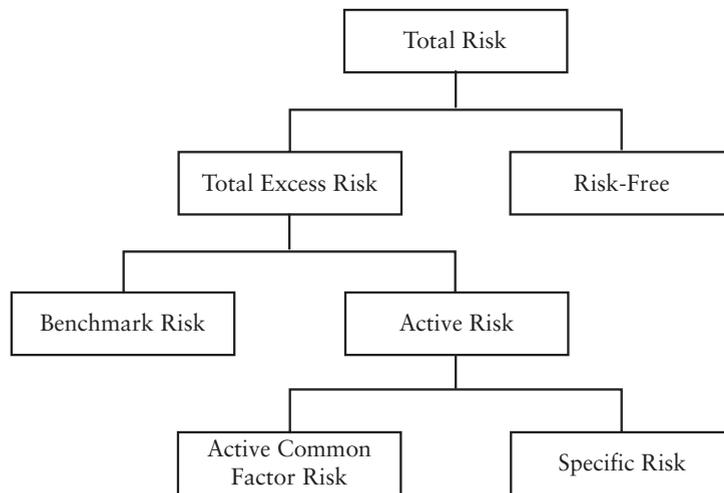


Figure 3 Active Risk Decomposition
 Source: Figure 4.4 in Barra (1998, p. 34). Reprinted with permission.

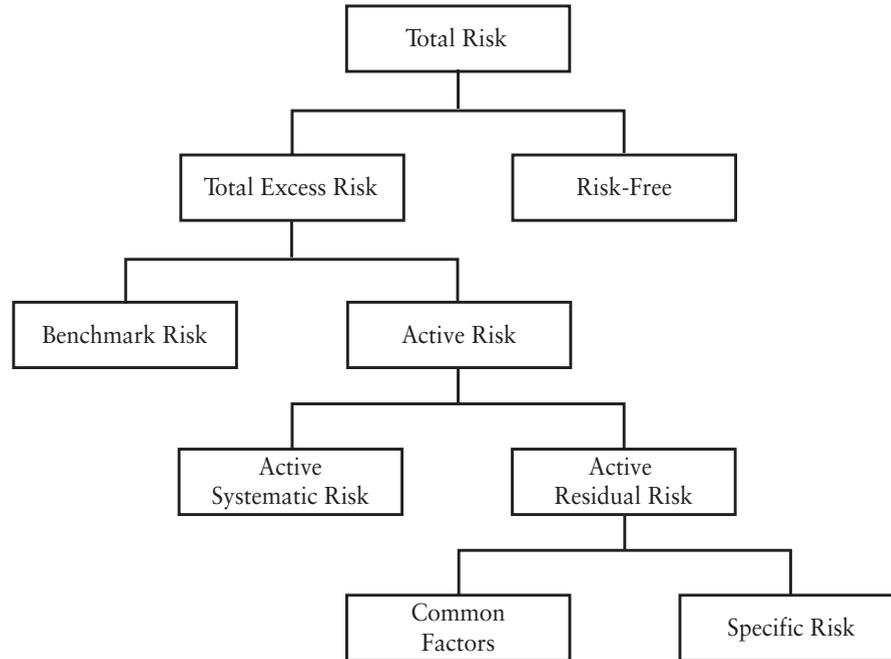


Figure 4 Active Systematic-Active Residual Risk Decomposition
Source: Figure 4.5 in Barra (1998, p. 37). Reprinted with permission.

Active risk is the risk that results from the manager's attempt to generate a return that will outperform the benchmark. Another name for active risk is tracking error. The active risk is further partitioned into common factor risk and specific risk. This decomposition is useful for managers of index funds and traditionally managed active funds.

Active Systematic-Active Residual Risk Decomposition

There are managers who overlay a market-timing strategy on their stock selection. That is, they not only try to select stocks they believe will outperform but also try to time the purchase of the acquisition. For a manager who pursues such a strategy, it will be important in evaluating performance to separate market risk from common factor risks. In the active risk decomposition approach just discussed, there is no market risk identified as one of the risk factors.

Since market risk (i.e., systematic risk) is an element of active risk, its inclusion as a source of risk is preferred by managers. When market risk is included, we have the active systematic-active residual risk decomposition approach shown in Figure 4. Total excess risk is again divided into benchmark risk and active risk. However, active risk is further divided into active systematic risk (i.e., active market risk) and active residual risk. Then active residual risk is divided into common factor risks and specific risk.

Summary of Risk Decomposition

The four approaches to risk decomposition are just different ways of slicing up risk to help a manager in constructing and controlling the risk of a portfolio and for a client to understand how the manager performed. Figure 5 provides an overview of the four approaches to carving up risk into specific/common factor, systematic/residual, and benchmark/active risks.

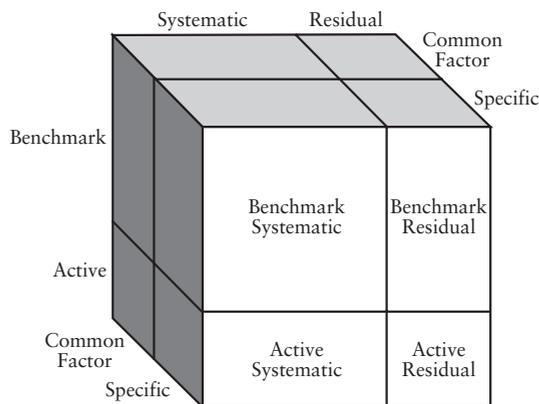


Figure 5 Risk Decomposition Overview
 Source: Figure 4.6 in Barra (1998, p. 38). Reprinted with permission.

APPLICATIONS IN PORTFOLIO CONSTRUCTION AND RISK CONTROL

The power of a multifactor risk model is that given the risk factors and the risk factor sensitivities, a portfolio’s risk exposure profile can be quantified and controlled. The three examples below show how this can be done so that a manager can avoid making unintended bets. In the examples, we use the Barra E3 factor model.⁴

A fundamental multifactor risk model can be used to assess whether the current portfolio is consistent with a manager’s strengths. Table 2 is a list of the top 15 holdings of Portfolio ABC as of December 31, 2008. Table 3 is a summary risk decomposition report for the same portfolio. The portfolio had a total market value of \$5.4 billion, 868 holdings, and a predicted beta of 1.15. The risk report also shows that the portfolio had an active risk of 6.7%. This is its tracking error with respect to the benchmark, the S&P 500 index. Notice that nearly 93% of the active risk variance (which is 44.8) came from common factor variance (which is 41.6), and only a small proportion came from stock-specific risk variance (also known as asset selection variance, which is 3.2). Clearly, the manager of this portfolio had placed fairly large factor bets.

The top portion of Table 4 lists the factor risk exposures of Portfolio ABC relative to those of the S&P 500 index, its benchmark. The first column shows the exposures of the portfolio, and the second column shows the exposures of the benchmark. The last column shows the active exposure, which is the difference between the portfolio exposure and the benchmark exposure. The exposures to the risk index factors are measured in units of standard deviation,

Table 2 Portfolio ABC’s Holdings (only the top 15 holdings shown)

Ticker	Security Name	Shares	Price (\$)	Weight	Beta	Industry
XOM	Exxon Mobil Corp.	3,080,429	79.83	4.56	0.92	Oil Refining
MSFT	Microsoft Corp.	6,235,154	19.44	2.25	0.95	Computer Software
CVX	Chevron Corp.	1,614,879	73.97	2.21	0.98	Energy Reserves & Production
IBM	International Business Machines Corp.	1,100,900	84.16	1.72	0.83	Computer Software
T	AT&T Inc.	3,226,744	28.50	1.70	0.80	Telephone
HPQ	Hewlett-Packard Co.	2,464,100	36.29	1.66	0.84	Computer Hardware & Business Machines
INTC	Intel Corp.	5,997,300	14.66	1.63	0.87	Semiconductors
COP	ConocoPhillips	1,634,986	51.80	1.57	1.24	Energy Reserves & Production
CSCO	Cisco Systems Inc.	5,186,400	16.30	1.57	0.95	Computer Hardware & Business Machines
JNJ	Johnson & Johnson	1,403,544	59.83	1.56	0.54	Medical Products & Supplies
OXY	Occidental Petroleum Corp.	1,324,426	59.99	1.47	1.26	Energy Reserves & Production
PG	Procter & Gamble Co.	1,249,446	61.82	1.43	0.57	Home Products
GE	General Electric Co.	4,762,984	16.20	1.43	1.41	Heavy Electrical Equipment
PFE	Pfizer Inc.	4,339,092	17.71	1.42	0.61	Drugs
TWX	Time Warner Inc.	1,948,880	30.18	1.09	1.32	Media

Table 3 Portfolio ABC's Summary Risk Decomposition Report

Number of Securities	868
Number of Shares	298,371,041
Average Share Price	\$24.91
Weighted Average Share Price	\$35.30
Portfolio Ending Market Value	\$5,396,530,668
Predicted Beta (vs. Benchmark, S&P 500)	1.15
Barra Risk Decomposition (Variance)	
Asset Selection Variance	3.2
Common Factor Variance:	
Risk Indexes	22.5
Industries	11.7
Covariance $\times 2$	7.5
Common Factor Variance	41.6
Active Variance	44.8
Benchmark Variance	749.4
Total Variance	1,016.6
Barra Risk Decomposition (Std. Dev.)	
Asset Selection Risk	1.8
Common Factor Risk:	
Risk Indexes	4.7
Industries	3.4
Covariance $\times 2$	
Common Factor Risk	6.5
Active Risk	6.7
Benchmark Risk	27.4
Total Risk	31.9

while the exposures to the industry factors are measured in percentages. The portfolio had a high active exposure to the Volatility risk index factor. That is, the stocks in the portfolio were far more volatile than the stocks in the benchmark. On the other side, the portfolio had a low active exposure to the Size risk index. That is, the stocks in the portfolio were smaller than the benchmark average in terms of market capitalization. The lower portion of Table 4 is an abbreviated list of the industry factor exposures.

An important use of such risk reports is the identification of portfolio bets, both explicit and implicit. If, for example, the manager of Portfolio ABC did not intend to place the large bet on the Volatility risk index, then he has to make appropriate changes in the portfolio holdings until the active exposure to this factor is closer to zero.

Risk Control against a Stock Market Index

The objective of equity indexing is to match the performance of some specified stock market

Table 4 Analysis of Portfolio ABC's Factor Exposures

Risk Indexes (std. dev. units)	Managed ^a	Benchmark ^b	Active ^c
U.S. Volatility	0.321	-0.089	0.410
U.S. Value	0.199	-0.024	0.223
U.S. Earnings Variation	0.149	-0.053	0.202
U.S. Earnings Yield	0.243	0.053	0.191
U.S. Trading Activity	0.161	0.052	0.109
U.S. Leverage	-0.036	-0.110	0.074
U.S. Growth	0.004	-0.069	0.073
U.S. Non-Estimation Universe	0.027	0.000	0.027
U.S. Currency Sensitivity	-0.013	0.007	-0.019
U.S. Momentum	-0.183	-0.043	-0.139
U.S. Yield	-0.115	0.078	-0.194
U.S. Size Non-Linearity	-0.107	0.123	-0.230
U.S. Size	-0.244	0.356	-0.600
Top Three Industries (percentages)	Managed	Benchmark	Active
U.S. Energy Reserves	0.098	0.064	0.033
U.S. Semiconductors	0.052	0.023	0.028
U.S. Mining and Metals	0.036	0.009	0.027

^a Managed return.

^b Benchmark return (S&P 500).

^c Active return = Managed return - Benchmark return.

Table 5 Factor Exposures of a 50-Stock Portfolio that Optimally Matches the S&P 500 Index

Risk Indexes (std. dev. units)	Managed ^a	Benchmark ^b	Active ^c
U.S. Volatility	-0.153	-0.089	-0.063
U.S. Momentum	-0.062	-0.043	-0.018
U.S. Size	0.795	0.356	0.440
U.S. Size Non-Linearity	0.164	0.123	0.041
U.S. Trading Activity	-0.001	0.052	-0.053
U.S. Growth	-0.052	-0.069	0.016
U.S. Earnings Yield	0.076	0.053	0.023
U.S. Value	-0.019	-0.024	0.005
U.S. Earnings Variation	-0.122	-0.053	-0.069
U.S. Leverage	-0.176	-0.110	-0.066
U.S. Currency Sensitivity	-0.048	0.007	-0.055
U.S. Yield	0.140	0.078	0.061
U.S. Non-Estimation Universe	0.000	0.000	0.000

^a Managed return.

^b Benchmark return (S&P 500).

^c Active return = Managed return – Benchmark return.

index with little tracking error. To do this, the risk profile of the indexed portfolio must match the risk profile of the designated stock market index. Put in other terms, the factor risk exposure of the indexed portfolio must match as closely as possible the exposure of the designated stock market index to the same factors. Any differences in the factor risk exposures result in tracking error. Identification of any differences allows the indexer to rebalance the portfolio to reduce tracking error.

To illustrate this, suppose that an index manager has constructed a portfolio of 50 stocks to match the S&P 500 index. Table 5 lists the exposures to the Barra risk indexes of the 50-stock portfolio and the S&P 500 index. The last column in the exhibit shows the difference in exposures. The differences are very small except for the exposures to the Size risk index factor. Though not shown in this exhibit, there is a similar list of exposures to the 55 industry factors.

The illustration in Table 5 uses price data as of December 31, 2008. It demonstrates how a multifactor risk model can be combined with an optimization model to construct an indexed portfolio when a given number of holdings are sought. Specifically, the portfolio analyzed in the exhibit is the result of an application

in which the manager wants a portfolio constructed that matches the S&P 500 index with only 50 stocks and that minimizes tracking error. The optimization model uses the multifactor risk model to construct a 50-stock portfolio with a tracking error versus S&P 500 index of just 2.75%. Since this is the optimal 50-stock portfolio to replicate the S&P 500 index with a minimum tracking error risk, this tells the index manager that if he seeks a lower tracking error, then more stocks must be held. Note, however, that the optimal portfolio changes as time passes and prices move.

Tilting a Portfolio

Now let's look at how an active manager can construct a portfolio to make intentional bets. Suppose that a portfolio manager seeks to construct a portfolio that generates superior returns relative to the S&P 500 by tilting it toward low P/E stocks. At the same time, the manager does not want to increase tracking error significantly. An obvious approach may seem to be to identify all the stocks in the universe that have a lower than average P/E. The problem with this approach is that it introduces unintentional bets with respect to the other risk indexes.

Table 6 Factor Exposures of a Portfolio Tilted Toward Earnings Yield

Risk Indexes (std. dev. units)	Managed ^a	Benchmark ^b	Active ^c
U.S. Volatility	-0.050	-0.089	0.039
U.S. Momentum	-0.096	-0.043	-0.052
U.S. Size	0.284	0.356	-0.072
U.S. Size Non-Linearity	0.096	0.123	-0.027
U.S. Trading Activity	0.114	0.052	0.062
U.S. Growth	-0.096	-0.069	-0.027
U.S. Earnings Yield	0.553	0.053	0.500
U.S. Value	0.076	-0.024	0.100
U.S. Earnings Variation	-0.091	-0.053	-0.038
U.S. Leverage	-0.153	-0.110	-0.043
U.S. Currency Sensitivity	0.066	0.007	0.059
U.S. Yield	0.179	0.078	0.100
U.S. Non-Estimation Universe	0.000	0.000	0.000

^a Managed return.

^b Benchmark return (S&P 500).

^c Active return = Managed return - Benchmark return.

Instead, an optimization method combined with a multifactor risk model can be used to construct the desired portfolio. The necessary inputs to this process are the tilt exposure sought and the benchmark stock market index. Additional constraints can be placed, for example, on the number of stocks to be included in the portfolio. The Barra optimization model can also handle additional specifications such as forecasts of expected returns or alphas on the individual stocks.

In our illustration, the tilt exposure sought is toward low P/E stocks, that is, toward high earnings yield stocks (since earnings yield is the inverse of P/E). The benchmark is the S&P 500. We seek a portfolio that has an average earnings yield that is at least 0.5 standard deviations more than that of the earnings yield of the benchmark. We do not place any limit on the number of stocks to be included in the portfolio. We also do not want the active exposure to any other risk index factor (other than earnings yield) to be more than 0.1 standard deviations in magnitude. This way we avoid placing unintended bets. While we do not report the holdings of the optimal portfolio here, Table 6 provides an analysis of that portfolio by comparing the risk exposure of the 50-stock optimal portfolio to that of the S&P 500. Though

not shown in Table 6, there is a similar list of exposures to the 55 industry factors.

KEY POINTS

- There are three types of multifactor equity risk models that are used in practice: statistical, macroeconomic, and fundamental. The most popular is the fundamental model.
- A multifactor equity risk model assumes that stock returns (and hence portfolio returns) can be explained by a linear model with multiple factors, consisting of “risk index” factors such as company size, volatility, momentum, and so on, and “industry” factors. The portion of the stock return that is not explained by this model is the stock-specific return.
- The risk index factors are measured in standard deviation units, while the industry factors are measured in percentages.
- The real usefulness of a linear multifactor model lies in the ease with which the risk (i.e., the volatility) of a portfolio with several assets can be estimated. Instead of estimating the variance-covariance matrix of its assets, it is only necessary to estimate the portfolio’s factor exposures and the variance-covariance matrix of the factors, a computationally much easier task.

- The variance-covariance matrix of the factors and the factor exposures of stocks are calculated based on a mix of historical and current data and are updated periodically.
 - Total risk of a portfolio can be decomposed in several ways. The partitioning method chosen is based on what is useful given the manager's strategy. The active risk decomposition method is useful for managers of index funds and traditionally managed active funds.
 - The level of active risk of a portfolio and the split of the tracking error variance between the common factor portion and the stock-specific portion are useful in assessing if the portfolio is constructed in a way that is consistent with the manager's strengths.
 - The list of active factor exposures of a portfolio helps the manager identify its bets, both explicit and implicit. If a manager discovers some unintended bets, then the portfolio can be rebalanced so as to minimize such bets.
 - Using a multifactor risk model and an optimization model, a portfolio that has the minimum active risk relative to its benchmark for a given number of assets held can be constructed. This application is useful for passive managers.
- Similarly, a manager can construct a portfolio that tilts toward a specified factor and has no material active exposure to any other factor. This application is useful for active managers.

NOTES

1. For a discussion of the different types of factor models, see Connor (1995).
2. For a more detailed description of each descriptor, see Appendix A in Barra (1998). A listing of the 55 industry groups is provided in Table 4 in this entry.
3. See Chapter 4 in Barra (1998). The discussion to follow in this section follows that in the Barra publication.
4. The illustrations were created by the authors based on applications suggested in Chapter 6 of Barra (1996).

REFERENCES

- Barra (1996). *United States Equity Model Handbook*. Berkeley, CA: Barra.
- Barra (1998). *Risk Model Handbook, United States Equity: Version 3*. Berkeley, CA: Barra.
- Connor, G. (1995). The three types of factor models. *Financial Analysts Journal* 15: 42–57.

Multifactor Fixed Income Risk Models and Their Applications

ANTHONY LAZANAS, PhD
Managing Director, Barclays

ANTÓNIO BALDAQUE DA SILVA, PhD
Managing Director, Barclays

RADU GĂBUDEAN, PhD
Vice President, Barclays

ARNE D. STAAL, PhD
Director, Barclays

Abstract: Multifactor risk models seek to estimate and characterize the risk of a portfolio, either in absolute value or when compared against a benchmark. Risk is typically decomposed into a systematic and an idiosyncratic component. Systematic risk captures the exposures the portfolio has to broad risk factors, such as interest rates or spreads. This risk is driven by the exposure of the portfolio to these risk factors, their volatility, and the correlation between these different sources of risk. Idiosyncratic risk captures the uncertainty associated with the particular issuers in the portfolio. Idiosyncratic risk is diversifiable by spreading the exposure to a large number of individual issuers. Multifactor risk models allow for the decomposition of the total risk by risk factor (or sets of risk factors). If the factors are economically meaningful, the risk model can provide relevant intuition regarding the major variables influencing the volatility of the portfolio and be a useful tool in portfolio construction.

In this entry, we discuss risk modeling construction and applications to fixed income portfolios. Although they share a similar framework, multifactor models in fixed income use different building blocks and provide a different analysis of the risk of a portfolio.

When analyzing their holdings, portfolio managers constantly monitor their exposures, typically net of a benchmark: What is the portfolio net duration? How risky is the overweight to credit? How does it relate to the exposure to mortgages? What is the exposure to specific

The authors would like to thank Andy Sparks for his valuable comments.

issuers? Even when portfolio holdings and exposures are well known, portfolio managers increasingly rely on quantitative techniques to translate this information into a common risk language. *Risk models* can present a coherent view of the portfolio, its exposures, and how they correlate to each other. They can quantify the risk of each exposure and its contribution to the overall risk of the portfolio.

Fixed income securities are exposed to many different types of risk. Multifactor risk models in this area capture these risks by first identifying common sources along different dimensions, the systematic risk factors. All risk not captured by systematic factors is considered idiosyncratic or security-specific. Typically, fixed income systematic risk factors are divided into two sets: those that influence securities across asset classes (e.g., yield curve risk) and those specific to a particular asset class (e.g., prepayment risk in securitized products).

There are many ways to define systematic *risk factors*. For instance, they can be defined purely by statistical methods, observed in the markets, or estimated from asset returns. In fixed income, the standard approach is to use pricing models to calculate the analytics that are the natural candidates for risk factor loadings (also called sensitivities). In this setting, the risk factors are estimated from cross-sectional asset returns. This is the approach taken in the Barclays Global Risk Model,¹ which is the model used for illustration throughout this entry.

In this risk model, the forecasted risk of the portfolio is driven by both a systematic and an idiosyncratic (also called specific, nonsystematic, and concentration) component. The forecasted systematic risk is a function of the mismatch between the portfolio and the benchmark in the exposures to the risk factors, such as yield curve or spreads. The exposures are aggregated from security-level analytics. The systematic risk is also a function of the volatility of the risk factors, as well as the correlations between them. In this setting, the correlation

of returns across securities is driven by the correlation of systematic risk factors these securities load on. As the model uses security-level returns and analytics to estimate the factors, we can recover the idiosyncratic return for each security. This is the return net of all systematic factors. The model uses these idiosyncratic returns to estimate rich specifications for the idiosyncratic risk.

APPROACHES USED TO ANALYZE RISK

In what follows, we turn to the analysis of the risk of a particular portfolio, going through the different approaches typically used. Specifically, consider a portfolio manager that is benchmarked against the Barclays US Aggregate Index. Moreover, suppose she believes interest rates are coming down—so she wants to be long duration—and that she wants some extra yield in her portfolio—meaning investing in bonds with relatively higher spreads. Finally, let us assume that she is mandated to keep the difference between the returns of the portfolio and the benchmark at around 15 basis points, on a monthly basis. Therefore, she has to track a benchmark, but is allowed to deviate from it up to a point in order to express views that hopefully lead to superior returns. A portfolio manager with such a mandate is called an enhanced indexer. The amount of deviation allowed is called the risk budget (15 basis points in our example) and can be quantified using a risk model. The risk model produces an estimate of the volatility of the difference of the portfolio and benchmark returns, called *tracking error volatility* (TEV). (In this entry, we refer to TEV, risk, and the standard deviation of the portfolio net returns interchangeably.) The portfolio manager should keep the TEV at a level equal to or less than her risk budget. For illustration, we construct a portfolio with 50 securities that is consistent with the portfolio manager's views

Table 1 Market Weights for Portfolio and Benchmark

Asset Class	Portfolio	Benchmark	Difference
Total	100.0	100.0	0.0
Treasury	30.2	32.1	-1.9
Government-related	5.8	12.3	-6.5
Corporate industrials	9.0	9.7	-0.7
Corporate utilities	2.9	2.1	0.8
Corporate financials	18.6	6.4	12.2
MBS agency	28.4	34.1	-5.8
ABS	0.0	0.3	-0.3
CMBS	5.2	3.1	2.1

and risk budget and analyze it throughout this entry.

Market Structure and Exposure Contributions

The first level of analysis that any portfolio manager usually performs is to compare the portfolio holdings in terms of market value with the holdings from the benchmark. For instance, Table 1 shows that the composition of the portfolio has several important mismatches when compared with the benchmark. The portfolio is underweighted in Treasuries and government-related securities by 8.4%. This is compensated with an overweight of 12.3% in corporates, especially in the financials sector. Other mismatches include an underweight in mortgage-backed securities (MBS) (-5.8%) and an overweight in commercial mortgage-backed securities (CMBS) (+2.1%).

Interestingly, for an equity manager, this kind of information—for example, applied to the different industries or sectors of the portfolio—would be of paramount importance to the analysis of the risk of her portfolio. For a fixed income portfolio, this is not the case. Although important, this analysis tells us very little about the true active exposures of a fixed income portfolio. What if the Treasuries in the portfolio have significantly longer duration than those in the benchmark—would we

Table 2 Aggregate Analytics

Analytics	Portfolio	Benchmark	Difference
Duration	4.55	4.30	0.25
Spread duration	4.67	4.56	0.11
Convexity	-0.15	-0.29	0.13
Vega	-0.02	-0.01	-0.01
Spread	157	57	100

be really “short” in this asset class? What if the spreads from financials in the portfolio are much smaller than those in the benchmark—the weight mismatch that important?

To answer this kind of questions, we turn to another typical dimension of analysis—the exposure of the portfolio to major sources of risk. An example of such a risk exposure is the duration of the portfolio. Other exposures typically monitored are the spread duration, convexity, spread level, and vega (if the portfolio has many securities with optionality, such as mortgages or callable bonds).

Table 2 shows these analytics at the aggregate level for our portfolio, benchmark, and the difference between the two. In particular, we can see that the portfolio is long duration (+0.25 years), consistent with the forecast the manager has regarding yield curve moves. In terms of spread duration, the mismatch is somewhat smaller. We can also see that the portfolio has significantly lower negative convexity than the benchmark (-0.15 versus -0.29), probably coming from the smaller weight MBS securities have in the portfolio. The portfolio has also a higher negative vega, but the number is reasonably small for both universes. Finally, the portfolio has significantly higher spreads (100 basis points) than the benchmark. This mismatch is consistent with the manager’s goal of having a higher yield in her portfolio, when compared with the benchmark.

The analysis in the Tables 1 and 2 can be combined to deliver a more detailed picture of where the different exposures are coming from. Table 3 shows that analysis for the duration of the portfolio. This exhibit shows that the majority of the mismatch in duration contribution

Table 3 Duration Contribution per Asset Class

Duration Contribution	Portfolio	Benchmark	Difference
Total	4.55	4.30	0.25
Treasury	1.92	1.71	0.21
Government-related	0.40	0.49	-0.09
Corporate	1.31	1.19	0.11
Securitized	0.92	0.90	0.02

(market-weighted duration exposures) comes from the Treasury component of our portfolio (+0.21). Interestingly, even though we are short in Treasuries, we are actually long in duration for that asset class. This means that our Treasury portfolio will be negatively impacted, when compared with the benchmark, by an increase in interest rates. Because we are short in Treasuries, this result must mean that our Treasury portfolio is longer in duration than the Treasury component of the benchmark. Conversely, we have a relatively small contribution to excess duration coming from our very large overexposure to corporates. This means that on average the corporate bonds in the portfolio are significantly shorter in duration than those in the benchmark.

Adding Volatility and Correlations into the Analysis

The analysis above gives us some basic understanding of our exposures to different kinds of risk. However, it is still hard to understand how we can compare the level of risk across these different exposures. What is more risky, the long duration exposure of 0.25 years, or the extra spread of 100 basis points? How can we quantify how serious is the vega mismatch on my portfolio? Specifically, the risk of the portfolio is a function of the exposures to the risk factors, but also of how volatile (how “risky”) each of the factors is. So to enhance the analysis we bring volatilities into the picture. Table 4 shows the outcome of this addition to our example. In

Table 4 Isolated Risk per Category

Risk Factors Categories	Risk
Curve	8.5
Volatility	1.7
Spread government-related	3.0
Spread corporate	5.1
Spread securitized	3.0

particular, it displays the risk of the different exposures of the portfolio in isolation (that is, if the only active imbalances were those from that particular set of risk factors).

For example, in Table 4 one can see that if the only active weight in the portfolio were the mismatch in the yield curve exposures, the risk of the portfolio would be 8.5 basis points per month. By adding volatilities into the analysis, we can now quantify that the mismatch of +0.25 years in duration “costs” the portfolio 8.5 basis points per month of extra volatility, when taken in isolation.² Similarly, if the only mismatch were the exposure to corporate spreads, the risk of the portfolio would be 5.1 basis points. Interestingly, we also see that both government-related and securitized sectors have nontrivial risk, despite having smaller imbalances in terms of market weights. By bringing volatilities into the analysis, we can now compare and quantify the impact of each of the imbalances in the portfolio.

For future reference, consider the volatility of the portfolio if all these sources of risk were independent (e.g., correlations were zero). That number would be 10.9 basis points per month.³ Of course, this scenario is unrealistic, as these sources of risk are not independent. Also, this analysis does not allow us to understand the interplay between the different imbalances. For instance, we know that the isolated risk associated with the curve is 8.5. But this value can be achieved both by being long or short duration. So the isolated number does not allow us to understand the impact of the curve imbalance to the total risk of the portfolio. The net impact certainly depends on the sign of the imbalance. For instance, if the long exposure in

Table 5 Correlated Risk per Category

Risk Factors Categories	Risk
Total	9.3
Curve	5.9
Volatility	0.1
Spread government-related	0.1
Spread corporate	2.4
Spread securitized	0.7

curve is diversified away by a long exposure in credit (due, for instance, to negative correlation between rates and credit spreads), a symmetric (short) curve exposure would add to the risk of the long exposure in credit. The risk is clearly smaller in the first case.

To alleviate these shortcomings, we bring correlations into the picture. They allow us to understand the net impact of the different exposures to the portfolio’s total risk and to detect potential sources of diversification among the imbalances in the portfolio. Table 5 reports the contribution of each of the risk factor groups to the total risk, once all correlations are taken into account. The total risk (9.3 bps/month) is smaller than the zero-correlation risk calculated before (10.9 bps/month) due to generally negative correlations between the curve and the spread factors. The exhibit also allows us to isolate the main sources of risk as being curve (5.9 bps/month) and credit spreads (2.4 bps/month), in line with the evidence from the earlier analysis. In particular, the risk of the government-related and securitized spreads is significantly smaller once correlations are taken into account.

The difference in analysis between the isolated and correlated risks reported in Tables 4 and 5 deserves a bit more discussion. For simplicity, assume there are only two sources of risk in the portfolio—yield curve (Y) and spreads (S). The total systematic variance of the portfolio (P) can be illustrated as follows:

$$\begin{aligned} VAR(P) &= VAR(Y + S) \\ &= VAR(Y) + VAR(S) + 2COV(Y, S) \\ &= Y \times Y + S \times S + 2(Y \times S) \end{aligned}$$

where we use the product (\times) to represent variances and covariances. Another way to represent this summation is using the following matrix:

$$\begin{bmatrix} Y \times Y & Y \times S \\ Y \times S & S \times S \end{bmatrix}$$

The sum of the four elements in the matrix is the variance of the portfolio. The isolated risk (in standard deviation units) reported in Table 4 is the square root of the diagonal terms. So the isolated risk due to spreads is represented as

$$Risk_{Spreads}^{Isolated} = \sqrt{S \times S}$$

It would be a function of the exposure to all spread factors, the volatilities of all these factors, and the correlations among them.

The correlated risk reported in Table 5 is

$$Risk_{Spreads}^{Correlated} = [Y \times S + S \times S] / \sqrt{VAR(P)}$$

that is, we sum all elements in the row of interest (row 1 for Y , row 2 for S) from the matrix above, and normalize it by the standard deviation of the portfolio. This statistic (1) takes into account correlations and (2) ensures that the correlated risks of all factors add up to the total risk of the portfolio ($Risk_{Curve}^{Correlated} + Risk_{Spreads}^{Correlated} = \sqrt{VAR(P)} = STD(P)$).⁴

The generic analysis we just performed constitutes the first step into the description of the risk associated with a portfolio. The analysis refers to categories of risk factors (such as “curve” or “spreads”). However, a factor-based risk model allows for a significantly deeper analysis of the imbalances the portfolio may have. Each of the risk categories referred to above can be described with a rich set of detailed risk factors. Typically in a fixed income factor model, each asset class has a specific set of risk factors, in addition to the potential set of factors common to all (e.g., curve factors). These asset-specific risk factors are designed to capture the particular sources of risk the asset class is exposed to. In the following section, we

go through a *risk report* built in such a way, emphasizing risk factors that are common or particular to the different asset classes. Along the way, we demonstrate how the report offers insights from both a risk management and a *portfolio construction* perspective.

A Detailed Risk Report

In this section, we continue the analysis of the portfolio introduced previously, a 50-bond portfolio benchmarked against the Barclays US Aggregate Index. The report package we present was generated using POINT[®], Barclays cross-asset portfolio analysis and construction system, and gives a very detailed picture of the risk embedded in the portfolio. The package is divided into four types of reports: summary reports, factor exposure reports, issue/issuer level reports, and scenario analysis reports. Some of the information we reviewed earlier can be thought of as summary reports.

Summary Report

Table 6 illustrates a typical risk summary statistics report. It shows that the portfolio has 50 positions, but from only 27 issuers. This number implies limited ability to diversify idiosyncratic risk, as we will see below. The report confirms that the portfolio is long duration (OAD of 4.55 years versus 4.30 years for the benchmark) and has higher yield (yield to worst of 3.71% versus 2.83% for the benchmark) and coupon (4.73% versus 4.46% for the benchmark).

The table also reports that the total volatility of the portfolio (163.3 bps/month) is higher than that of the benchmark (158.1 bps/month). This is not surprising: longer duration, higher spread and less diversification all tend to increase the volatility of a portfolio. Because of its higher volatility, we refer to the portfolio as riskier than the benchmark. Looking into the different components of the portfolio's total volatility, the table reports that the idiosyncratic volatility of the portfolio is significantly smaller

Table 6 Summary Statistics Report

	Portfolio	Benchmark	
A. Parameter			
Positions	50	8,191	
Issuers	27	787	
Currencies	1	1	
Market value (\$ millions)	200	14,762	
Notional (\$ millions)	187	13,750	
B. Analytics			
	Portfolio	Benchmark	Difference
Coupon	4.73	4.46	0.27
Average life	6.63	6.35	0.27
Yield to worst	3.71	2.83	0.88
Spread	157	57	100
Duration	4.55	4.30	0.25
Vega	-0.02	-0.01	-0.01
Spread duration	4.67	4.56	0.11
Convexity	-0.15	-0.29	0.13
C. Volatility			
	Portfolio	Benchmark	TEV
Systematic	162.9	158.0	9.3
Idiosyncratic	11.1	5.6	10.1
Total	163.3	158.1	13.7
D. Portfolio Beta			
			1.03

than that of the systematic (11.1 bps/month versus 162.9 bps/month, respectively). This is also expected from a portfolio of investment-grade bonds. Given the fact that by construction the systematic and idiosyncratic components of risk are independent, we can calculate the total volatility of the portfolio as

$$TEV_{PTF} = \sqrt{162.9^2 + 11.1^2} = 163.3$$

There are two interesting observations regarding this number: first, the total volatility is smaller than the sum of the volatilities of the two components. This is the diversification benefit that comes from combining independent sources of risk. Second, the total volatility is very close to the systematic one. This may suggest that the idiosyncratic risk is irrelevant. That is an erroneous and dangerous conclusion. In particular, when managing against a benchmark, the focus should be on the net exposures and risk, not on their absolute

Table 7 Factor Partition—Risk Analysis

Risk Factor Group	Isolated TEV	Contribution to TEV	Liquidation Effect on TEV	TEV Elasticity (%)
Total	13.7	13.7	-13.7	1.0
Systematic risk	9.3	6.3	-3.6	0.5
Curve	8.5	4.0	-1.5	0.3
Volatility	1.7	0.1	0.0	0.0
Government-related spreads	3.0	0.1	0.2	0.0
Corporate spreads	5.1	1.6	-0.7	0.1
Securitized spreads	3.0	0.5	-0.2	0.1
Idiosyncratic risk	10.1	7.4	-4.4	0.5

counterparts. In Table 6 the total TEV is reported as 13.7 bps/month. This means that the model forecasts the portfolio return to be typically no more than 14 bps/month higher or lower than the return of the benchmark. This number is in line with the risk budget of our manager. The exhibit also reports idiosyncratic TEV of 10.1 bps/month, which is greater than the systematic TEV (9.3). When measured against the benchmark, our major source of risk is idiosyncratic, contrary to the conclusion one could draw by looking only at the portfolio's volatility. The TEV of our portfolio is also bigger than the difference between the volatilities of the portfolio and benchmark. Again, this is not surprising: The volatility depends on the absolute exposures, while the TEV measures imbalances between these absolute exposures from the portfolio and the benchmark. For the TEV what matters most is the correlation between these absolute exposures. Depending on this correlation, the TEV may be smaller or bigger than the difference in volatilities.

Finally, the report estimates the portfolio to have a beta of 1.03 to the benchmark. This statistic measures the co-movement between the portfolio and the benchmark. We can read it as follows: The model forecasts that a movement of 10 bps in the benchmark leads to a movement of 10.3 bps in the portfolio in the same direction. Note that a beta of less than one does not mean that the portfolio is less risky than the benchmark. In the limit, if the portfolio and benchmark are uncorrelated, the port-

folio beta is zero but obviously that does not mean that the portfolio has zero risk. Finally, one can compute many different "betas" for the portfolio or subcomponents of it.⁵ A simple and widely used one is the "duration beta," given by the ratio of the portfolio duration to that of the benchmark. In our case this ratio is $4.55/4.30 = 1.06$. This implies that the portfolio has a return from yield curve movements around 1.06 times larger than that of the benchmark. This beta is larger than the portfolio beta (1.03), meaning that net exposures to other factors (e.g., spreads) "hedge" the portfolio's curve risk.

This first summary report (Table 6) allows us to get a glimpse into the risk of the portfolio. However, we want to know in more detail what the source of this risk is. To do that, we turn to the next two summary reports. In the first, risk is partitioned across different groups of risk factors. In the second, the partition is across groups of securities/asset classes.

Table 7 shows four different statistics associated with each set of risk factors. The first two were somewhat explored in Tables 4 and 5.⁶ The exhibit reports in the first column the isolated TEV, that is, the risk associated with that particular set of risk factors only. We see that in an isolated analysis, the systematic and idiosyncratic risks are balanced, at 9.3 and 10.1 respectively. The report also shows the isolated risk associated with the major components of systematic risk. As discussed before, all components of systematic risk have nontrivial isolated risk, but only curve and credit spreads

Table 8 Security Partition—Risk Analysis I

Security Partition Bucket	NMW (%)	Contribution to TEV		
		Systematic	Idiosyncratic	Total
Total	0.0	6.3	7.4	13.7
Treasuries	-2.0	2.9	0.2	3.1
Government agencies	-5.4	0.5	0.4	0.9
Government nonagencies	-1.0	-1.4	0.1	-1.3
Corporates	12.4	3.4	4.3	7.7
MBS	-5.8	0.9	0.8	1.7
ABS	-0.3	0.0	0.0	0.0
CMBS	2.1	0.0	1.6	1.6

are significant when we look into the *contributions* to TEV. If we look across factors, the major contributors are idiosyncratic risk, curve, and credit spreads. Other systematic exposures are relatively small.

Another look into the correlation comes when we analyze the liquidation effect reported in the table. This number represents the change in TEV when we completely hedge that particular group of risk factors. For instance, if we hedge the curve component of our portfolio, our TEV drops by 1.5 bps/month, from 13.7 to 12.2. One may think that the drop is rather small, given the magnitude of isolated risk the curve represents. However, if we hedge the curve, we also eliminate the beneficial effect the negative correlation between curve and spreads have on the overall risk of the portfolio. Therefore, we have a more limited impact when hedging the curve risk. In fact, for this portfolio we see that hedging any particular set of risk factors has a limited effect in the overall risk.

The TEV elasticity reported in the last column gives another perspective into how the TEV in the portfolio changes when we change the risk loadings. Specifically, it tells us what the percentage change in TEV would be if we changed our exposure to that particular set of factors by 1%. We can see that if we reduce our exposure to corporate spreads by 1%, our TEV would decrease by 0.1%.

We perform a similar analysis in Table 8, but applied to a security partition. That is, instead of looking at individual sources of risk (e.g.,

curve) across all securities, we now aggregate all sources of risk within a security and report analytics for different groups of these securities (e.g., subportfolios). In particular, Table 8 reports the results by asset class. We can see that the majority of risk (7.7 bps/month) is coming from the corporate component of the portfolio.⁷ Corporates are also the primary contributors to the portfolio's systematic and idiosyncratic components of risk. This is not surprising, given the portfolio's large net market weight (NMW) to this sector. There are two other important sources of risk. The first is the Treasuries subportfolio, with 3.1 bps/month of risk. This risk comes mainly from the mismatch in duration. The second comes from the idiosyncratic risk of the CMBS component of the portfolio. Even though the NMW and systematic risk are not significant for this asset class, the relatively small number of (risky) CMBS positions in the portfolio causes it to have significant idiosyncratic risk (three securities in the portfolio versus 1,735 in the index). Since the portfolio manager is trying to replicate a very large benchmark with only 50 positions, she has to be very confident in the issuers selected. This report highlights the significant name risk the portfolio is exposed to.

Table 9 completes the analysis, reporting other important risk statistics about the different asset classes within the portfolio. These statistics mimic the analysis done in terms of risk factor partitions in Table 7, so we will not repeat their definitions. We focus on the

Table 9 Security Partition—Risk Analysis II

Security Partition Bucket	Isolated TEV	Liquidation Effect on TEV	TEV Elasticity (%)
Total	13.7	−13.7	1.0
Treasuries	7.4	−1.1	0.2
Government agencies	9.1	2.0	0.1
Government nonagencies	6.7	2.7	−0.1
Corporates	15.2	0.6	0.6
MBS	5.8	−0.5	0.1
ABS	1.1	0.1	0.0
CMBS	5.1	−0.7	0.1

numbers. In particular, the isolated TEV from the corporate sector is 15.2 bps/month, higher than the total risk of the portfolio. This means that the exposures to the other asset classes, on average, hedge our credit portfolio. The exhibit also reports that the agencies isolated risk is very large. This is due to the large negative net exposure (−5.4%) we have to this asset class. But the risk is fully hedged by the other exposures of the portfolio (e.g., long exposure to credit or long duration on Treasuries), so overall the risk contribution of this asset class is small, as previously discussed. We can even take the analysis a bit further: Table 9 shows us through the liquidation effect that if we eliminate the imbalance the portfolio has on agencies, we actually would increase the total risk of the portfolio by 2.0 bps/month. In short, we would be eliminating the hedge this asset class provides to the global portfolio, therefore increasing its risk. The exposures to this asset class were clearly built to counteract other exposures in the portfolio. Finally, Table 9 also reports the TEV elasticity of the different components of the portfolio. This number represents the percentage change in TEV if the NMW to that sub-portfolio changes by 1%, so we need to read the numbers with an opposite sign if the NMW is negative. In particular, if we increase the weight of the agency portfolio in absolute value (making it “more short”) by 1%, we would actually increase the TEV by 0.1%. This result shows that

the position in agencies provides hedging “on average,” but marginally it is already increasing the risk of the portfolio. In other words, the hedging went beyond its optimal value.

This set of summary reports gives us a very clear picture of the major sources of risk and how they relate to each other. In what follows, we focus on the more detailed analysis of the individual systematic sources of risk.

Factor Exposure Reports

At the heart of a multifactor risk model is the definition of the set of systematic factors that drive risk across the portfolio. As described above, there are different types of risk a fixed income portfolio is exposed to. In what follows, we focus on the three major types: curve, credit, and prepayment risk. Specifically in what regards the latter two, we use the credit and MBS component of the portfolio, respectively, to illustrate how to measure risks along these dimensions. Moreover, to keep the example simple, we show only a partial view of all relevant factors for these sources of risk. Later in this section we refer briefly to other sources of risk a fixed income portfolio may be exposed to.

Curve Risk As the previous analysis shows (e.g., Table 7), curve is the major source of risk in our portfolio. This kind of risk is embedded in virtually all fixed income securities (exceptions are, for instance, floaters and distressed securities), therefore mismatches are very penalizing.

When analyzing curve risk, we should use the curve of reference we are interested in. Depending on the portfolio and circumstances, this is typically the government or swap curve.⁸ In calm periods, the behavior of the swap curve tends to match that of the government curve. However, during liquidity crises (e.g., the Russian crisis in 1998 or the credit crisis in 2008), they can diverge significantly. To capture these different behaviors adequately, we analyze curve risk using the following decomposition: For government products, the curve

risk is assessed using the government curve. For all other products in our portfolio (that usually trade off the swap curve), this risk is measured using both the Treasury curve and swap spreads (i.e., the spreads between the swap and the government curve). Other decompositions are also possible.

The risk associated with each of these curves can be described by the exposure the portfolio has to different points along the curve and how volatile and correlated the movement in these points of the curve are. A additional convexity term is sometimes used to capture the non-linear components of curve risk. For a typical portfolio, a good description of the curve can be achieved by looking at a relatively small number of points along the curve (called key rates), for example, 6-month, 2-year, 5-year, 10-year, 20-year, and 30-year. An alternative set of factors used to capture yield curve risk can be defined using statistical analysis of the historical realizations of the various yield curve points. The statistical method used most often is called principal component analysis (PCA). This method defines factors that are statistically independent of each other. Typically three or four such factors are sufficient to explain the risk associated with changes of yields across the yield curve. PCA analysis has several shortcomings and must be used with caution. Using a larger set of economic factors, such as the key rate points described above, is more intuitive and captures the risk of specialized portfolios better. In our analysis, we follow the key rates approach.

Table 10 details the risk in our portfolio associated with the US Treasury curve. It starts by describing all risk factors our portfolio or benchmark load on. As discussed above, we identify the six key rate (KR) points in the curve plus the convexity term as the risk factors associated with US Treasury risk. They are described in the first column of panel A in the exhibit. They measure the risk associated with moves in that particular point in the curve. Exposure to these risk factors is measured by the key rate dura-

tions (KRD) for each of the six points. The description of the loading is in the second column of the exhibit, while its value for the portfolio, benchmark, and the difference is displayed in the next columns. Key rate durations are also called partial durations, as they add up to approximately the duration of the portfolio. Their loadings are constructed by aggregating partial durations across (virtually) all the securities. For instance, for our portfolio, the sum of the key rate durations is $0.14 + 0.86 + 1.30 + 0.77 + 1.02 + 0.47 = 4.56$, very close to the total duration of our portfolio.

Looking at the table, we see significant mismatches in the duration profiles between our portfolio and its benchmark, namely at the 10-year and 20-year points on the curve. Specifically, we are short 0.41 years at the 10-year point and long 0.53 years at the 20-year point. How serious is this mismatch? Looking at the factor volatility column, it can be seen that these points on the curve have been very volatile at around 40 bps/month. If we interpret this volatility as a typical move, the first two columns of panel B show us the potential impact of such a movement in the return of our portfolio, net of benchmark. For instance, a typical move up (+44.2 bps/month) in the 10-year point of the Treasury curve, when considered in isolation, will deliver a positive net return of 15.9 bp.⁹ In isolation, the positive impact is expected because we are short that point of the curve. More interesting may be the correlated number on the exhibit. It states the return impact but in a correlated fashion. In the scenario under analysis, a movement in the 10-year point will almost certainly involve a movement of the neighboring points in the curve. So, contrary to the positive isolated effect documented above, the correlated impact of a change up in the 10-year point is actually negative, at -5.0 bps. This result is in line with the overall positive duration exposure the portfolio has: General (correlated) movements up in the curve have negative impact in the portfolio's performance.¹⁰ Finally, and broadly

Table 10 Treasury Curve Risk

A. Exposures and Factor Volatility					
Factor Name	Units	Exposure			Factor Volatility
		Portfolio	Benchmark	Net	
USD 6M key rate	KRD (Yr)	0.14	0.15	-0.01	36.0
USD 2Y key rate	KRD (Yr)	0.86	0.70	0.15	38.0
USD 5Y key rate	KRD (Yr)	1.30	1.25	0.05	44.3
USD 10Y key rate	KRD (Yr)	0.77	1.13	-0.36	44.2
USD 20Y key rate	KRD (Yr)	1.02	0.53	0.49	39.6
USD 30Y key rate	KRD (Yr)	0.47	0.53	-0.06	39.7
USD convexity	OAC	-0.15	-0.29	0.13	8.4

B. Other Risk Statistics				
Factor Name	Return Impact of a Typical Move		Marginal Contribution to TEV	TEV Elasticity (%)
	Isolated	Correlated		
USD 6M key rate	0.5	-2.4	6.3	0.0
USD 2Y key rate	-5.8	-4.5	12.2	0.1
USD 5Y key rate	-2.0	-4.5	14.5	0.0
USD 10Y key rate	15.9	-5.0	15.9	-0.4
USD 20Y key rate	-19.5	-5.2	14.9	0.5
USD 30Y key rate	2.5	-5.2	14.8	-0.1
USD convexity	1.1	2.0	1.2	0.0

speaking, the (negative of the) ratio of the correlated impact to the factor volatility gives us the model-implied partial empirical duration of the portfolio. For instance, if we focus on the 10-year point, we get $-(-5.0/44.2) = 0.11$. This smaller empirical duration is typical in portfolios with spread exposure. The spread exposure tends to empirically hedge some of the curve exposure, given the negative correlation between these two sources of risk. Finally, the exhibit shows the risk associated with convexity. We can see that the benchmark is significantly more negatively convex, so the portfolio is less responsive than the benchmark to higher order changes in the yield curve.

There are many other statistics of interest one can analyze regarding the Treasury curve risk of the portfolio. Portfolio managers frequently have questions such as: If I want to reduce the risk of my portfolio by manipulating my Treasury curve exposure, what should I change? What is the most effective move? By how much would my risk actually change? The

statistics reported in the columns “Marginal Contribution to TEV” and “TEV Elasticity (%)” of panel B are typically used to answer these questions. Regarding the marginal contributions, the 10-year point has the largest value, indicating that an increase (reduction) of one unit of exposure (in this case one year of duration) to the 10-year point leads to an increase (reduction) of around 16 bps in the TEV.¹¹ In other words, if we want to reduce risk by manipulating our exposure to the yield curve, the 10-year point seems to present the fastest track. In addition, the exhibit shows that all Treasury risk factors are associated with positive marginal contributions. This means that an increase in the exposure to any of these factors increases the risk (TEV) of the portfolio. This conclusion holds, even for factors for which we have negative exposure (e.g., the 10-year key rate). The reason behind this result is our overall long duration exposure. If we add exposure to it, regardless of the specific point where we add it, we extend our duration

Table 11 Swap Spread (SS) Risk

Factor Name	Exposure (SS-KRD)			Factor Volatility	Return Impact Correlated	Marginal Contribution to TEV
	Portfolio	Benchmark	Net			
6M SS	0.14	0.13	0.01	39.1	-2.1	5.8
2Y SS	0.52	0.47	0.04	20.4	-2.1	3.0
5Y SS	0.84	0.75	0.09	9.6	-2.0	1.4
10Y SS	0.71	0.68	0.03	14.1	1.7	-1.8
20Y SS	0.34	0.33	0.01	17.0	2.2	-2.7
30Y SS	0.06	0.20	-0.15	20.1	2.4	-3.5

even further, increasing the mismatch our portfolio has in terms of duration, and so increasing its risk.¹² This result holds because we take into consideration the correlations between the different points in the Treasury curve. Without correlations, the analysis would be significantly less clear. The exhibit also reports the TEV elasticity of each of the risk factors, a concept introduced earlier. The interpretation is similar to the marginal contribution, but with normalized changes (percentage changes). This normalization makes the numbers more comparable across risk factors of very different nature. It is also useful when considering leveraging the entire portfolio proportionally. In our case, if we increase the exposure to the 10-year key rate point by 10%, from -0.36 to something around -0.40 (effectively reducing our long duration exposure), our TEV would be reduced by 4% (from 13.7 to 13.2 bps/month).

We now turn the analysis to the other component of the curve risk described above: the risk embedded into the portfolio exposure to the swap spread, that is, the spread between the swap and the Treasury curves. All securities that trade against the swap curve (e.g., all typical credit and securitized bonds) are exposed to this risk. Its analysis follows very closely that of the Treasury curve, so we only highlight the major risk characteristics of the portfolio along this dimension. Table 11 shows that in general our exposure to the swap spreads is smaller than the exposure to the Treasury curve. Remember that Treasuries do not load on this set of risk factors, so the market-weighted exposures are conse-

quently smaller. Looking at the profile of factor volatilities, one can see that its term structure of volatilities is U-shaped, with the short end extremely volatile and the five-year point having the lowest volatility. When comparing with the Treasury curve volatility profile (see Table 10), we can see significant differences, the aftermath of a strong liquidity crisis. Regarding net exposures, the exhibit shows that our largest mismatch is at the 30-year point, where we are short by 0.15 years. Interestingly, this is not the most expensive mismatch in terms of risk: When looking at the last column, we see that we would be able to change risk the most by manipulating the short end of our exposure to the swap spread curve, namely the six-month point.

The previous tables allow us to understand our exposures to the different types of curve risk and their impact both on the return and risk of our portfolios. They also guide us regarding what changes we can introduce to modify the risk profile of the portfolio. We now turn our attention to sources of risk that are more specific to particular asset classes. In particular, we start with the analysis of credit risk.

Credit Risk Instruments in the portfolio issued by corporations or entities that may default are said to have credit risk. The holders of these securities demand some extra yield—on top of the risk-free yield—to compensate for that risk. The extra yield is usually measured as a spread to a reference curve. For instance, for corporate bonds the reference curve is usually the swap curve. The level of credit spreads

determines to a large extent the credit risk exposure associated with the portfolio.¹³

There are several characteristics of credit bonds that are naturally associated with systematic sources of credit spread risk. For instance, depending on the business cycle, particular industries may be going through especially tough times. So industry membership is a natural systematic source of risk. Similarly, bonds with different credit ratings are usually treated as having different levels of credit risk. Credit rating could be another dimension we can use to measure systematic exposure to credit risk. Given these observations, it is common to see factor models for credit risk using industry and rating as the major systematic risk factors. Recent research suggests that risk models that directly use the spreads of the bonds instead of their ratings to assess risk perform

better for relatively short/medium horizons of analysis.¹⁴ Under this approach, the loading of a particular bond to a credit risk factor would be the commonly used spread duration multiplied by the bond's spread (the loading is termed $DTS = \text{Duration Times Spread} = OASD \times OAS$). By directly using the spread of the bond in the definition of the loading to the credit risk factors we do not need to assign specific risk factors to capture the rating or any similar quality-like effect. It will be automatically captured by the bond's loading to the credit risk factor and will adjust as the spread of the bond changes. We use different systematic risk factors only to distinguish among credit risk coming from different industries.¹⁵

The results of such an approach to the analysis of our portfolio are displayed in Table 12, which shows the typical industry risk factors

Table 12 Credit Spread Risk

Factor Name	Exposure (DTS)			Factor Volatility	Return Impact Correlated	Marginal Contribution to TEV
	Portfolio	Benchmark	Net			
IND Chemicals	0.00	0.03	-0.03	15.01	-0.39	0.43
IND Metals	0.00	0.06	-0.06	20.01	-0.16	0.23
IND Paper	0.00	0.01	-0.01	17.04	-0.40	0.49
IND Capital Goods	0.00	0.05	-0.05	14.98	-0.02	0.02
IND Div. Manufacturing	0.00	0.03	-0.03	14.21	-0.62	0.64
IND Auto	0.00	0.01	-0.01	22.18	-0.53	0.85
IND Consumer Cyclical	0.10	0.05	0.06	17.05	-0.26	0.32
IND Retail	0.00	0.05	-0.05	16.95	0.14	-0.17
IND Cons. Non-cyclical	0.00	0.13	-0.13	14.62	-0.22	0.24
IND Health Care	0.00	0.02	-0.02	14.07	0.13	-0.13
IND Pharmaceuticals	0.19	0.06	0.12	15.13	-0.34	0.37
IND Energy	0.12	0.20	-0.07	16.39	-0.29	0.34
IND Technology	0.00	0.06	-0.06	15.52	-0.11	0.12
IND Transportation	0.00	0.05	-0.05	15.09	-0.26	0.29
IND Media Cable	0.24	0.06	0.18	15.83	0.51	-0.58
IND Media Non-cable	0.00	0.04	-0.04	15.94	0.20	-0.23
IND Wirelines	0.09	0.17	-0.08	15.26	0.41	-0.45
IND Wireless	0.00	0.03	-0.03	14.87	1.06	-1.13
UTI Electric	0.28	0.20	0.08	15.79	-0.16	0.18
UTI Gas	0.09	0.10	-0.01	18.51	-0.41	0.55
FIN Banking	0.88	0.56	0.32	18.61	1.19	-1.59
FIN Brokerage	0.00	0.02	-0.02	15.90	1.47	-1.68
FIN Finance Companies	0.08	0.10	-0.02	20.64	0.68	-1.01
FIN Life & Health Insurance	0.12	0.11	0.01	19.96	0.58	-0.84
FIN P&C Insurance	0.00	0.06	-0.06	11.76	0.34	-0.29
FIN Reits	0.14	0.04	0.10	17.68	0.80	-1.02
Non Corporate	0.06	0.23	-0.17	25.27	0.28	-0.50

Table 13 Risk per Rating

Rating	NMW (%)	TEV				Systematic Beta
		Contribution	Isolated	Liquidation	Elasticity (%)	
Total	0.0	13.7	13.7	-13.7	1.0	1.03
AAA	-7.2	10.9	37.4	22.2	0.8	1.12
AA1	-0.3	-0.2	1.0	0.2	0.0	0.00
AA2	0.2	0.3	3.3	0.1	0.0	1.10
AA3	-2.3	-1.3	6.7	2.6	-0.1	0.00
A1	-0.5	0.3	4.2	0.4	0.0	1.51
A2	7.1	3.6	11.2	1.0	0.3	0.77
A3	4.7	1.7	5.8	-0.5	0.1	0.65
BAA1	-0.1	0.3	3.7	0.2	0.0	1.51
BAA2	-3.3	-2.3	11.5	5.9	-0.2	0.00
BAA3	1.7	0.3	7.7	1.7	0.0	0.37

associated with credit risk. The portfolio has net positions in 27 industries, spanning all three major sectors: Industrials (IND), Utilities (UTI) and Financials (FIN). We saw before that we have a significant net exposure to financials in terms of market weights (12.2%, see Table 1). In terms of risk exposure, Table 12 shows that the net DTS exposure to the Banking industry is 0.32, clearly the highest across all sectors.¹⁶ However, the marginal contribution to TEV that comes from that industry, although high, is comparable to other industries, namely Brokerage, for which the net exposure is close to zero. This means that these two industries are close substitutes in terms of the current portfolio holdings. Actually, what is very interesting is the fact that the marginal contribution is negative for these industries, even though we are significantly overweighting them. The analysis suggests that if we increase our risk exposure to Banking, our risk would actually decrease. This result is again driven by the strong negative correlation between spreads in financials and the yield curve. Therefore, the exposure in banking is actually helping hedge out our (more risky) long duration position. This kind of analysis is only possible when you account for the correlations across factors. It is of course also dependent on the quality of the correlation estimations the model has.

Although the risk factors used to measure risk are predetermined in a linear factor model,

there is extreme flexibility in the way the risk numbers can be aggregated and reported.¹⁷ For example, as explained above, the risk model we use to generate the current risk reports does not use credit ratings as drivers of systematic credit risk. Instead, it relies on the DTS concept. However, once generated, the risk numbers can be reported using any portfolio partition. As an example, Table 13 shows the risk breakdown by rating. As reported in this table, the majority of risk is coming from our AAA exposure (10.9 bps/month), the bucket with the biggest mismatch in terms of net weight (-7.2%). This bucket includes Treasury and government-related securities, sectors that are underweighted in the portfolio leading to significant risk. This is even clearer when we look into the isolated TEV numbers. If we had mismatches only on AAAs, the risk of our portfolio would be 37.4 bps/month, instead of the actual 13.7: our other exposures (namely the one to single As) hedge the risk from AAAs. This table also reports the systematic betas associated with each of the rating subportfolios. These betas add up to the portfolio beta, when we use the portfolio weights (not NMW) as weights in the summation. Systematic betas of zero identify buckets for which the portfolio has (close to) no holdings. The table shows that a movement of 10 basis points in the benchmark leads to a 11.2 basis points return in the AAA subcomponent of the portfolio. The beta of 0.37

Table 14 MBS (spread) Prepayment Risk

Factor Name	Exposure (OASD)			Factor Volatility	Return Impact Correlated	Marginal Contribution to TEV
	Portfolio	Benchmark	Net			
MBS New Discount	0.00	0.00	0.00	36.8	-1.2	3.3
MBS New Current	0.00	0.04	-0.04	24.5	-0.3	0.6
MBS New Premium	0.38	0.59	-0.21	29.7	-0.1	0.3
MBS Seasoned Current	0.00	0.00	0.00	25.5	-0.6	1.2
MBS Seasoned Premium	0.65	0.46	0.19	29.8	0.1	-0.2
MBS Ginnie Mae 30Y	0.31	0.21	0.10	6.1	-0.1	0.0
MBS Fannie Mae 15Y	0.00	0.11	-0.11	15.7	0.4	-0.4
MBS Ginnie Mae 15Y	0.00	0.01	-0.01	12.3	0.5	-0.4

for the BAA3 component of the portfolio does not signal low volatility for this subportfolio. It indicates mainly low correlation with the benchmark. This is probably due to a larger component of idiosyncratic risk for this set of bonds.

Prepayment Risk Securitized products are generally exposed to prepayment risk. The most common of the securitized products are the residential MBS (RMBS or simply MBS). These securities represent pools of deals that allow the borrower to prepay their debt before the maturity of the loan/deal, typically when prevailing lending rates are lower. This option means an extra risk to the holder of the security, the risk of holding cash exactly when reinvestment rates are low. Therefore, these securities have two major sources of risk: interest rates (including convexity) and prepayment risk.

Some part of the prepayment risk can be expressed as a function of interest rates via a prepayment model. This risk will be captured as part of interest-rate risk using the key rate durations and the convexity. These securities usually have negative convexity because usually prepayments increase (decrease) with decreasing (increasing) interest rates, thereby reducing price appreciation (increase price depreciation). The remaining part of prepayment risk—that is not captured by the prepayment model—must be modeled with additional systematic risk factors. Typically, the volatility of prepayment speeds (and therefore of risk) on MBS securities depends on three characteristics:

program/term of the deal, if the bond is priced at discount or premium (e.g., if the coupon on the bond is bigger than the current mortgage rates) and how seasoned the bond is. This analysis suggests that the systematic risk factors in a risk model should span these three characteristics of the securities.

Table 14 shows a potential set of risk factors that capture the three characteristics discussed above. Programs identified as having different prepayment characteristics are the conventional (Fannie Mae) 30-year bonds (the base case used for the analysis), the 15-year conventional (Fannie Mae) bonds, as well as the Ginnie Mae 30- and 15-year bonds. The age of bonds is captured by factors distinguishing between new and aged deals. Finally, each bond is also classified by the price of the security—discount, current, or premium. In this example there are no seasoned discounted bonds, given the unprecedented level of mortgage rates as of June 2010. In terms of risk exposures, the exhibit shows that we are currently underweighting 15-year conventional bonds, and overweighting 30-year Ginnie Mae bonds.

Interaction between Sources of Risk So far we analyzed the major sources of spread risk: credit and prepayment. To do this, we conveniently used two asset classes—credit and agency RMBS, respectively—where one can argue that these sources of risk appear relatively isolated. However, recent developments have made very clear that these sources of risk appear simultaneously in other major asset

classes, including non-agency RMBS, home equity loans and CMBS.¹⁸ When designing a risk model for a particular asset class, one should be able to anticipate the nature of the risks the asset class exhibits currently or may encounter in the future. The design and ability to segregate between these two kinds of risk depends also on the richness of the bond indicatives and analytics available to the researcher. For this last point, it is imperative that the researcher understands well the pricing model and assumptions made to generate the analytics typically used as inputs in a risk model. This allows the user to fully understand the output of the model, as well as its applicability and shortcomings.

Other Sources of Risk There are other sources of systematic risk we did not detail in this section. They may be important sources of risk for particular portfolios. Specific risk models can be designed to address them. We now mention some of them briefly.

Implied Volatility Risk Many fixed income securities have embedded options (e.g., callable bonds). This means that the expected future volatility (implied volatility¹⁹) of the interest rate or other discount curves used to price the security plays a role in the value of that option. If expected volatility increases, options generally become more expensive, affecting the prices of bonds with embedded options. For example, callable bonds will become cheaper with increasing implied volatility since the bond holder is short optionality (the right of the issuer to call the bond). Therefore, the exposure of the portfolio to the implied volatility of the yield curve is also a source of risk that should be accounted for. The sensitivity of securities to changes of implied volatilities is measured by vega, which is calculated using the security pricing model. Implied volatility factors can be either calculated by the market prices of liquid fixed income options (caps, floors, and swaptions), or implied by the returns of bonds with embedded options within each asset class.

Liquidity Risk Many fixed income securities are traded over-the-counter, in decentralized markets. Some trade infrequently, making them illiquid. It is therefore hard to establish their fair price. These bonds are said to be exposed to liquidity risk. The holder of illiquid bonds would have to pay a higher price to liquidate its position, usually meaning selling at a discount. This discount is uncertain and varies across the business cycle. For instance, the discount can be significant in a liquidity crisis, such as the one we experienced in 2008. The uncertainty about this discount means that, everything equal, a more illiquid bond will be riskier. This extra risk can be captured through liquidity risk factors. For instance, in the Treasury markets, one generally refers to the difference in volatility between an on-the-run and an off-the-run Treasury bond as liquidity risk.

Inflation Risk Inflation-linked securities are priced based on the expectation of future inflation. Uncertainty about this variable adds to the volatility of the bond over and above the volatility from other sources of risk, such as the nominal interest rates. Expected inflation is not an observed variable in the marketplace but can be extracted from the prices on inflation-linked government bonds and inflation swaps. Expected inflation risk factors can be constructed by summarizing this information. The sensitivity of securities to expected inflation is calculated using a specialized pricing model and is usually called inflation duration.

Tax-Policy Risk Many municipal securities are currently tax-exempt. This results in added benefit to their holders. This benefit—incorporated in the price of the security—depends on the level of exemption allowed. Uncertainty around tax policy—tax-policy risk—adds to the risk of these securities. Once again, tax-policy risk factors cannot be observed in the marketplace and must be extracted from the prices of municipal securities. The return of municipal securities in excess of interest rates is driven

partially by tax-policy expectations changes. However, it is also driven by changes in the creditworthiness of the municipal issuers as well as other factors. In this case it is difficult to separate tax-policy risk factors from other factors driving municipal bond spreads. Therefore, instead of specific tax-policy factors we usually extract factors representing the overall spread risk of municipal securities. This exercise is performed in a similar way to the credit risk model, where securities are partitioned into groups of “similar” risk by geography, bond-type (general obligation versus revenue), tax-status, and the like.²⁰

Issue-Level Reports

The previous analysis focused on the systematic sources of risk. We now turn our attention to the idiosyncratic or security-specific risk embedded in our portfolio. This risk measures the volatility the portfolio has due to news or demand–supply imbalances specific to the individual issues/issuers it holds. Therefore, the idiosyncratic risk is independent across issuers and diversifies away as the number of issues in the portfolio increases: Negative news about some issuers is canceled by positive news about others. For relatively small portfolios, the idiosyncratic risk may be a substantial compo-

nent of the total risk. This can be seen in our example, as our portfolio has only 27 issuers. Table 6 shows that the idiosyncratic volatility of our portfolio is 11.1 bps/month, more than twice the idiosyncratic volatility of the benchmark (5.6 bps/month). When looking at the tracking error volatility net of benchmark, Table 6 shows that our specific risk is 10.1 bps/month and larger than the systematic component (9.3 bps/month). This means that, typically, a major component of the monthly net return is driven by events affecting only individual issues or issuers. Therefore, monitoring these individual exposures is of paramount importance.

The idiosyncratic risk of each bond is a function of two variables: its net market weight and its idiosyncratic volatility. This last parameter depends on the nature of the bond issuer. For instance, a bond from a distressed firm has much higher idiosyncratic volatility than one from a government-related agency.

Table 15 provides a summary of the idiosyncratic risk for the top 10 positions by market weight in our portfolio. Not surprisingly, our top seven holdings are Treasuries and MBS securities, in line with the constitution of the index we are using as benchmark. Moreover, these positions have significant market weights, given that our portfolio contains

Table 15 Issue Specific Risk

Identifier	Ticker	Description	Maturity	Spread (bps)	Market Weight (%)		Idiosyncratic TEV
					Portfolio	Net	
912828KF	US/T	US Treasury Notes	2/28/2014	4	5.4	5.2	0.4
912828KJ	US/T	US Treasury Notes	3/31/2014	3	5.0	4.8	0.4
912828JW	US/T	US Treasury Notes	12/31/2013	1	4.7	4.5	0.4
912828KN	US/T	US Treasury Notes	4/30/2014	2	3.8	3.6	0.3
FNA04409	FNMA	FNMA Conventional Long T. 30yr	3/1/2039	20	3.2	1.1	0.4
FGB04409	FHLMC	FHLM Gold Guar Single F. 30yr	3/1/2039	25	2.7	1.1	0.4
912810FT	US/T	US Treasury Bonds	2/15/2036	−1	2.3	2.1	0.7
20029PAG	CMCSA	Comcast Cable Communication	5/1/2017	222	2.2	2.2	2.4
59018YSU	BAC	Merrill Lynch & Co.	2/3/2014	300	2.1	2.1	2.9
912828KV	US/T	US Treasury Notes	5/31/2014	1	2.1	1.9	0.2

only 50 positions. Even though we see large concentrations, the idiosyncratic TEV for the top holdings is small, as they are not exposed to significant name risk. The last column of the table shows that from this group the largest idiosyncratic risk comes from two corporate bonds (issued by Comcast Cable Communication “CMCSA” and Merrill Lynch “BAC”). This is not surprising, as these are the type of securities with larger event risk. Even within corporates, idiosyncratic risk can be quite diverse. In particular, it usually depends on the industry, duration, and level of distress of the issuer (usually proxied by rating, but in our model by the spread of the bond). For instance, the net position for both the CMCSA and BAC bonds is similar (2.2% and 2.1% respectively), but even though the maturity of the BAC bond is significantly shorter, its spread is higher, delivering a higher idiosyncratic risk (2.9 versus 2.4 bps/month). The fact that BAC is a firm from an industry (Financials) that experienced significant volatility in the recent past also contributes to higher idiosyncratic volatility. To manage the idiosyncratic risk in the portfolio one should pay particular attention to mismatches between the portfolio and benchmark for bonds with large spreads or long durations. These would tend to affect disproportionately the idiosyncratic risk of the portfolio.

Although important, the information in Table 15 is not enough to fully assess the idiosyncratic risk embedded in the portfolio. For instance, one could buy credit protection to BAC through a credit default swap (CDS). In this case, our exposure to this issuer may not be significant, even though, taken separately, the position reported in this exhibit is relevant. More generally, idiosyncratic risk is independent across issuers, but what happens within a particular issuer? A good risk model should have the ability to account for the fact that the idiosyncratic risk of two securities from the same issuer is correlated, as they are both subject to the same company-specific events. This is especially the case for corporates and emerg-

ing market securities. Moreover, it is important to note that the correlation between issues from the same issuer is not constant either. For an issuer in financial distress, all claims to their assets (bonds, equities, convertibles, etc.) tend to move together, in the absence of specific circumstances. This means that the idiosyncratic correlation between issues from that issuer should be high. Therefore, adding more issues from that issuer to the portfolio does not deliver additional diversification. On the other end, securities from firms that enjoy very strong financial wealth can move quite differently, driven by liquidity or other factors. In this case, one can have some diversification of idiosyncratic risk (although limited) even when adding issues from that same issuer into the portfolio.

To help us understand the net effect of all these points, we need to know the issuers that contribute the most to idiosyncratic risk. When aggregating risk from the issue (as shown in Table 15) to the issuer level, the correlations referred to above should be fully taken into account. Table 16 shows the results of this exercise for the 10 issuers with the highest idiosyncratic TEV. Our riskiest exposure comes from Johnson & Johnson (JNJ), with 3.7 bps/month of issuer risk. We can also observe that idiosyncratic TEV is not monotonic in the NMW: We have JNJ and President & Fellows of Harvard “HARVRD” with the same NMW, but the former is significantly more risky (3.7 versus 2.0 bps/month). It is possible to have important issuer risk even for names we do not have in our portfolio, if they have significant market weight in the benchmark. Finally, note that because the idiosyncratic risk across issuers is independent, we can easily calculate the cumulative risk of several issuers. For example, the total idiosyncratic risk of the first two issuers is given by

$$TEV_{idio}^{JNJ+D} = \sqrt{3.7^2 + 2.8^2} = 4.6$$

Another important interpretation from Table 16 is that these are our biggest name exposures in our portfolio. In this case, we are overweight in all of them. Therefore, we should

Table 16 Issuer Specific Risk

Ticker	Name	Sector	NMW (%)	Idiosyncratic TEV
JNJ	Johnson & Johnson	Pharmaceuticals	2.0	3.7
D	Dominion Resources Inc	Electric	1.8	2.8
CMCSA	Comcast Cable Communication	Media_cable	2.0	2.1
BBT	BB&T Corporation	Banking	2.0	2.1
HARVRD	Pres&Fellows of Harvard	Industrial_other	2.0	2.0
AXP	American Express Credit	Banking	1.7	1.8
MS	Morgan Stanley Dean Witter	Banking	1.3	1.7
C	Citigroup Inc	Banking	1.5	1.7
BAC	Merrill Lynch & Co.	Banking	1.6	1.6
RBS	Charter One Bank Fsb	Banking	1.6	1.4

not have negative views about any of them. If this is not the case, then we are assuming an unintended name risk. This risk should be promptly taken out of the portfolio, in favor of another issuer with similar characteristics and for which we do not have negative views about. This interactive exercise can easily be performed with a good and flexible optimizer.

Scenario Analysis Report

Scenario analysis is another useful way to gain additional perspective on the portfolio’s risk. There are many ways to perform this exercise. For instance, one may want to reprice the whole portfolio under a particular interest rate or spread scenario, and look at the hypothetical return under that scenario. Alternatively, one may look at the holdings of the portfolio

and see how they would have performed under particular stressed historical scenarios (e.g., the 1987 equity crash or the Asian crisis in 1997). One particular problem with this approach is the fact that, given the dynamic nature of the securities, the current portfolio did not exist with the current characteristics along all these historical episodes. A solution may be to try to price the current securities with the market variables at the time. Another solution is to represent the current portfolio as the set of loadings to all systematic risk factors in the factor risk model. We can then multiply these loadings by the historical realizations of the risk factors. The result is a set of historical systematic simulated returns. Figure 1 presents these returns for our portfolio over the last five years. As expected, the largest volatility came with the crisis of 2008, when the portfolio registered returns between –200

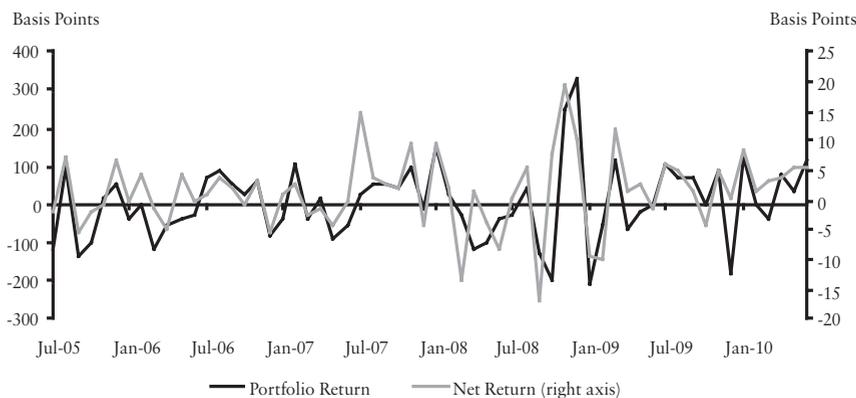


Figure 1 Historical Systematic Simulated Returns (basis points)

and +300 basis points. The largest underperformance against the benchmark appeared in September 2008, followed by the largest outperformance two months after, both at around 20 basis points.

This analysis has some limitations, especially for the portfolio under consideration, where idiosyncratic exposure is a major source of risk. This kind of risk is always very hard to pin down and obviously less relevant from an historical perspective, as the issuers in our current portfolio may have not witnessed any particular major idiosyncratic event in the past. However, these and other kinds of historical scenario analysis are very important, as they give us some indication of the magnitude of historical returns our portfolio might have encountered. They are usually the starting point for any stress testing. The researcher should always complement these with other nonhistorical scenarios relevant for the particular portfolio under analysis. One way to use the risk model to express such scenarios is discussed in the following section.

APPLICATIONS OF RISK MODELING

In this section, we illustrate several risk model applications typically employed for portfolio management. All applications make use of the fact that the risk model translates into a common, comparable set of numbers the imbalances the portfolio may have across many different dimensions. In some of the applications—*risk budgeting* and portfolio rebalancing—an optimizer that uses the risk model as an input is the optimal setting to perform the exercise.

Portfolio Construction and Risk Budgeting

Portfolio managers can be divided broadly into indexers (those that measure their returns rela-

tive to a benchmark index) and absolute return managers (typically hedge fund managers). In between stand the enhanced indexers we introduced previously in the entry. All are typically subject to a risk budget that prescribes how much risk they are allowed to take to achieve their objectives: Minimize transaction costs and match the index returns for the pure indexers, maximize the net return for the enhanced indexers, or maximize absolute returns for absolute return managers. In any of these cases, the manager has to merge all her views and constraints into a final portfolio. When constructing the portfolio, how can she manage the competing views, while respecting the risk budget? How can the views be combined to minimize the risk? What trade-offs can be made? Many different techniques can be used to structure portfolios in accordance with the manager's views. In particular, risk models are widely used to perform this exercise. They perform this task in a simple and objective manner: They can measure how risky each view is and how correlated they are. The manager can then compare the risk with the expected return of each of the views and decide on the optimal allocation across her views.

An example of a portfolio construction exercise using the risk model is the one we performed to construct the portfolio analyzed in the previous section.²¹ Figure 2 shows the exact problem we asked the optimizer to solve. We start the problem by defining an initial portfolio (empty in our case) and a tradable universe—the set of securities we allow the optimizer to buy or sell from. In our case, this is the Barclays US Aggregate index with issues having at least \$300 million of amount outstanding (in the Tradable Universe Options pane of the POINT[®] Optimizer window shown in Figure 2). The selection of this universe allows us to avoid having small issues in our portfolio, potentially increasing its liquidity. Pertaining to the risk model use (in the Objectives pane of the POINT[®] Optimizer window), the objective function used in the problem is to minimize Total TEV. This means that we are giving

Tradable Universe Options				
No.	Name	Type	Trade/Buy/Sell	Long/Short
1	Initial Portfolio	Initial Portfolio	Buy and Sell	Long/Short
2	US Agg 300 Mln (System)	Index	Buy and Sell	Long Only

Objectives				
<input checked="" type="radio"/> Minimize <input type="radio"/> Maximize				
No.	Attribute	Measure	Weight	Unit
1	Total TEV	Net vs Bmark	1.00	bps / mo
2	Systematic TEV	Net vs Bmark	0.00	bps / mo
3	Idiosyncratic TEV	Net vs Bmark	0.00	bps / mo

Common Constraints				
Final Portfolio Cash (base currency): <input type="radio"/> Long/Short <input type="radio"/> Long Only <input type="radio"/> Short Only <input checked="" type="radio"/> No Cash				
E...	Description	Measure	Bound	Unit
<input checked="" type="checkbox"/>	Budget: Final Portfolio Market Value	Change	100,000,000	USD
<input checked="" type="checkbox"/>	Final portfolio maximum gross size	Target		USD
<input checked="" type="checkbox"/>	Turnover: Maximum gross size of trades	Target		USD
<input checked="" type="checkbox"/>	Maximum number of securities in final portfolio			
<input checked="" type="checkbox"/>	Maximum number of trades		50	
<input checked="" type="checkbox"/>	Minimum trade size	Target		USD

Constraints on values aggregated by Buckets								
No.	Soft	...	Attribute	Universe	Measure	Lower Bound	Upper Bound	Unit
1	<input type="checkbox"/>		OAD	Final Portfolio	Net vs Bmark	0.25	0.30	yrs
2	<input type="checkbox"/>		OAS	Final Portfolio	Net vs Bmark	100.0	150.0	bps / yr

Constraints on each Issue/Issuer/Ticker									
Universe: Final Portfolio									
No.	Soft	Penalty	Attribute	Universe	For Each	Measure	Lower Bound	Upper Bound	Unit
1	<input type="checkbox"/>		Market Value [%]	Final Portfolio	Ticker	Net vs Bmark		2.00000	%

Figure 2 Portfolio Construction Optimization Setup in the POINT® Optimizer

leeway to the risk model to choose a portfolio from the tradeable universe that minimizes the risk relative to the benchmark, in our case the Barclays US Aggregate index. In the Common Constraints pane, additional generic constraints have been imposed: a \$100 million final portfolio with a maximum number of 50 securities. In the Constraints on values aggregated by Buckets pane, we force the optimizer to tilt our portfolio to respect the portfolio manager’s views: long duration against the benchmark between 0.25 and 0.30 years and spreads between 100 and 150 bps higher than the benchmark. In the Constraints on each Issue/Issuer/Ticker pane, we impose a maximum under-/overweight of 2% per issuer, to ensure proper diversification.²² The characteristics of the portfolio resulting from this optimization problem were extensively analyzed in the previous section.

Portfolio Rebalancing

Managers need to rebalance their portfolios regularly. For instance, as time goes by, the characteristics of the portfolio may drift away from targeted levels. This may be due to the aging of its holdings, market moves, or issuer-specific events such as downgrades or defaults. The periodic re-alignment of a portfolio to its investment guidelines is called portfolio rebalancing. Similar needs arise in many different contexts: when managers receive extra cash to invest, get small changes to their mandates, want to tilt their views, and the like. Similar to portfolio construction, a risk model is very useful in the rebalancing exercise. During rebalancing, the portfolio manager typically seeks to sell bonds currently held and replace them with others having properties more consistent with the overall portfolio goals. Such buy and sell transactions are costly, and their cost must

Common Constraints								
Final Portfolio Cash (base currency): <input type="radio"/> Long/Short <input type="radio"/> Long Only <input type="radio"/> Short Only <input checked="" type="radio"/> No Cash								
E...	Description	Measure	Bound	Unit				
<input checked="" type="checkbox"/>	Budget: Final Portfolio Market Value	Change		0 USD				
<input checked="" type="checkbox"/>	Final portfolio maximum gross size	Target		USD				
<input checked="" type="checkbox"/>	Turnover: Maximum gross size of trades	Target		USD				
<input checked="" type="checkbox"/>	Maximum number of securities in final portfolio							
<input checked="" type="checkbox"/>	Maximum number of trades			10				
<input checked="" type="checkbox"/>	Minimum trade size	Target		1,000,000 USD				
Constraints on values aggregated by Buckets								
No.	Soft	...	Attribute	Universe	Measure	Lower Bound	Upper Bound	Unit
1	<input type="checkbox"/>		OAD	Final Portfolio	Net vs Bmark	0.25	0.30	yrs
2	<input type="checkbox"/>		OAS	Final Portfolio	Net vs Bmark	100.0	150.0	bps / yr
3	<input type="checkbox"/>		Market Value [%]	Financial Inst. Banki...	Net vs Bmark	0.00000	5.00000	%

Figure 3 Portfolio Rebalancing Optimization Setup in the POINT[®] Optimizer

be weighted against the benefit from moving the portfolio closer to its initial specifications. A risk model can tell the manager how much risk reduction (or increase) a particular set of transactions can achieve so that she can evaluate the risk adjustment benefits relative to the transaction cost.

As an example, suppose our portfolio manager wants to tone down the heavy overweight she has on banking. She wants to cap that overweight to 5% and wants to do it with no more than 10 trades. Finally, assume she wants no change to the market value of the final portfolio. We can use a setup similar to that of Figure 2, but adjusting some of the constraints. Figure 3 shows two of the constraints option panes in the POINT[®] Optimizer window, changed to allow for the new constraints. Specifically, in the first panel, we allow for 10 trades and, in the second, included an extra constraint for the banking industry.

Table 17 shows the trading list suggested by the POINT[®] Optimizer. Not surprisingly, the biggest sells are of financial companies. To replace them, the optimizer—using the risk model—recommends more holdings of Treasury and corporate bonds. (We need these last to keep the net yield of the portfolio high.) Remember that we concluded that our financial holdings were highly correlated with Treasuries, so the proposed swap is not surprising.

Interestingly, the extra constraint imposed on the optimization problem did not materially

change the risk of the portfolio. Results show that the risk actually decreased to around 13 bps/month. This is due to the extra three positions added to the portfolio that now has 53 securities. These extra securities allowed the portfolio to reduce both its systematic as well as its idiosyncratic risk.

Scenario Analysis

As described in the previous section, scenario analysis is a very popular tool both for risk management and portfolio construction. In this section, we illustrate another way to construct scenarios, this time using the covariance matrix of the risk model. In this context, users express views on the returns of particular financial variables, indexes, securities, or risk factors, and the scenario analysis tool (using the risk model) calculates their impact on the portfolio's (net) return.

Typically in this kind of scenario analysis, the views one has are only partial views. This means we can have specific views on how particular macro variables, asset classes, or risk factors will behave; but we hardly have views on all risk factors the portfolio under analysis is exposed to. This is when risk models may be useful. At the heart of the linear factor models described in this entry is a set of risk factors and the covariance matrix between them. They are being increasingly used in the context of scenario analysis as a way to “complete”

Table 17 Proposed Trading List

BUYS			
Identifier	Description	Position Amount	Market Value
912828KV	US Treasury Notes	967,403	1,000,000
126650BK	CVS Corp-Global	1,696,069	1,518,408
98385XAJ	XTO Energy Inc	2,097,746	2,508,567
FNA05009	FNMA Conventional Long T. 30yr	2,547,359	2,708,258
912828KF	US Treasury Notes	3,786,070	3,882,263
Total			11,617,497
SELLS			
Identifier	Description	Position Amount	Market Value
16132NAV	Charter One Bank FSB	-3,229,847	-3,370,981
05531FAF	BB&T Corporation	-2,425,413	-2,499,505
0258M0BZ	American Express Credit	-2,021,013	-2,208,231
3133XN4B	Federal Home Loan Bank	-1,818,417	-2,085,812
740816AB	Pres&Fellows of Harvard	-1,281,616	-1,452,968
Total			-11,617,497

specific partial views or scenarios, delivering a full picture of the impact of the scenario in the return of the portfolio. Mechanically, what happens is the following: First, one translates the views into realizations of a subset of risk factors. Then the scenario is completed—using the risk model covariance matrix—to get the realizations of all risk factors. Finally, the portfolio's (net) loadings to all risk factors are used to get its (net) return under that scenario (by multiplying the loadings by the factor realizations under the scenario). This construction implies a set of assumptions that should be carefully understood. For instance, we assume that we can represent or translate our views as risk factor returns. So, if we have a view about the unemployment rate, and this is not a risk factor,²³ we cannot use this procedure to test our scenario. Also, to “complete” the scenario, we generally assume a stationary and normal multivariate distribution between all factors. These assumptions make this analysis less appropriate for looking at extreme events or regime shifts, for instance. But the analysis can be very useful in many circumstances.

As an example, consider using the scenario analysis to compute the model-implied empir-

ical durations (MED) of the portfolio we analyzed in detail previously in this entry. To do this, we express our views as changes in the curve factors. In our risk model, these are represented by the six key rate factors illustrated in Table 10. In particular, to calculate the model-implied empirical duration, we are going to assume that all six decrease by 25 bps/month, broadly in line with our managers' views.

Panel a of Table 18 shows that under this scenario, the portfolio returns 99 basis points, against the 93 of the benchmark. As expected given our longer duration, we outperform the benchmark. Due to the other exposures present in the portfolio and benchmark (e.g., spreads) and their average negative correlation with the curve factors, the duration implied by the scenario (MED) for our portfolio is only 3.96 (= 99/25) against the analytical 4.55. The scenario shows a similar decrease in the benchmark's duration.

Another characteristic imposed while constructing the portfolio was a targeted higher spread. As shown in Table 2, this resulted in an OAS for the portfolio of 157 bps against the 57 of the benchmark. It would be

Table 18 Spread Analysis

a. Analytical and Model-Implied Durations			
Universe	Return under Scenario (bp)	Durations	
		MED (scenario)	Analytical
Portfolio	99	3.96	4.55
Benchmark	93	3.72	4.30

b. Spread Contraction of 10%			
Universe		Restriction on YC movement	
		No Movement	Correlated
Portfolio		31	-3
Benchmark		32	0

interesting to evaluate the impact to the portfolio (net) return of a credit spread contraction of 10%. The portfolio is long spread duration (net OASD = 0.11, see Table 2), so we may expect our portfolio to outperform in this scenario. To do so, we analyze the results under two spread contraction scenarios: imposing no change in the yield curve (that is, an unchanged yield curve is part of the view) or allowing this change to be implied by the correlation matrix. (That is, the change in the yield curve is not part of the scenario. We have no views about it, but we allow it to change in a way historically consistent with our spread view.) Contrary to what one might expect, panel b of Table 18 shows that the effect in the net return is minimal under both scenarios. The higher spreads deliver no return advantage under this scenario. However, the absolute returns are quite different across the scenarios. When one allows the rates to move in a correlated fashion the net return drops close to zero: All positive return from the spread contraction is cancelled by the probable increase in the level of the curve and our long-duration exposure.

These very simple examples illustrate how one can look at reasonable scenarios to study the behavior of the portfolio or the benchmark under different environments. This scenario analysis does increase significantly the intuition the portfolio manager may have regarding the results from the risk model.

KEY POINTS

- Risk models describe the different imbalances of a portfolio using a common language. The imbalances are combined into a consistent and coherent analysis reported by the risk model.
- Risk models provide important insights regarding the different trade-offs existing in the portfolio. They provide guidance regarding how to balance them.
- Risk models in fixed income are unique in two different ways: First, the existence of good pricing models allows us to robustly calculate important analytics regarding the securities. These analytics can be used confidently as inputs into a risk model. Second, returns are not typically used directly to calibrate risk factors. Instead returns are first normalized into more invariant series (e.g., returns normalized by the duration of the bond).
- The fundamental systematic risk of all fixed income securities is interest rate and term structure risk. This is captured by factors representing risk-free rates and swap spreads of various maturities.
- Excess (of interest rates) systematic risk is captured by factors specific to each asset class. The most important components of such risk are credit risk and prepayment risk. Other risk factors that can be important are implied volatility, liquidity, inflation, and tax policy.

- Idiosyncratic risk is diversified away in large portfolios and indices but can become a very significant component of the total risk in small portfolios. The correlation of idiosyncratic risk of securities of the same issuer is nonzero and must be modeled very carefully.
- A good risk model provides detailed information about the exposures of a complex portfolio and can be a valuable tool for portfolio construction and management. It can help managers construct portfolios tracking a particular benchmark, express views subject to a given risk budget, and rebalance a portfolio while avoiding excessive transaction costs. Further, by identifying the exposures where the portfolio has the highest risk sensitivity it can help a portfolio manager reduce (or increase) risk in the most effective way.

NOTES

1. The Barclays Global Risk Model is available through POINT[®], Barclays portfolio management tool. It is a multi-currency cross-asset model that covers many different asset classes across the fixed income and equity markets, including derivatives in these markets. At the heart of the model is a *covariance matrix* of risk factors. The model has more than 700 factors, many specific to a particular asset class. The asset class models are periodically reviewed. Structure is imposed to increase the robustness of the estimation of such large covariance matrix. The model is estimated from historical data. It is calibrated using extensive security-level historical data and is updated on a monthly basis.
2. Later in this entry, we refer to this risk number as Isolated TEV.
3. We arrive at this number by taking the square root of the sum of squares of all the numbers in the table: $10.9 = (8.5^2 + 1.7^2 + 3.0^2 + 5.1^2 + 3.0^2)^{0.5}$. Moreover, this number would represent the total systematic risk of the portfolio. This definition is developed later in the entry.
4. In this example, we focus only on the systematic component of risk. Later, the normalization is with respect to the total risk of the portfolio, including idiosyncratic risk.
5. For example, see Table 13 later in this entry.
6. Note that the contribution numbers are different from those from Table 5 because there we reported the contribution to systematic—not total—risk.
7. This result does not contradict the findings in Table 7, where we see that curve is the major source of risk. Remember that the curve risk can come from our corporate subportfolio.
8. Other curves that can be used are, for instance, the municipals (tax free) curve, derivatives-based curves, and the like.
9. This number is obtained by simply multiplying the net exposure by the factor volatility. The sign of the move depends on the interpretation of the factor. In the case of the yield curve movements we know that $R = -KRD \times \Delta KR$. In our example $-(-0.36) \times 44.2 = 15.9$.
10. This reversal is clearly related to the fact that the 10-year and the 20-year points in the curve are usually highly correlated. In our case, our short position on the 10-year point is more than compensated by the positive exposure in the 20-year. Netting out, the 20-year effect (long duration) dominates when all changes are taken in a correlated fashion.
11. The marginal contribution is the derivative of the TEV with respect to the loading of each factor, so its interpretation holds only locally. Therefore, a more realistic reading may be that if we reduce the exposure to the 10-year by 0.1 years, the TEV would be reduced by around 1.6 basis points.
12. This is a rationale very similar to the one used before, where we see all correlated impacts with the same sign.
13. Spreads are also compensation for sources of risk other than credit (e.g., liquidity), but

- for the sake of our argument, we treat them primarily as major indicators of credit risk.
14. For details, see Ben Dor et al. (2010).
 15. The general principle of a risk model is that the historical returns of assets contain information that can be used to estimate the future volatility of portfolio returns. However, good risk models must have the ability to interpret the historical asset returns in the context of the current environment. This translation is made when designing a particular risk model/factor and delivers risk factors that are as invariant as possible. This invariance makes the estimation of the factor distribution much more robust. In the particular case of the DTS, by including the spread in the loading (instead of using only the typical spread duration), we change the nature of the risk factor being estimated. The factor now represents percentage change in spreads, instead of absolute changes in spreads. The former has a significantly more invariant distribution. For more details, see Silva (2009a).
 16. The DTS units used in the report are based on an OASD stated in years and an OAS in percentage points. Therefore, a bond with an OASD = 5 and an OAS = 200 basis points would have a DTS of $5 \times 2 = 10$. The DTS industry exposures are the weighted sum of the DTS of each of the securities in that industry, the weights being the market weight of each security.
 17. For a detailed methodology on how to perform this customized analysis, Silva (2009b).
 18. For a further discussion, see Gabudean (2009).
 19. The volatility is called implied because it is calculated from the market prices of liquid options with the help of an option-pricing model.
 20. For more discussion, see Staal (2009).
 21. The example is constructed using the POINT[®] Optimizer. For more details, refer to Kumar (2010).
 22. Another way to ensure diversification would be to include the minimization of the idiosyncratic TEV as a specific goal in the objective function.
 23. Unemployment rate is not used as a factor in most short- and medium-term risk models.

REFERENCES

- Ben Dor, A., Dynkin, L., Houweling, P., Hyman, J., van Leeuwen, E., and Penninga, O. (2010). A new measure of spread exposure in credit portfolios. *Barclays Publication*, February.
- Gabudean, R. (2009). U.S. home equity ABS risk model. *Barclays Publication*, October.
- Kumar, A. (2010). The POINT optimizer. *Barclays Publication*, June.
- Silva, A. (2009a). A note on the new approach to credit in the Barclays Global Risk Model. *Barclays Capital Publication*, September.
- Silva, A. (2009b). Risk attribution with custom-defined risk factors. *Barclays Publication*, August.
- Staal, A. (2009). US municipal bond risk model. *Barclays Publication*, July.

Financial Econometrics

Scope and Methods of Financial Econometrics

SERGIO M. FOCARDI, PhD

Partner, The Intertek Group

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Financial econometrics is the econometrics of financial markets. It is a quest for models that describe financial time series such as prices, returns, interest rates, financial ratios, defaults, and so on. The economic equivalent of the laws of physics, econometrics represents the quantitative, mathematical laws of economics. The development of a quantitative, mathematical approach to economics started at the end of the 19th century in a period of great enthusiasm for the achievements of science and technology. Robert Engle and Clive Granger, two econometricians who shared the 2003 Nobel Prize in Economics Sciences, have contributed greatly to the field of financial econometrics.

Econometrics is the branch of economics that draws heavily on statistics for testing and analyzing economic relationships. Within econometrics, there are theoretical econometricians who analyze statistical properties of estimators of models. Several recipients of the Nobel Prize in Economic Sciences received the award as a result of their lifetime contribution to this branch of economics. To appreciate the importance of econometrics to the discipline of economics, when the first Nobel Prize in Economic Sciences was awarded in 1969, the co-recipients were two econometricians, Jan Tinbergen and Ragnar Frisch (who is credited with first us-

ing the term “econometrics” in the sense that it is known today). Further specialization within econometrics, and the subject of this entry, is financial econometrics.

As Jianqing Fan (2004) writes, financial econometrics uses statistical techniques and economic theory to address a variety of problems from finance. These include building financial models, estimation and inferences of financial models, volatility estimation, risk management, testing financial economics theory, capital asset pricing, derivative pricing, portfolio allocation, risk-adjusted returns, simulating financial systems, and hedging strategies, among others.

In this entry, we provide an overview of financial econometrics and the methods employed.

THE DATA GENERATING PROCESS

The basic principles for formulating quantitative laws in financial econometrics are the same as those that have characterized the development of quantitative science over the last four centuries. We write mathematical models, that is, relationships between different variables and/or variables in different moments and different places. The basic tenet of quantitative science is that there are relationships that do not change regardless of the moment or the place under consideration. For example, while sea waves might look like an almost random movement, in every moment and location the basic laws of hydrodynamics hold without change. Similarly, asset price behavior might appear to be random, but econometric laws should hold in every moment and for every set of assets.

There are similarities between financial econometric models and models of the physical sciences, but there are also important differences. The physical sciences aim at finding immutable laws of nature; econometric models model the economy or financial markets—artifacts subject to change. For example, financial markets in the form of stock exchanges have been in operation for two centuries. During this period, they have changed significantly both in the number of stocks listed and the type of trading. And the information available on transactions has also changed. Consider that in the 1950s, we had access only to daily closing prices and this typically the day after; now we have instantaneous information on every single transaction. Because the economy and financial markets are artifacts subject to change, econometric models are not unique representations valid throughout time; they must adapt to the changing environment.

While basic physical laws are expressed as differential equations, financial econometrics uses both continuous-time and discrete-time models. For example, continuous-time models are used in modeling derivatives where both the underlying and the derivative price are represented by stochastic (i.e., random) differential equations. In order to solve stochastic differential equations with computerized numerical methods, derivatives are replaced with finite differences. (Note that the stochastic nature of differential equations introduces fundamental mathematical complications. The definition of stochastic differential equations is a delicate mathematical process invented, independently, by the mathematicians Ito and Stratonovich. In the Ito-Stratonovich definition, the path of a stochastic differential equation is not the solution of a corresponding differential equation. However, the numerical solution procedure yields a discrete model that holds pathwise. See Focardi and Fabozzi [2004].) This process of discretization of time yields discrete time models. However, discrete time models used in financial econometrics are not necessarily the result of a process of discretization of continuous time models.

Let's focus on models in discrete time, the bread-and-butter of econometric models used in asset management. There are two types of discrete-time models: static and dynamic. *Static models* involve different variables at the same time. The well-known capital asset pricing model (CAPM), for example, is a static model. *Dynamic models* involve one or more variables at two or more moments. (This is true in discrete time. In continuous time, a dynamic model might involve variables and their derivatives at the same time.) Momentum models, for example, are dynamic models.

In a dynamic model, the mathematical relationship between variables at different times is called the *data generating process* (DGP). This terminology reflects the fact that, if we know the DGP of a process, we can simulate the process recursively, starting from initial conditions.

Consider the time series of a stock price p_t , that is, the series formed with the prices of that stock taken at fixed points in time, say daily. Let's now write a simple econometric model of the prices of a stock as follows:

$$p_{t+1} = \mu + \rho p_t + \varepsilon_{t+1} \quad (1)$$

This model tells us that if we consider any time $t+1$, the price of that stock at time $t+1$ is equal to a constant plus the price in the previous moment t multiplied by ρ plus a zero-mean random disturbance independent from the past, which always has the same statistical characteristics. If we want to apply this model to real-world price processes, the constants μ and ρ must be estimated. The parameter μ determines the trend and ρ defines the dependence between the prices. Typically ρ is less than but close to 1. A random disturbance of the type shown in the above equation is called a *white noise*.

If we know the initial price p_0 at time $t = 0$, using a computer program to generate random numbers, we can simulate a path of the price process with the following recursive equations:

$$\begin{aligned} p_1 &= \mu + \rho p_0 + \varepsilon_1 \\ p_2 &= \mu + \rho p_1 + \varepsilon_2 \end{aligned}$$

That is, we can compute the price at time $t = 1$ from the initial price p_0 and a computer-generated random number ε_1 and then use this new price to compute the price at time $t = 2$, and so on. The ε_i are independent and identically distributed random variables with zero mean. Typical choices for the distribution of ε are normal distribution, t -distribution, and stable non-Gaussian distribution. The distribution parameters are estimated from the sample.

It is clear that if we have a DGP we can generate any path. An econometric model that involves two or more different times can be regarded as a DGP. However, there is a more general way of looking at econometric models that encompasses both static and dynamic models. That is, we can look at econometric models from a perspective other than that of the

recursive generation of stochastic paths. In fact, we can rewrite our previous model as follows:

$$p_{t+1} - \mu - \rho p_t = \varepsilon_{t+1} \quad (2)$$

This formulation shows that, if we consider any two consecutive instants of time, there is a combination of prices that behave as random noise. More in general, an econometric model can be regarded as a mathematical device that reconstructs a noise sequence from empirical data.

This concept is visualized in Figure 1, which shows a time series of numbers p_t generated by a computer program according to the rule given by (2) with $\rho = 0.9$ and $\mu = 1$ and the corresponding time series ε_t . If we choose any pair of consecutive points in time, say $(t+1, t)$, the difference $p_{t+1} - \mu - \rho p_t$ is always equal to the series ε_{t+1} . For example, consider the points $p_{13} = 10.2918$, $p_{14} = 12.4065$. The difference $p_{14} - 0.9p_{13} - 1 = 2.1439$ has the same value as ε_{14} . If we move to a different pair we obtain the same result, that is, if we compute $p_{t+1} - 1 - 0.9p_t$, the result will always be the noise sequence ε_{t+1} .

To help intuition, imagine that our model is a test instrument: Probing our time series with our test instrument, we always obtain the same reading. Actually, what we obtain is not a constant reading but a random reading with mean zero and fixed statistical characteristics. The objective of financial econometrics is to find possibly simple expressions of different financial variables such as prices, returns, or financial ratios in different moments that always yield, as a result, a zero-mean random disturbance.

Static models (i.e., models that involve only one instant) are used to express relationships between different variables at any given time. Static models are used, for example, to determine exposure to different risk factors. However, because they involve only one instant, static models cannot be used to make forecasts; forecasting requires models that link variables in two or more instants in time.

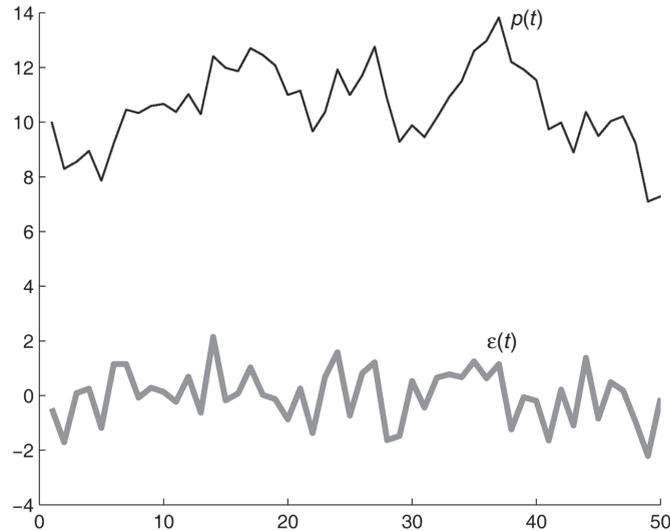


Figure 1 DGP and Noise Terms

FINANCIAL ECONOMETRICS AT WORK

Applying financial econometrics involves three key steps: (1) *model selection*, (2) *model estimation*, and (3) *model testing*.

In the first step, model selection, the modeler chooses (or might write *ex novo*) a family of models with given statistical properties. This entails the mathematical analysis of the model properties as well as economic theory to justify the model choice. It is in this step that the modeler decides to use, for example, regression on financial ratios or other variables to model returns.

In general, models include a number of free parameters that have to be estimated from sample data, the second step in applying financial econometrics. Suppose that we have decided to model returns with a regression model. This requires the estimation of the regression coefficients, performed using historical data. Estimation provides the link between reality and models. As econometric models are probabilistic models, any model can in principle describe our empirical data. We choose a family of models in the model selection phase and then de-

termine the optimal model in the estimation phase.

As mentioned, model selection and estimation are performed on historical data. As models are adapted (or fitted) to historical data there is always the risk that the fitting process captures ephemeral features of the data. Thus there is the need to test the models on data different from the data on which the models were estimated. This is the third step in applying financial econometrics, model testing. We assess the performance of models on fresh data.

We can take a different approach to model selection and estimation, namely statistical learning. *Statistical learning* combines the two steps—model selection and model estimation—insofar as it makes use of a class of universal models that can fit any data. Neural networks are an example of universal models. The critical step in the statistical learning approach is estimation. This calls for methods to restrict model complexity (i.e., the number of parameters used in a model).

Within this basic scheme for applying financial econometrics, we can now identify a number of modeling issues, such as:

- How do we apply statistics given that there is only one realization of financial series?
- Given a sample of historical data, how do we choose between linear and nonlinear models, or the different distributional assumptions or different levels of model complexity?
- Can we exploit more data using, for example, high-frequency data?
- How can we make our models more robust, reducing model risk?
- How do we measure not only model performance but also the ability to realize profits?

Implications of Empirical Series with Only One Realization

As mentioned, econometric models are *probabilistic models*: Variables are random variables characterized by a probability distribution. Generally speaking, probability concepts cannot be applied to single “individuals” (at least, not if we use a frequentist concept of probability). Probabilistic models describe “populations” formed by many individuals. However, empirical financial time series have only one realization. For example, there is only one historical series of prices for each stock—and we have only one price at each instant of time. This makes problematic the application of probability concepts. How, for example, can we meaningfully discuss the distribution of prices at a specific time given that there is only one price observation? Applying probability concepts to perform estimation and testing would require populations made up of multiple time series and samples made up of different time series that can be considered a random draw from some distribution.

As each financial time series is unique, the solution is to look at the single elements of the time series as the individuals of our population. For example, because there is only one realization of each stock’s price time series, we have to look at the price of each stock at different moments. However, the price of a stock (or of any other asset) at different moments is not a random in-

dependent sample. For example, it makes little sense to consider the distribution of the prices of a single stock in different moments because the level of prices typically changes over time. Our initial time series of financial quantities must be transformed; that is, a unique time series must be transformed into populations of individuals to which statistical methods can be applied. This holds not only for prices but for any other financial variable.

Econometrics includes transformations of the above type as well as tests to verify that the transformation has obtained the desired result. The DGP is the most important of these transformations. Recall that we can interpret a DGP as a method for transforming a time series into a sequence of noise terms. The DGP, as we have seen, constructs a sequence of random disturbances starting from the original series; it allows one to go backwards and infer the statistical properties of the series from the noise terms and the DGP. However, these properties cannot be tested independently.

The DGP is not the only transformation that allows statistical estimates. Differencing time series, for example, is a process that may transform *nonstationary time series* into *stationary time series*. A stationary time series has a constant mean that, under specific assumptions, can be estimated as an empirical average.

Determining the Model

As we have seen, econometric models are mathematical relationships between different variables at different times. An important question is whether these relationships are linear or nonlinear. Consider that every econometric model is an approximation. Thus the question is: Which approximation—linear or nonlinear—is better?

To answer this, it is generally necessary to consider jointly the linearity of models, the distributional assumptions, and the number of time lags to introduce. The simplest models are linear models with a small number of lags under

the assumption that variables are normal variables. A widely used example of normal linear models are regression models where returns are linearly regressed on lagged factors under the assumption that noise terms are normally distributed. A model of this type can be written as:

$$r_{t+1} = \beta f_t + \varepsilon_{t+1} \quad (3)$$

where r_t are the returns at time t and f_t are factors, that is, economic or financial variables. Given the linearity of the model, if factors and noise are jointly normally distributed, returns are also normally distributed.

However, the distribution of returns, at least at some time horizons, is not normal. If we postulate a nonlinear relationship between factors and returns, normally distributed factors yield a non-normal return distribution. However, we can maintain the linearity of the regression relationship but assume a non-normal distribution of noise terms and factors. Thus nonlinear models transform normally distributed noise into non-normal variables but it is not true that non-normal distributions of variables implies nonlinear models.

If we add lags (i.e., a time space backward), the above model becomes sensitive to the shape of the factor paths. For example, a regression model with two lags will behave differently if the factor is going up or down. Adding lags makes models more flexible but more brittle. In general, the optimal number of lags is dictated not only by the complexity of the patterns that we want to model but also by the number of points in our sample. If sample data are abundant, we can estimate a rich model.

Typically there is a trade-off between model flexibility and the size of the data sample. By adding time lags and nonlinearities, we make our models more flexible, but the demands in terms of estimation data are greater. An optimal compromise has to be made between the flexibility given by nonlinear models and/or multiple lags and the limitations due to the size of the data sample.

TIME HORIZON OF MODELS

There are trade-offs between model flexibility and precision that depend on the size of sample data. To expand our sample data, we would like to use data with small time spacing in order to multiply the number of available samples. *High-frequency data* or HFD (i.e., data on individual transactions) have the highest possible frequency (i.e., each individual transaction) and are irregularly spaced. To give an idea of the ratio in terms of numbers, consider that there are approximately 2,100 ticks per day for the median stock in the Russell 3000 (see Falkenberry, 2002). Thus the size of the HDF data set of one day for a typical stock in the Russell 3000 is 2,100 times larger than the size of closing data for the same day!

In order to exploit all available data, we would like to adopt models that work over time intervals of the order of minutes and, from these models, compute the behavior of financial quantities over longer periods. Given the number of available sample data at high frequency, we could write much more precise laws than those established using longer time intervals. Note that the need to compute solutions over forecasting horizons much longer than the time spacing is a general problem that applies at any time interval. For example, in asset allocation we need to understand the behavior of financial quantities over long time horizons. The question we need to ask is if models estimated using daily intervals can correctly capture the process dynamics over longer periods, such as years.

It is not necessarily true that models estimated on short time intervals, say minutes, offer better forecasts at longer time horizons than models estimated on longer time intervals, say days. This is because financial variables might have a complex short-term dynamics superimposed on a long-term dynamics. It might be that using high-frequency data, one captures the short-term dynamics without any improvement in the estimation of the long-term dynamics. That

is, with high-frequency data it might be that models get more complex (and thus more data-hungry) because they describe short-term behavior superimposed on long-term behavior. This possibility must be resolved for each class of models.

Another question is if it is possible to use the same model at different time horizons. To do so is to imply that the behavior of financial quantities is similar at different time horizons. This conjecture was first made by Benoit Mandelbrot (1963), who observed that long series of cotton prices were very similar at different time aggregations.

Model Risk and Model Robustness

Not only are econometric models probabilistic models, as we have already noted; they are only approximate models. That is, the probability distributions themselves are only approximate and uncertain. The theory of *model risk* and *model robustness* assumes that all parameters of a model are subject to uncertainty and attempts to determine the consequence of model uncertainty and strategies for mitigating errors.

The growing use of models in finance has heightened the attention to model risk and model-risk mitigation techniques. Asset management firms are beginning to address the need to implement methodologies that allow both robust estimation and robust optimization in the portfolio management process.

Performance Measurement of Models

It is not always easy to understand *ex ante* just how well (or how poorly) a forecasting model will perform. Because performance evaluations made on training data are not reliable, the evaluation of *model performance* requires separate data sets for training and for testing. Models are estimated on training data and tested on the test data. Poor performance might be due to

model misspecification, that is, models might not reflect the true DGP of the data (assuming one exists), or there might simply be no DGP.

Various measures of model performance have been proposed. For example, one can compute the correlation coefficient between the forecasted variables and their actual realizations. Each performance measure is a single number and therefore conveys only one aspect of the forecasting performance. Often it is crucial to understand if errors can become individually very large or if they might be correlated. Note that a simple measure of model performance does not ensure the profitability of strategies. This can be due to a number of reasons, including, for example, the risk inherent in apparently profitable forecasts, market impact, and transaction costs.

APPLICATIONS

There has been a greater use of econometric models in investment management since the turn of the century. Application areas include:

- Portfolio construction and optimization
- Risk management
- Asset and liability management

Each type of application requires different modeling approaches.

Portfolio Construction and Optimization

Portfolio construction and optimization require models to forecast returns: There is no way to escape the need to predict future returns. Passive strategies apparently eschew the need to forecast future returns of individual stocks by investing in broad indexes. They effectively shift the need to forecast to a higher level of analysis and to longer time horizons.

Until recently, the mainstream view was that financial econometric models could perform dynamic forecasts of volatility but not of

expected returns. However, volatility forecasts are rarely used in portfolio management. With the exception of some proprietary applications, the most sophisticated models used in portfolio construction until recently were factor models where forecasts are not dynamic but consist in estimating a drift (i.e., a constant trend) plus a variance-covariance matrix.

Since the late 1990s, the possibility of making dynamic forecasts of both volatility and expected returns has gained broad acceptance. During the same period, it became more widely recognized that returns are not normally distributed, evidence that had been reported by Mandelbrot (1963). Higher moments of distributions are therefore important in portfolio management and therefore require representation and estimation of nonnormal distributions.

As observed above, the ability to correctly forecast expected returns does not imply, per se, that there are profit opportunities. In fact, we have to take into consideration the interplay between expected returns, higher moments, and transaction costs. As dynamic forecasts typically involve higher portfolio turnover, transaction costs might wipe out profits. As a general comment, portfolio management based on dynamic forecasts calls for a more sophisticated framework for optimization and risk management with respect to portfolio management based on static forecasts.

Regression models appear to form the core of the modeling efforts to predict future returns at many asset management firms. Regression models regress returns on a number of predictors. Stated otherwise, future returns are a function of the value of present and past predictors. Predictors include financial ratios such as earning-to-price ratio or book-to-price ratio and other fundamental quantities; predictors might also include behavioral variables such as market sentiment. A typical formula of a regressive model is the following:

$$r_{i,t+1} = \alpha_i + \sum_{j=1} \beta_{ij} f_{j,t} + \varepsilon_{i,t+1} \quad (4)$$

where

$$r_{i,t+1} = \frac{P_{i,t+1} - P_{i,t}}{P_{i,t}}$$

is the return at time $t+1$ of the i -th asset, and the $f_{j,t}$ are factors observed at time t . While regressions are generally linear, nonlinear models are also used.

In general, the forecasting horizon in asset management varies from a few days for actively managed or hedge funds to several weeks for more traditionally managed funds. Dynamic models typically have a short forecasting horizon as they capture short-term dynamics. This contrasts with static models, such as the widely used multifactor models, which tend to capture long-term trends and ignore short-term dynamics.

The evolution of forecasting models over the last two decades has also changed the way forecasts are used. A basic utilization of forecasts is in stock picking/ranking systems, which have been widely implemented at asset management firms. The portfolio manager builds his or her portfolio combining the model ranking with the manager's personal views and within the constraints established by the firm. A drawback in using such an approach is the difficulty in properly considering the structure of correlations and the role of higher moments.

Alternatively, forecasts can be fed to an optimizer that automatically computes the portfolio weights. But because an optimizer implements an optimal trade-off between returns and some measure of risk, the forecasting model must produce not only returns forecasts but also measures of risk. If risk is measured by portfolio variance or standard deviation, the forecasting model must be able to provide an estimated variance-covariance matrix.

Estimating the variance-covariance matrix is the most delicate of the estimation tasks. Here is why. The number of entries of a variance-covariance matrix grows with the square of the number of stocks. As a consequence, the number of entries in a variance-covariance matrix

rapidly becomes very large. For example, the variance-covariance matrix of the stocks in the S&P 500 is a symmetric matrix that includes some 125,000 entries. If our universe were the Russell 5000, the variance-covariance matrix would include more than 12 million entries. The problem with estimating matrices of this size is that estimates are very noisy because the number of sample data is close to the number of parameters to estimate. For example, if we use three years of data for estimation, we have, on average, less than three data points per estimated entry in the case of the S&P 500; in the case of the Russell 5000, the number of data points would be one fourth of the number of entries to estimate! Robust estimation methods are called for.

Note that if we use forecasting models we typically have (1) an equilibrium variance-covariance matrix that represents the covariances of the long-run relationships between variables, plus (2) a short-term, time-dependent, variance-covariance matrix. If returns are not normally distributed, optimizers might require the matrix of higher moments.

A third utilization of forecasting models and optimizers is to construct model portfolios. In other words, the output of the optimizer is used to construct not an actual but a model portfolio. This model portfolio is used as input by portfolio managers.

Risk Management

Risk management has different meanings in different contexts. In particular, when optimization is used, risk management is intrinsic to the optimization process, itself a risk-return trade-off optimization. In this case, risk management is an integral part of the portfolio construction process.

However, in most cases, the process of constructing portfolios is entrusted to human portfolio managers who might use various inputs including, as noted above, ranking systems

or model portfolios. In these cases, portfolios might not be optimal from the point of view of risk management and it is therefore necessary to ensure independent risk oversight. This oversight might take various forms. One form is similar to the type of risk oversight adopted by banks. The objective is to assess potential deviations from expectations. In order to perform this task, the risk manager receives as input the composition of portfolios and makes return projections using static forecasting models.

Another form of risk oversight, perhaps the most diffused in portfolio management, assesses portfolio exposures to specific risk factors. As portfolio management is often performed relative to a benchmark and risk is defined as underperformance relative to the benchmark, it is important to understand the sensitivity of portfolios to different risk factors. This type of risk oversight does not entail the forecasting of returns. The risk manager uses various statistical techniques to estimate how portfolios move in function of different risk factors. In most cases, linear regressions are used. A typical model will have the following form:

$$r_{i,t} = \alpha_i + \sum_{j=1}^s \beta_{ij} f_{j,t} + \varepsilon_{i,t} \quad (5)$$

where

$$r_{i,t} = \frac{P_{i,t} - P_{i,t-1}}{P_{i,t-1}}$$

is the return observed at time t of the i -th asset, and the $f_{j,t}$ are factors observed at time t . Note that this model is fundamentally different from a regressive model with time lags as given by (4).

Asset-Liability Management

Asset-liability management (ALM) is typical of those asset management applications that require the optimization of portfolio returns at some fixed time horizon plus a stream of consumption throughout the entire life of the portfolio. ALM is important for managing portfolios

of institutional investors such as pension funds or foundations. It is also important for wealth management, where the objective is to cover the investor's financial needs over an extended period.

ALM requires forecasting models able to capture the asset behavior at short-, medium-, and long-term time horizons. Models of the long-term behavior of assets exist but are clearly difficult to test. Important questions related to these long-term forecasting models include:

- Do asset prices periodically revert to one or many common trends in the long run?
- Can we assume that the common trends (if they exist) are deterministic trends such as exponentials or are common trends stochastic (i.e., random) processes?
- Can we recognize regime shifts over long periods of time?

KEY POINTS

- Financial econometrics employs the same basic principles for formulating quantitative laws that characterized the development of quantitative science.
- Although there are similarities between financial econometric models and models of the physical sciences, important differences exist. For example, physical sciences seek immutable laws of nature, while econometric models model the economy or financial markets, which are artifacts subject to change.
- Econometric models are mathematical relationships between different variables at different times, and every econometric model is an approximation.
- Both continuous-time and discrete-time models are used in financial econometrics.
- Static models express relationships between different variables at any given time. Because

they involve only one instant in time, static models cannot be used to make forecasts since to do so models that link variables in two or more instants in time are required.

- Dynamic models involve one or more variables at two or more points in time; the data generating process in dynamic models is the mathematical relationship between variables at different times.
- Applying financial econometrics involves three key steps: (1) model selection, (2) model estimation, and (3) model testing.
- In model selection, the modeler selects the model based on an assessment of a model's properties and its fit to economic theory.
- Estimation provides the link between reality and models. In model estimation, the modeler applies financial econometric techniques to estimate the model's free parameters from sample data.
- The evaluation of model performance requires separate data sets for training and for testing because performance evaluations made on training data are not reliable.
- In investment management there has been increased use of econometric models in portfolio construction and optimization, risk management, and asset and liability management. A different modeling approach is needed for each type of application.

REFERENCES

- Falkenberry, T. N. (2002). High frequency data filtering. Tick Data Inc.
- Fan, J. (2004). An introduction to financial econometrics. Unpublished paper, Department of Operations Research and Financial Engineering, Princeton University.
- Focardi, S. M., and Fabozzi, F. J. (2004). *The Mathematics of Financial Modeling and Investment Management*. Hoboken, NJ: John Wiley & Sons.
- Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business* 36: 394–419.

Regression Analysis: Theory and Estimation

SERGIO M. FOCARDI, PhD

Partner, The Intertek Group

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: The tools of financial econometrics play an important role in financial model building. The most basic tool in financial econometrics is regression analysis. The purpose in regression analysis is to estimate the relationship between a random variable and one or more independent variables. To understand and apply regression analysis one must understand the theory and the methodologies for estimating the parameters of the regression model. Moreover, when the assumptions underlying the model are violated, it is necessary to know how to remedy the problem.

Our first basic tool in econometrics is regression analysis. In regression analysis, we estimate the relationship between a random variable Y and one or more variables X_i . The variables X_i can be either deterministic variables or random variables. The variable Y is said to be the dependent variable because its value is assumed to be dependent on the value of the X_i 's. The X_i 's are referred to as the independent variables, regressor variables, or explanatory variables. Our primary focus is on the *linear regression* model. We will be more precise about what we mean by a "linear" regression model later in this entry. Let's begin with a discussion of the concept of dependence.

THE CONCEPT OF DEPENDENCE

Regressions are about *dependence* between variables. In this section we provide a brief discussion of how dependence is represented in both a deterministic setting and a probabilistic setting. In a deterministic setting, the concept of dependence is embodied in the mathematical notion of function. A function is a correspondence between the individuals of a given domain A and the individuals of a given range B . In particular, numerical functions establish a correspondence between numbers in a domain A and numbers in a range B .

In quantitative science, we work with variables obtained through a process of observation or measurement. For example, price is the observation of a transaction, time is the reading of a clock, position is determined with measurements of the coordinates, and so on. In quantitative science, we are interested in numerical functions $y = f(x_1, \dots, x_n)$ that link the results of measurements so that by measuring the independent variables (x_1, \dots, x_n) we can predict the value of the dependent variable y . Being the results of measurements, variables are themselves functions that link a set Ω of unobserved “states of the world” to observations. Different states of the world result in different values for the variables but the link among the variables remains constant. For example, a column of mercury in a thermometer is a physical object that can be in different “states.” If we measure the length and the temperature of the column (in steady conditions), we observe that the two measurements are linked by a well-defined (approximately linear) function. Thus, by measuring the length, we can predict the temperature.

In order to model uncertainty, we keep the logical structure of variables as real-valued functions defined on a set Ω of unknown states of the world. However, we add to the set Ω the structure of a probability space. A probability space is a triple formed by a set of individuals (the states of the world), a structure of events, and a probability function: $(\Omega, \mathfrak{S}, P)$. Random variables represent measurements as in the deterministic case, but with the addition of a probability structure that represents uncertainty. In financial econometrics, a “state of the world” should be intended as a complete history of the underlying economy, not as an instantaneous state.

Our objective is to represent dependence between random variables, as we did in the deterministic case, so that we can infer the value of one variable from the measurement of the other. In particular, we want to infer the future values of variables from present and past observations. The probabilistic structure offers different possibilities. For simplicity, let’s consider only two

variables X and Y ; our reasoning extends immediately to multiple variables. The first case of interest is the case when the dependent variable Y is a random variable while the independent variable X is deterministic. This situation is typical of an experimental setting where we can fix the conditions of the experiment while the outcome of the experiment is uncertain.

In this case, the dependent variable Y has to be thought of as a family of random variables Y_x , all defined on the same probability space $(\Omega, \mathfrak{S}, P)$, indexed with the independent variable x . Dependence means that the probability distribution of the dependent random variable depends on the value of the deterministic independent value. To represent this dependence we use the notation $F(y|x)$ to emphasize the fact that x enters as a parameter in the distribution. An obvious example is the dependence of a price random variable on a time variable in a stochastic price process.

In this setting, where the independent variable is deterministic, the distributions $F(y|x)$ can be arbitrarily defined. Important for the discussion of linear regressions in this entry is the case when the shape of the distribution $F(y|x)$ remains constant and only the mean of the distribution changes as a function of x .

Consider now the case where both X and Y are random variables. For example, Y might be the uncertain price of IBM stock tomorrow and X the uncertain level of the S&P 500 tomorrow. One way to express the link between these two variables is through their joint distribution $F(x,y)$ and, if it exists, their joint density $f(x,y)$. We define the joint and marginal distributions as follows:

$$\begin{aligned}
 F_{XY}(x, y) &= P(X \leq x, Y \leq y), \quad F_X(x) = P(X \leq x), \\
 & \quad F_Y(y) = P(Y \leq y) \\
 F_{XY}(x, y) &= \int_{-\infty}^x \int_{-\infty}^y f(x, y) dx dy \\
 F_X(x) &= \int_{-\infty}^x \int_{-\infty}^{-\infty} f(u, y) du dy = \int_{-\infty}^x \left(\int_{-\infty}^{-\infty} f(u, y) dy \right) du \\
 &= \int_{-\infty}^x f_X(u) du
 \end{aligned}$$

$$\begin{aligned}
 F_Y(x) &= \int_{-\infty}^{-\infty} \int_{-\infty}^y f(x, v) dx dv = \int_{-\infty}^y \left(\int_{-\infty}^{-\infty} f(x, v) dx \right) dv \\
 &= \int_{-\infty}^x f_Y(v) dv \\
 f(x|y) &= \frac{f(x, y)}{f_Y(y)}, f(y|x) = \frac{f(x, y)}{f_X(x)}
 \end{aligned}$$

We will also use the short notation:

$$\begin{aligned}
 f_X(x) &= f(x), f_Y(y) = f(y), f_{X|Y}(x|y) \\
 &= f(x|y), f_{Y|X}(y|x) = f(y|x)
 \end{aligned}$$

Given a joint density $f(x, y)$, we can also represent the functional link between the two variables as the dependence of the distribution of one variable on the value assumed by the other variable. In fact, we can write the joint density $f(x, y)$ as the product of two factors, the conditional density $f(y|x)$ and the marginal density $f_X(x)$:

$$f(x, y) = f(y|x)f_X(x) \quad (1)$$

This *factorization*—that is, expressing a joint density as a product of a marginal density and a conditional density—is the conceptual basis of financial econometrics. There are significant differences in cases where both variables X and Y are random variables, compared to the case where the variable X is deterministic. First, as both variables are uncertain, we cannot fix the value of one variable as if it were independent. We have to adopt a framework of conditioning where our knowledge of one variable influences our knowledge of the other variable.

The impossibility of making experiments is a major issue in econometrics. In the physical sciences, the ability to create the desired experimental setting allows the scientist to isolate the effects of single variables. The experimenter tries to create an environment where the effects of variables other than those under study are minimized. In economics, however, all the variables change together and cannot be controlled. Back in the 1950s, there were serious doubts that econometrics was possible. In fact, it was believed that estimation required the independence of samples while economic samples are never independent.

However, the framework of conditioning addresses this problem. After conditioning, the joint densities of a process are factorized into initial and conditional densities that behave as independent distributions. An econometric model is a probe that extracts independent samples—the noise terms—from highly dependent variables.

Let's briefly see, at the heuristic level, how conditioning works. Suppose we learn that the random variable X has the value x , that is, $X = x$. Recall that X is a random variable that is a real-valued function defined over the set Ω . If we know that $X = x$, we do not know the present state of the world but we do know that it must be in the subspace ($\omega \in \Omega: X(\omega) = x$). We call $(Y|X = x)$ the variable Y defined on this subspace. If we let x vary, we create a family of random variables defined on the family of subspaces ($\omega \in \Omega: X(\omega) = x$) and indexed by the value assumed by the variable X .

It can be demonstrated that the sets ($\omega \in \Omega: X(\omega) = x$) can be given a structure of probability space, that the variables $(Y|X = x)$ are indeed random variables on these probability spaces, and that they have (if they exist) the conditional densities:

$$f(y|x) = \frac{f(x, y)}{f_X(x)} \quad (2)$$

for $f_X(x) > 0$. In the discrete setting we can write

$$\begin{aligned}
 f(y|x) &= P(Y = y|X = x) \\
 f(x, y) &= P(X = x, Y = y)
 \end{aligned}$$

The conditional expectation $E[Y|X = x]$ is the expectation of the variable $(Y|X = x)$. Consider the previous example of the IBM stock price tomorrow and of the S&P 500 level tomorrow. Both variables have unconditional expectations. These are the expectations of IBM's stock tomorrow and of S&P 500's level tomorrow considering every possible state of the world. However, we might be interested in computing the expected value of IBM's stock price tomorrow if we know S&P 500's value tomorrow. This is the case if, for example, we are creating scenarios based on S&P 500's value.

If we know the level of the S&P 500, we do not know the present state of the world but we do know the subset of states of the world in which the present state of the world is. If we only know the value of the S&P 500, IBM's stock price is not known because it is different in each state that belongs to this restricted set. IBM's stock price is a random variable on this restricted space and we can compute its expected value.

If we consider a discrete setting, that is, if we consider only a discrete set of possible IBM stock prices and S&P 500 values, then the computation of the conditional expectation can be performed using the standard definition of conditional probability. In particular, the conditional expectation of a random variable Y given the event B is equal to the unconditional expectation of the variable Y set to zero outside of B and divided by the probability of B : $E[Y|B] = E[1_B Y]/P(B)$, where 1_B is the indicator function of the set B , equal to 1 for all elements of B , zero elsewhere. Thus, in this example,

$$\begin{aligned} E[\text{IBM stock price} \mid \text{S\&P 500 value} = s] \\ = E[1_{(\text{S\&P 500 value}=s)}(\text{IBM stock price})] / \\ P(\text{S\&P 500 value} = s) \end{aligned}$$

However, in a continuous-state setting there is a fundamental difficulty: The set of states of the world corresponding to any given value of the S&P 500 has probability zero; therefore we cannot normalize dividing by $P(B)$. As a consequence we cannot use the standard definition of conditional probability to compute directly the conditional expectation.

To overcome this difficulty, we define the conditional expectation indirectly, using only unconditional expectations. We define the conditional expectation of IBM's stock price given the S&P 500 level as that variable that has the same unconditional expectation as IBM's stock price on each set that can be identified by for the value of the S&P 500. This is a random variable which is uniquely defined for each state of the world up to a set of probability zero.

If the conditional density exists, conditional expectation is computed as follows:

$$E[Y|X = x] = \int_{-\infty}^{+\infty} yf(y|x)dy \quad (3)$$

We know from probability theory that the *law of iterated expectations* holds

$$E[E[Y|X = x]] = E[Y] \quad (4)$$

and that the following relationship also holds

$$E[XY] = E[XE[Y|X]] \quad (5)$$

Rigorously proving all these results requires a considerable body of mathematics and the rather difficult language and notation of σ -algebras. However, the key ideas should be sufficiently clear.

What is the bearing of the above on the discussion of regressions in this entry? Regressions have a twofold nature: They can be either (1) the representation of dependence in terms of conditional expectations and conditional distributions or (2) the representation of dependence of random variables on deterministic parameters. The above discussion clarifies the probabilistic meaning of both.

REGRESSIONS AND LINEAR MODELS

In this section we discuss regressions and, in particular, linear regressions.

Case Where All Regressors Are Random Variables

Let's start our discussion of regression with the case where all regressors are random variables. Given a set of random variables $\mathbf{X} = (Y, X_1, \dots, X_N)'$, with a joint probability density $f(y, x_1, \dots, x_N)$, consider the conditional expectation of Y given the other variables (X_1, \dots, X_N) :

$$Y = E[Y|X_1, \dots, X_N]$$

As we saw in the previous section, the conditional expectation is a random variable. We can therefore consider the residual:

$$\varepsilon = Y - E[Y|X_1, \dots, X_N]$$

The residual is another random variable defined over the set Ω . We can rewrite the above equation as a *regression equation*:

$$Y = E[Y|X_1, \dots, X_N] + \varepsilon \quad (6)$$

The deterministic function $y = \varphi(\mathbf{z})$ where

$$y = \varphi(\mathbf{z}) = E[Y|X_1 = z_1, \dots, X_N = z_N] \quad (7)$$

is called the *regression function*.

The following *properties of regression* equations hold.

Property 1. The conditional mean of the residual is zero: $E[\varepsilon|X_1, \dots, X_N] = 0$. In fact, taking conditional expectations on both sides of equation (7), we can write

$$E[Y|X_1, \dots, X_N] = E[E[Y|X_1, \dots, X_N] | X_1, \dots, X_N] + E[\varepsilon|X_1, \dots, X_N]$$

Because

$$E[E[Y|X_1, \dots, X_N]|X_1, \dots, X_N] = E[Y|X_1, \dots, X_N]$$

is a property that follows from the law of iterated expectations, we can conclude that $E[\varepsilon|X_1, \dots, X_N] = 0$.

Property 2. The unconditional mean of the residual is zero: $E[\varepsilon] = 0$. This property follows immediately from the multivariate formulation of the law of iterated expectations (4): $E[E[Y|X_1, \dots, X_N]] = E[Y]$. In fact, taking expectation of both sides of equation (7) we can write

$$E[Y] = E[E[Y|X_1, \dots, X_N]] + E[\varepsilon]$$

hence $E[\varepsilon] = 0$.

Property 3: The residuals are uncorrelated with the variables X_1, \dots, X_N : $E[\varepsilon X] = 0$. This follows from equation (6) by multiplying both sides of

equation (7) by X_1, \dots, X_N and taking expectations. Note however, that the residuals are not necessarily independent of the regressor X .

If the regression function is linear, we can write the following *linear regression equation*:

$$Y = a + \sum_{i=1}^N b_i X_i + \varepsilon \quad (8)$$

and the following linear regression function:

$$y = a + \sum_{i=1}^N b_i x_i \quad (9)$$

The rest of this entry deals with linear regressions. If the vector $\mathbf{Z} = (Y, X_1, \dots, X_N)'$ is jointly normally distributed, then the regression function is linear. To see this, partition \mathbf{z} , the vector of means $\boldsymbol{\mu}$, and the covariance matrix Σ conformably in the following way:

$$\mathbf{Z} = \begin{pmatrix} Y \\ \mathbf{X} \end{pmatrix}, \quad \mathbf{z} = \begin{pmatrix} y \\ \mathbf{x} \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_y \\ \boldsymbol{\mu}_x \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} \sigma_{yy} & \sigma_{xy} \\ \sigma_{yx} & \Sigma_{xx} \end{pmatrix}$$

where $\boldsymbol{\mu}$ is the vector of means and Σ is the covariance matrix. It can be demonstrated that the conditional density $(Y|\mathbf{X} = \mathbf{x})$ has the following expression:

$$(Y|\mathbf{X} = \mathbf{x}) \sim N(\alpha + \boldsymbol{\beta}'\mathbf{x}, \sigma^2) \quad (10)$$

where

$$\boldsymbol{\beta} = \Sigma_{xx}^{-1} \sigma_{xy}$$

$$\alpha = \mu_y - \boldsymbol{\beta}'\boldsymbol{\mu}_x \quad (11)$$

$$\sigma^2 = \sigma_{yy}^2 - \sigma_{yx} \Sigma_{xx}^{-1} \sigma_{xy}$$

The regression function can be written as follows:

$$y = \alpha + \boldsymbol{\beta}'\mathbf{x}, \text{ or explicitly: } y = \alpha + \sum_{i=1}^N \beta_i x_i \quad (12)$$

The normal distribution is not the only joint distribution that yields linear regressions. Spherical and elliptical distributions also yield linear regressions. Spherical distributions extend the multivariate normal distribution $N(0, \mathbf{I})$

(i.e., the joint distribution of independent normal variables). Spherical distributions are characterized by the property that their density is constant on a sphere, so that their joint density can be written as

$$f(x_1, \dots, x_N) = g(x_1^2 + \dots + x_N^2)$$

for some function g .

Spherical distributions have the property that their marginal distributions are uncorrelated but not independent, and can be viewed as multivariate normal random variables, with a random covariance matrix. An example of a spherical distribution used in financial econometrics is the multivariate t -distribution with m degrees of freedom, whose density has the following form:

$$f(x_1, \dots, x_N) = c \left[1 + \frac{1}{m}(x_1^2 + \dots + x_N^2) \right]^{-\frac{m+N}{2}}$$

The multivariate t -distribution is important in econometrics for several reasons. First, some sampling distributions are actually a t -distribution entries. Second, the t -distribution proved to be an adequate description of fat-tailed error terms in some econometrics models (although not as good as the stable Paretian distribution).

Elliptical distributions generalize the multivariate normal distribution $N(0, \Sigma)$. (See Bradley and Taqqu [2003].) Because they are constant on an ellipsoid, their joint density can be written as

$$f(\mathbf{x}) = g((\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma} (\mathbf{x} - \boldsymbol{\mu})), \mathbf{x}' = (x_1, \dots, x_N)$$

where $\boldsymbol{\mu}$ is a vector of constants and $\boldsymbol{\Sigma}$ is a strictly positive-definite matrix. Spherical distributions are a subset of elliptical distributions. Conditional distributions and linear combinations of elliptical distributions are also elliptical.

The fact that elliptical distributions yield linear regressions is closely related to the fact that the linear correlation coefficient is a meaningful measure of dependence only for elliptical distributions. There are distributions that do not factorize as linear regressions. The linear

correlation coefficient is not a meaningful measure of dependence for these distributions. The copula function of a given random vector $X = (X_1, \dots, X_N)'$ completely describes the dependence structure of the joint distribution of random variables $X_i, i = 1, \dots, N$. (See Embrechts, McNeil, and Straumann [2002].)

Linear Models and Linear Regressions

Let's now discuss the relationship between linear regressions and linear models. In applied work, we are given a set of multivariate data that we want to explain through a model of their dependence. Suppose we want to explain the data through a linear model of the type:

$$Y = \alpha + \sum_{i=1}^N \beta_i X_i + \varepsilon$$

We might know from theoretical reasoning that linear models are appropriate or we might want to try a linear approximation to nonlinear models. A linear model such as the above is not, per se, a linear regression unless we apply appropriate constraints. In fact, linear regressions must satisfy the three properties mentioned above. We call linear regressions linear models of the above type that satisfy the following set of assumptions such that

$$\alpha + \sum_{i=1}^N \beta_i X_i$$

is the conditional expectation of Y .

Assumption 1. The conditional mean of the residual is zero: $E[\varepsilon | X_1, \dots, X_N]$.

Assumption 2. The unconditional mean of the residual is zero: $E[\varepsilon] = 0$.

Assumption 3: The correlation between the residuals and the variables X_1, \dots, X_N is zero: $E[\varepsilon X] = 0$.

The above set of assumptions is not the full set of assumptions used when estimating a linear model as a regression but only consistency

conditions to interpret a linear model as a regression. We will introduce additional assumptions relative to how the model is sampled in the section on estimation. Note that the linear regression equation does not fully specify the joint conditional distribution of the dependent variables and the regressors. (This is a rather subtle point related to concept of exogeneity of variables. See Hendry [1995] for a further discussion.)

Case Where Regressors Are Deterministic Variables

In many applications of interest to the financial modeler, the regressors are deterministic variables. Conceptually, regressions with deterministic regressors are different from cases where regressors are random variables. In particular, as we have seen in a previous section, one cannot consider the regression as a conditional expectation. However, we can write a linear regression equation:

$$Y = \alpha + \sum_{i=1}^N \beta_i x_i + \varepsilon \quad (13)$$

and the following linear regression function:

$$y = \alpha + \sum_{i=1}^N \beta_i x_i \quad (14)$$

where the regressors are deterministic variables. As we will see in the following section, in both cases the least squares estimators are the same though the variances of the regression parameters as functions of the samples are different.

ESTIMATION OF LINEAR REGRESSIONS

In this section, we discuss how to estimate the linear regression parameters. We consider two main estimation techniques: maximum likelihood method and least squares method. A discussion of the sampling distributions of linear

regression parameters follow. The method of moments and the instrumental variables method are other methods that are used but are not discussed in this entry.

Maximum Likelihood Estimates

Let's reformulate the regression problem in a matrix form that is standard in regression analysis and that we will use in the following sections. Let's start with the case of a dependent variable Y and one independent regressor X . This case is referred to as the bivariate case or the simple linear regression. Suppose that we are empirically given T pairs of observations of the regressor and the independent variable. In financial econometrics these observations could represent, for example, the returns Y of a stock and the returns X of a factor taken at fixed intervals of time $t = 1, 2, \dots, T$. Using a notation that is standard in regression estimation, we place the given data in a vector \mathbf{Y} and a matrix \mathbf{X} :

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_T \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_T \end{pmatrix} \quad (15)$$

The column of 1s represents constant terms. The regression equation can be written as a set of T samples from the same regression equation, one for each moment:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ &\vdots \\ Y_T &= \beta_0 + \beta_1 X_T + \varepsilon_T \end{aligned}$$

that we can rewrite in matrix form,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients,

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

and $\boldsymbol{\varepsilon}$ are the residuals.

We now make a set of assumptions that are standard in regression analysis and that we will progressively relax. The assumptions for the linear regression model with normally

distributed residuals are:

1. The residuals are zero-mean, normally distributed independent variables $\varepsilon \sim N(0, \sigma_\varepsilon^2 \mathbf{I})$, where σ_ε^2 is the common variance of the residuals and \mathbf{I} is the identity matrix.
2. \mathbf{X} is distributed independently of the residuals ε .

(16)

The regression equation can then be written: $E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\beta$. The residuals form a sequence of independent variables. They can therefore be regarded as a strict white-noise sequence. As the residuals are independent draws from the same normal distribution, we can compute the log-likelihood function as follows:

$$\log L = -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(\sigma_\varepsilon^2) - \sum_{t=1}^T \left[\frac{(Y_t - \beta_0 - \beta_1 X_t)^2}{2\sigma_\varepsilon^2} \right] \quad (17)$$

The *maximum likelihood* (ML) principle requires maximization of the log-likelihood function. Maximizing the log-likelihood function entails first solving the equations:

$$\frac{\partial \log L}{\partial \beta_0} = 0, \quad \frac{\partial \log L}{\partial \beta_1} = 0, \quad \frac{\partial \log L}{\partial \sigma_\varepsilon^2} = 0$$

These equations can be explicitly written as follows:

$$\begin{aligned} \sum_{t=1}^T (Y_t - \beta_0 - \beta_1 X_t) &= 0 \\ \sum_{t=1}^T X_t (Y_t - \beta_0 - \beta_1 X_t) &= 0 \\ T\sigma_\varepsilon^2 - \sum_{t=1}^T [(Y_t - \beta_0 - \beta_1 X_t)^2] &= 0 \end{aligned}$$

A little algebra shows that solving the first two equations yields

$$\begin{aligned} \hat{\beta}_1 &= \frac{\overline{XY} - \overline{X}\overline{Y}}{\sigma_\varepsilon^2} \\ \hat{\beta}_0 &= (\overline{Y} - \beta_1 \overline{X}) \end{aligned} \quad (18)$$

where

$$\overline{X} = \frac{1}{T} \sum_{t=1}^T X_t, \quad \overline{XY} = \frac{1}{T} \sum_{t=1}^T X_t Y_t$$

and where $\bar{\sigma}_x, \bar{\sigma}_y$ are the empirical standard deviations of the sample variables X, Y respectively. Substituting these expressions in the third equation

$$\frac{\partial \log L}{\partial \sigma_\varepsilon^2} = 0$$

yields the variance of the residuals:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{T} \sum_{t=1}^T [(Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t)^2] \quad (19)$$

In the matrix notation established above, we can write the estimators as follows:

$$\text{For parameters: } \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (20)$$

For the variance of the regression:

$$\hat{\sigma}^2 = \frac{1}{T} (\mathbf{Y} - \mathbf{X}\hat{\beta})' (\mathbf{Y} - \mathbf{X}\hat{\beta}) \quad (21)$$

A comment is in order. We started with T pairs of given data $(X_i, Y_i), i = 1, \dots, T$ and then attempted to explain these data as a linear regression $Y = \beta_1 X + \beta_0 + \varepsilon$. We estimated the coefficients (β_1, β_2) with maximum likelihood estimation (MLE) methods. Given this estimate of the regression coefficients, the estimated variance of the residuals is given by equation (22). Note that equation (22) is the empirical variance of residuals computed using the estimated regression parameters. A large variance of the residuals indicates that the level of noise in the process (i.e., the size of the unexplained fluctuations of the process) is high.

Generalization to Multiple Independent Variables

The above discussion of the MLE method generalizes to multiple independent variables, N . We are empirically given a set of T observations

that we organize in matrix form,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_T \end{pmatrix}, \mathbf{X} = \begin{pmatrix} X_{11} & \dots & X_{N1} \\ \vdots & \ddots & \vdots \\ X_{1T} & \dots & X_{NT} \end{pmatrix} \quad (22)$$

and the regression coefficients and error terms in the vectors,

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_N \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{pmatrix} \quad (23)$$

The matrix \mathbf{X} which contains all the regressors is called the design matrix. The regressors \mathbf{X} can be deterministic, the important condition being that the residuals are independent. One of the columns can be formed by 1s to allow for a constant term (intercept). Our objective is to explain the data as a linear regression:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

We make the same set of assumptions given by equation (17) as we made in the case of a single regressor. Using the above notation, the loglikelihood function will have the form

$$\log L = -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \quad (24)$$

The maximum likelihood conditions are written as

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = 0, \quad \frac{\partial \log L}{\partial \sigma_\varepsilon^2} = 0 \quad (25)$$

These equations are called normal equations. Solving the system of normal equations gives the same form for the estimators as in the univariate case:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \\ \hat{\sigma}^2 = \frac{1}{T} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \quad (26)$$

The variance estimator is not unbiased. It can be demonstrated that to obtain an unbiased estimator we have to apply a correction that takes into account the number of variables by replac-

ing T with $T - N$, assuming $T > N$:

$$\hat{\sigma}^2 = \frac{1}{T - N} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \quad (27)$$

The MLE method requires that we know the functional form of the distribution. If the distribution is known but not normal, we can still apply the MLE method but the estimators will be different. We will not here discuss further MLE for nonnormal distributions.

Ordinary Least Squares Method

We now establish the relationship between the MLE principle and the *ordinary least squares* (OLS) method. OLS is a general method to approximate a relationship between two or more variables. We use the matrix notation defined above for MLE method; that is, we assume that observations are described by equation (23) while the regression coefficients and the residuals are described by equation (24).

If we use the OLS method, the assumptions of linear regressions can be weakened. In particular, we need not assume that the residuals are normally distributed but only assume that they are uncorrelated and have finite variance. The residuals can therefore be regarded as a white-noise sequence (and not a strict white-noise sequence as in the previous section). We summarize the linear regression assumptions as follows:

- Assumptions for the linear regression model:
1. The mean of the residuals is zero: $E(\boldsymbol{\varepsilon}) = 0$
 2. The residuals are mutually uncorrelated: $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}$, where σ^2 is the variance of the residuals and \mathbf{I} is the identity matrix.
 3. \mathbf{X} is distributed independently of the residuals $\boldsymbol{\varepsilon}$.
- (28)

In the general case of a multivariate regression, the OLS method requires minimization of the sum of the squared residuals. Consider the

vector of residuals:

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{bmatrix}$$

The sum of the squared residuals ($SSR = (\varepsilon_1^2 + \dots + \varepsilon_T^2)$) can be written as $SSR = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$. As $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$, we can also write

$$SSR = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

The OLS method requires that we minimize the SSR. To do so, we equate to zero the first derivatives of the SSR:

$$\frac{\partial (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$$

This is a system of N equations. Solving this system, we obtain the estimators:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

These estimators are the same estimators obtained with the MLE method; they have an optimality property. In fact, the Gauss-Markov theorem states that the above OLS estimators are the best linear unbiased estimators (BLUE). "Best" means that no other linear unbiased estimator has a lower variance. It should be noted explicitly that OLS and MLE are conceptually different methodologies: MLE seeks the optimal parameters of the distribution of the error terms, while OLS seeks to minimize the variance of error terms. The fact that the two estimators coincide was an important discovery.

SAMPLING DISTRIBUTIONS OF REGRESSIONS

Estimated regression parameters depend on the sample. They are random variables whose distribution is to be determined. As we will see in this section, the sampling distributions differ depending on whether the regressors are

assumed to be fixed deterministic variables or random variables.

Let's first assume that the regressors are fixed deterministic variables. Thus only the error terms and the dependent variable change from sample to sample. The $\hat{\boldsymbol{\beta}}$ are unbiased estimators and $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ therefore holds. It can also be demonstrated that the following expression for the variance of $\hat{\boldsymbol{\beta}}$ holds

$$E[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \quad (29)$$

where an estimate $\hat{\sigma}^2$ of σ^2 is given by (27).

Under the additional assumption that the residuals are normally distributed, it can be demonstrated that the regression coefficients are jointly normally distributed as follows:

$$\hat{\boldsymbol{\beta}} \sim N_N[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}] \quad (30)$$

These expressions are important because they allow us to compute confidence intervals for the regression parameters.

Let's now suppose that the regressors are random variables. Under the assumptions set forth in (29), it can be demonstrated that the variance of the estimators $\hat{\boldsymbol{\beta}}$ can be written as follows:

$$V(\hat{\boldsymbol{\beta}}) = E[(\mathbf{X}'\mathbf{X})^{-1}]V(\mathbf{X}'\boldsymbol{\varepsilon})E[(\mathbf{X}'\mathbf{X})^{-1}] \quad (31)$$

where the terms $E[(\mathbf{X}'\mathbf{X})^{-1}]$ and $V(\mathbf{X}'\boldsymbol{\varepsilon})$ are the empirical expectation of $(\mathbf{X}'\mathbf{X})^{-1}$ and the empirical variance of $(\mathbf{X}'\boldsymbol{\varepsilon})$, respectively.

The following terms are used to describe this estimator of the variance: *sandwich estimator*, *robust estimator*, and *White estimator*. The term *sandwich estimator* is due to the fact that the term $V(\mathbf{X}'\boldsymbol{\varepsilon})$ is sandwiched between the terms $E[(\mathbf{X}'\mathbf{X})^{-1}]$. These estimators are *robust* because they take into account not only the variability of the dependent variables but also that of the independent variables. Consider that if the regressors are a large sample, the sandwich and the classical estimators are close to each other.

DETERMINING THE EXPLANATORY POWER OF A REGRESSION

The above computations to estimate regression parameters were carried out under the assumption that the data were generated by a linear regression function with uncorrelated and normally distributed noise. In general, we do not know if this is indeed the case. Though we can always estimate a linear regression model on any data sample by applying the estimators discussed above, we must now ask the question: When is a linear regression applicable and how can one establish the goodness (i.e., explanatory power) of a linear regression?

Quite obviously, a linear regression model is applicable if the relationship between the variables is approximately linear. How can we check if this is indeed the case? What happens if we fit a linear model to variables that have non-linear relationships, or if distributions are not normal? A number of tests have been devised to help answer these questions.

Intuitively, a measure of the quality of approximation offered by a linear regression is given by the variance of the residuals. Squared residuals are used because a property of the estimated relationship is that the sum of the residuals is zero. If residuals are large, the regression model has little explanatory power. However, the size of the average residual in itself is meaningless as it has to be compared with the range of the variables. For example, if we regress stock prices over a broad-based stock index, other things being equal, the residuals will be numerically different if the price is in the range of dollars or in the range of hundreds of dollars.

Coefficient of Determination

A widely used measure of the quality and usefulness of a regression model is given by the *coefficient of determination* denoted by R^2 or R -squared. The idea behind R^2 is the following.

The dependent variable Y has a total variation given by the following expression:

$$\text{Total variation} = S_Y^2 = \frac{1}{T-1} \sum_{t=1}^T (Y_t - \bar{Y})^2 \quad (32)$$

where

$$\bar{Y} = \frac{1}{T-1} \sum_{t=1}^T Y_t$$

This total variation is the sum of the variation of the variable Y due to the variation of the regressors plus the variation of residuals $S_Y^2 = S_R^2 + S_\varepsilon^2$. We can therefore define the coefficient of determination:

$$\begin{aligned} \text{Coefficient of determination} = R^2 &= \frac{S_R^2}{S_Y^2} \\ 1 - R^2 &= \frac{S_\varepsilon^2}{S_Y^2} \end{aligned} \quad (33)$$

as the portion of the total fluctuation of the dependent variable, Y , explained by the regression relation. R^2 is a number between 0 and 1: $R^2 = 0$ means that the regression has no explanatory power, $R^2 = 1$ means that the regression has perfect explanatory power. The quantity R^2 is computed by software packages that perform linear regressions.

It can be demonstrated that the coefficient of determination R^2 is distributed as the well-known Student F distribution. This fact allows one to determine intervals of confidence around a measure of the significance of a regression.

Adjusted R^2

The quantity R^2 as a measure of the usefulness of a regression model suffers from the problem that a regression might fit data very well in-sample but have no explanatory power out-of-sample. This occurs if the number of regressors is too high. Therefore an adjusted R^2 is sometimes used. The adjusted R^2 is defined as R^2 corrected by a penalty function that takes into account the number p of regressors in the

model:

$$\text{Adjusted } R^2 = \frac{T-1}{T-N-1} \frac{S_R^2}{S_Y^2} \quad (34)$$

Relation of R^2 to Correlation Coefficient

The R^2 is the squared *correlation coefficient*. The correlation coefficient is a number between -1 and $+1$ that measures the strength of the dependence between two variables. If a linear relationship is assumed, the correlation coefficient has the usual product-moment expression:

$$r = \sqrt{\frac{\overline{XY} - \bar{X}\bar{Y}}{S_y X_x}} \quad (35)$$

USING REGRESSION ANALYSIS IN FINANCE

This section provides several illustrations of regression analysis in finance as well as the data for each illustration. However, in order to present the data, we limit our sample size.

Characteristic Line for Common Stocks

The *characteristic line* of a security is the regression of the excess returns of that security on the market excess returns:

$$r_i = \alpha_i + \beta_i r_M$$

where

r_i = the security excess return of a security over the risk-free rate

r_M = the market excess return of the market over the risk-free rate

We computed the characteristic lines of two common stocks, Oracle and General Motors (GM), and a randomly created portfolio consisting of 20 stocks equally weighted. We used the S&P 500 Index as a proxy for the market returns and the 90-day Treasury rate as a proxy for the risk-free rate. The return and excess return data are shown in Table 1. Note that there are 60 monthly observations used to estimate the characteristic line from December 2000 to November 2005. The 20 stocks comprising the portfolio are shown at the bottom of Table 1.

Table 1 Return and Excess Return Data for S&P 500, Oracle, GM, and Portfolio^a: 12/1/2000–11/1/2005

Date	S&P 500 Return	Risk-Free Rate	S&P – Risk Free Rate	Oracle Return	Oracle Excess Return	GM Return	GM Excess Return	Portfolio Return	Portfolio Excess Return
12/1/2000	0.03464	0.00473	0.02990	0.00206	-0.00267	0.05418	0.04945	0.01446	0.00973
1/1/2001	-0.09229	0.00413	-0.09642	-0.34753	-0.35165	-0.00708	-0.01120	-0.07324	-0.07736
2/1/2001	-0.06420	0.00393	-0.06813	-0.21158	-0.21550	-0.02757	-0.03149	-0.07029	-0.07421
3/1/2001	0.07681	0.00357	0.07325	0.07877	0.07521	0.05709	0.05352	0.11492	0.11135
4/1/2001	0.00509	0.00321	0.00188	-0.05322	-0.05643	0.03813	0.03492	0.01942	0.01621
5/1/2001	-0.02504	0.00302	-0.02805	0.24183	0.23881	0.13093	0.12791	-0.03050	-0.03351
6/1/2001	-0.01074	0.00288	-0.01362	-0.04842	-0.05130	-0.01166	-0.01453	-0.03901	-0.04189
7/1/2001	-0.06411	0.00288	-0.06698	-0.32467	-0.32754	-0.13915	-0.14203	-0.08264	-0.08552
8/1/2001	-0.08172	0.00274	-0.08447	0.03030	0.02756	-0.21644	-0.21918	-0.13019	-0.13293
9/1/2001	0.01810	0.00219	0.01591	0.07790	0.07571	-0.03683	-0.03902	0.05969	0.05749
10/1/2001	0.07518	0.00177	0.07341	0.03466	0.03289	0.20281	0.20104	0.11993	0.11816
11/1/2001	0.00757	0.00157	0.00601	-0.01568	-0.01725	-0.02213	-0.02370	0.02346	0.02190
12/1/2001	-0.01557	0.00148	-0.01706	0.24982	0.24834	0.05226	0.05078	0.05125	0.04976
1/1/2002	-0.02077	0.00144	-0.02221	-0.03708	-0.03852	0.03598	0.03454	0.02058	0.01914
2/1/2002	0.03674	0.00152	0.03522	-0.22984	-0.23136	0.14100	0.13948	0.02818	0.02667
3/1/2002	-0.06142	0.00168	-0.06309	-0.21563	-0.21730	0.06121	0.05953	-0.00517	-0.00684
4/1/2002	-0.00908	0.00161	-0.01069	-0.21116	-0.21276	-0.03118	-0.03279	-0.02664	-0.02825
5/1/2002	-0.07246	0.00155	-0.07401	0.19571	0.19416	-0.13998	-0.14153	-0.04080	-0.04235
6/1/2002	-0.07900	0.00149	-0.08050	0.05702	0.05553	-0.12909	-0.13058	-0.05655	-0.05804
7/1/2002	0.00488	0.00142	0.00346	-0.04196	-0.04337	0.02814	0.02673	-0.01411	-0.01553

Table 1 (Continued)

Date	S&P 500 Return	Risk-Free Rate	S&P – Risk Free Rate	Oracle Return	Oracle Excess Return	GM Return	GM Excess Return	Portfolio Return	Portfolio Excess Return
8/1/2002	-0.11002	0.00133	-0.11136	-0.18040	-0.18173	-0.18721	-0.18855	-0.09664	-0.09797
9/1/2002	0.08645	0.00133	0.08512	0.29644	0.29510	-0.14524	-0.14658	0.06920	0.06787
10/1/2002	0.05707	0.00130	0.05577	0.19235	0.19105	0.19398	0.19268	0.08947	0.08817
11/1/2002	-0.06033	0.00106	-0.06139	-0.11111	-0.11217	-0.07154	-0.07259	-0.04623	-0.04729
12/1/2002	-0.02741	0.00103	-0.02845	0.11389	0.11286	-0.01438	-0.01541	-0.00030	-0.00134
1/1/2003	-0.01700	0.00100	-0.01800	-0.00582	-0.00682	-0.07047	-0.07147	-0.03087	-0.03187
2/1/2003	0.00836	0.00098	0.00737	-0.09365	-0.09463	-0.00444	-0.00543	-0.00951	-0.01049
3/1/2003	0.08104	0.00094	0.08010	0.09594	0.09500	0.07228	0.07134	0.06932	0.06838
4/1/2003	0.05090	0.00095	0.04995	0.09512	0.09417	-0.01997	-0.02092	0.06898	0.06803
5/1/2003	0.01132	0.00090	0.01042	-0.07686	-0.07776	0.01896	0.01806	0.00567	0.00477
6/1/2003	0.01622	0.00077	0.01546	-0.00167	-0.00243	0.03972	0.03896	0.03096	0.03019
7/1/2003	0.01787	0.00079	0.01708	0.07006	0.06927	0.09805	0.09726	0.03756	0.03677
8/1/2003	-0.01194	0.00086	-0.01280	-0.12315	-0.12401	-0.00414	-0.00499	-0.03145	-0.03231
9/1/2003	0.05496	0.00084	0.05412	0.06400	0.06316	0.04251	0.04167	0.07166	0.07082
10/1/2003	0.00713	0.00083	0.00630	0.00418	0.00334	0.00258	0.00174	0.00832	0.00749
11/1/2003	0.05077	0.00085	0.04992	0.10067	0.09982	0.24825	0.24740	0.06934	0.06849
12/1/2003	0.01728	0.00083	0.01645	0.04762	0.04679	-0.06966	-0.07049	0.00012	-0.00070
1/1/2004	0.01221	0.00081	0.01140	-0.07143	-0.07224	-0.03140	-0.03221	0.01279	0.01198
2/1/2004	-0.01636	0.00083	-0.01718	-0.06760	-0.06842	-0.01808	-0.01890	-0.03456	-0.03538
3/1/2004	-0.01679	0.00083	-0.01762	-0.06250	-0.06333	0.00360	0.00277	-0.00890	-0.00972
4/1/2004	0.01208	0.00091	0.01118	0.01333	0.01243	-0.04281	-0.04372	0.02303	0.02212
5/1/2004	0.01799	0.00109	0.01690	0.04649	0.04540	0.02644	0.02535	-0.00927	-0.01036
6/1/2004	-0.03429	0.00133	-0.03562	-0.11903	-0.12036	-0.07405	-0.07538	-0.05173	-0.05307
7/1/2004	0.00229	0.00138	0.00090	-0.05138	-0.05276	-0.04242	-0.04380	-0.00826	-0.00965
8/1/2004	0.00936	0.00143	0.00793	0.13139	0.12996	0.02832	0.02689	0.01632	0.01488
9/1/2004	0.01401	0.00156	0.01246	0.12234	0.12078	-0.09251	-0.09407	0.00577	0.00421
10/1/2004	0.03859	0.00167	0.03693	0.00632	0.00465	0.00104	-0.00063	0.05326	0.05159
11/1/2004	0.03246	0.00189	0.03057	0.07692	0.07503	0.03809	0.03620	0.02507	0.02318
12/1/2004	-0.02529	0.00203	-0.02732	0.00364	0.00162	-0.08113	-0.08315	-0.03109	-0.03311
1/1/2005	0.01890	0.00218	0.01673	-0.05955	-0.06172	-0.03151	-0.03369	0.01225	0.01008
2/1/2005	-0.01912	0.00231	-0.02143	-0.03629	-0.03860	-0.17560	-0.17790	-0.01308	-0.01538
3/1/2005	-0.02011	0.00250	-0.02261	-0.07372	-0.07622	-0.09221	-0.09471	-0.03860	-0.04110
4/1/2005	0.02995	0.00254	0.02741	0.10727	0.10472	0.18178	0.17924	0.04730	0.04476
5/1/2005	-0.00014	0.00257	-0.00271	0.03125	0.02868	0.07834	0.07577	-0.02352	-0.02609
6/1/2005	0.03597	0.00261	0.03336	0.02803	0.02542	0.08294	0.08033	0.04905	0.04644
7/1/2005	-0.01122	0.00285	-0.01407	-0.04274	-0.04559	-0.07143	-0.07428	-0.02185	-0.02470
8/1/2005	0.00695	0.00305	0.00390	-0.04542	-0.04847	-0.10471	-0.10776	0.00880	0.00575
9/1/2005	-0.01774	0.00306	-0.02080	0.02258	0.01952	-0.10487	-0.10793	-0.04390	-0.04696
10/1/2005	0.03519	0.00333	0.03186	-0.00631	-0.00963	-0.20073	-0.20405	0.01649	0.01316
11/1/2005	0.01009	0.00346	0.00663	-0.00714	-0.01060	0.01050	0.00704	0.01812	0.01466

^aPortfolio includes the following 20 stocks: Honeywell, Alcoa, Campbell Soup, Boeing, General Dynamics, Oracle, Sun, General Motors, Procter & Gamble, Wal-Mart, Exxon, ITT, Unilever, Hilton, Martin Marietta, Coca-Cola, Northrop Grumman, Mercury Interact, Amazon, and United Technologies.

The estimated parameters for the two stocks and the portfolios are reported in Table 2. As can be seen from the table, the intercept term is not statistically significant; however, the slope, referred to as the beta of the characteristic line, is statistically significant. Typically for individual stocks, the R^2 ranges from 0.15 to 0.65. For Oracle and GM the R^2 is 0.23 and 0.26, respectively.

In contrast, for a randomly created portfolio, the R^2 is considerably higher. For our 20-stock portfolio, the R^2 is 0.79.

Note that some researchers estimate a stock's beta by using returns rather than excess returns. The regression estimated is referred to as the single-index market model. This model was first suggested by Markowitz as a proxy

Table 2 Characteristic Line of the Common Stock of General Motors, Oracle, and Portfolio: 12/1/2000–11/1/2005

Coefficient	Coefficient Estimate	Standard Error	<i>t</i> -statistic	<i>p</i> -value
GM				
α	-0.005	0.015	-0.348	0.729
β	1.406	0.339	4.142	0.000
R^2	0.228			
<i>p</i> -value	0.000			
Oracle				
α	-0.009	0.011	-0.812	0.420
β	1.157	0.257	4.501	0.000
R^2	0.259			
<i>p</i> -value	0.000			
Portfolio				
α	0.003	0.003	1.027	0.309
β	1.026	0.070	14.711	0.000
R^2	0.787			
<i>p</i> -value	0.000			

measure of the covariance of a stock with an index so that the full mean-variance analysis need not be performed. While the approach was mentioned by Markowitz (1959) in a footnote in his book, it was Sharpe (1963) who investigated this further. It turns out that the beta estimated using both the characteristic line and the single-index market model do not differ materially. For example, for our 20-stock portfolio, the betas differed only because of rounding off.

Empirical Duration of Common Stock

A commonly used measure of the interest-rate sensitivity of an asset's value is its duration. Duration is interpreted as the approximate percentage change in the value of an asset for a 100-basis-point change in interest. Duration can be estimated by using a valuation model or empirically by estimating from historical returns the sensitivity of the asset's value to changes in interest rates. When duration is measured in the latter way, it is called empirical duration. Since it is estimated using regression analysis, it is sometimes referred to as regression-based duration.

A simple linear regression for computing empirical duration using monthly historical data (see Reilly, Wright, and Johnson, 2007) is

$$y_{it} = \alpha_i + \beta_i x_t + e_{it}$$

where

y_{it} = the percentage change in the value of asset i for month t

x_t = the change in the Treasury yield for month t

The estimated β_i is the empirical duration for asset i .

We will apply this linear regression to monthly data from October 1989 to October 2003 shown in Table 3¹ for the following asset indexes:

- Electric Utility sector of the S&P 500
- Commercial Bank sector of the S&P 500
- Lehman U.S. Aggregate Bond Index (now the Barclays Capital U.S. Aggregate Bond Index)

The yield change (x_t) is measured by the Lehman Treasury Index. The regression results are shown in Table 4. We report the empirical duration (β_i), the t -statistic, the p -value, the R^2 , and the intercept term. Negative values are reported for the empirical duration. In practice, however, the duration is quoted as a positive value. For the Electric Utility sector and the Lehman U.S. Aggregate Bond Index, the empirical duration is statistically significant at any reasonable level of significance.

A multiple regression model to estimate the empirical duration that has been suggested is

$$y_{it} = \alpha_i + \beta_{1i} x_{1t} + \beta_{2i} x_{2t} + e_{it}$$

where y_{it} and x_{1t} are the same as for the simple linear regression and x_{2t} is the return on the S&P 500. The results for this model are also shown in Table 4.

The results of the multiple regression indicate that the returns for the Electric Utility sector are affected by both the change in Treasury rates and the return on the stock market as proxied

Table 3 Data for Empirical Duration Illustration

Month	Change in Lehman Bros Treasury Yield	S&P500 Return	Monthly Returns for		
			Electric Utility Sector	Commercial Bank Sector	Lehman U.S. Aggregate Bond Index
Oct-89	-0.46	-2.33	2.350	-11.043	2.4600
Nov-89	-0.10	2.08	2.236	-3.187	0.9500
Dec-89	0.12	2.36	3.794	-1.887	0.2700
Jan-90	0.43	-6.71	-4.641	-10.795	-1.1900
Feb-90	0.09	1.29	0.193	4.782	0.3200
Mar-90	0.20	2.63	-1.406	-4.419	0.0700
Apr-90	0.34	-2.47	-5.175	-4.265	-0.9200
May-90	-0.46	9.75	5.455	12.209	2.9600
Jun-90	-0.20	-0.70	0.966	-5.399	1.6100
Jul-90	-0.21	-0.32	1.351	-8.328	1.3800
Aug-90	0.37	-9.03	-7.644	-10.943	-1.3400
Sep-90	-0.06	-4.92	0.435	-15.039	0.8300
Oct-90	-0.23	-0.37	10.704	-10.666	1.2700
Nov-90	-0.28	6.44	2.006	18.892	2.1500
Dec-90	-0.23	2.74	1.643	6.620	1.5600
Jan-91	-0.13	4.42	-1.401	8.018	1.2400
Feb-91	0.01	7.16	4.468	12.568	0.8500
Mar-91	0.03	2.38	2.445	5.004	0.6900
Apr-91	-0.15	0.28	-0.140	7.226	1.0800
May-91	0.06	4.28	-0.609	7.501	0.5800
Jun-91	0.15	-4.57	-0.615	-7.865	-0.0500
Jul-91	-0.13	4.68	4.743	7.983	1.3900
Aug-91	-0.37	2.35	3.226	9.058	2.1600
Sep-91	-0.33	-1.64	4.736	-2.033	2.0300
Oct-91	-0.17	1.34	1.455	0.638	1.1100
Nov-91	-0.15	-4.04	2.960	-9.814	0.9200
Dec-91	-0.59	11.43	5.821	14.773	2.9700
Jan-92	0.42	-1.86	-5.515	2.843	-1.3600
Feb-92	0.10	1.28	-1.684	8.834	0.6506
Mar-92	0.27	-1.96	-0.296	-3.244	-0.5634
Apr-92	-0.10	2.91	3.058	4.273	0.7215
May-92	-0.23	0.54	2.405	2.483	1.8871
Jun-92	-0.26	-1.45	0.492	1.221	1.3760
Jul-92	-0.41	4.03	6.394	-0.540	2.0411
Aug-92	-0.13	-2.02	-1.746	-5.407	1.0122
Sep-92	-0.26	1.15	0.718	1.960	1.1864
Oct-92	0.49	0.36	-0.778	2.631	-1.3266
Nov-92	0.26	3.37	-0.025	7.539	0.0228
Dec-92	-0.24	1.31	3.247	5.010	1.5903
Jan-93	-0.36	0.73	3.096	4.203	1.9177
Feb-93	-0.29	1.35	6.000	3.406	1.7492
Mar-93	0.02	2.15	0.622	3.586	0.4183
Apr-93	-0.10	-2.45	-0.026	-5.441	0.6955
May-93	0.25	2.70	-0.607	-0.647	0.1268
Jun-93	-0.30	0.33	2.708	4.991	1.8121
Jul-93	0.05	-0.47	2.921	0.741	0.5655
Aug-93	-0.31	3.81	3.354	0.851	1.7539
Sep-93	0.00	-0.74	-1.099	3.790	0.2746
Oct-93	0.05	2.03	-1.499	-7.411	0.3732
Nov-93	0.26	-0.94	-5.091	-1.396	-0.8502
Dec-93	0.01	1.23	2.073	3.828	0.5420

(Continued)

Table 3 (Continued)

Month	Change in Lehman Bros Treasury Yield	S&P500 Return	Monthly Returns for		
			Electric Utility Sector	Commercial Bank Sector	Lehman U.S. Aggregate Bond Index
Jan-94	-0.17	3.35	-2.577	4.376	1.3502
Feb-94	0.55	-2.70	-5.683	-4.369	-1.7374
Mar-94	0.55	-4.35	-4.656	-3.031	-2.4657
Apr-94	0.37	1.30	0.890	3.970	-0.7985
May-94	0.18	1.63	-5.675	6.419	-0.0138
Jun-94	0.16	-2.47	-3.989	-2.662	-0.2213
Jul-94	-0.23	3.31	5.555	2.010	1.9868
Aug-94	0.12	4.07	0.851	3.783	0.1234
Sep-94	0.43	-2.41	-2.388	-7.625	-1.4717
Oct-94	0.18	2.29	1.753	1.235	-0.0896
Nov-94	0.37	-3.67	2.454	-7.595	-0.2217
Dec-94	0.11	1.46	0.209	-0.866	0.6915
Jan-95	-0.33	2.60	7.749	6.861	1.9791
Feb-95	-0.41	3.88	-0.750	6.814	2.3773
Mar-95	0.01	2.96	-2.556	-1.434	0.6131
Apr-95	-0.18	2.91	3.038	4.485	1.3974
May-95	-0.72	3.95	7.590	9.981	3.8697
Jun-95	-0.05	2.35	-0.707	0.258	0.7329
Jul-95	0.14	3.33	-0.395	4.129	-0.2231
Aug-95	-0.10	0.27	-0.632	5.731	1.2056
Sep-95	-0.05	4.19	6.987	5.491	0.9735
Oct-95	-0.21	-0.35	2.215	-1.906	1.3002
Nov-95	-0.23	4.40	-0.627	7.664	1.4982
Dec-95	-0.18	1.85	6.333	0.387	1.4040
Jan-96	-0.13	3.44	2.420	3.361	0.6633
Feb-96	0.49	0.96	-3.590	4.673	-1.7378
Mar-96	0.31	0.96	-1.697	2.346	-0.6954
Apr-96	0.25	1.47	-4.304	-1.292	-0.5621
May-96	0.18	2.58	1.864	2.529	-0.2025
Jun-96	-0.14	0.41	5.991	-0.859	1.3433
Jul-96	0.08	-4.45	-7.150	0.466	0.2736
Aug-96	0.15	2.12	1.154	4.880	-0.1675
Sep-96	-0.23	5.62	0.682	6.415	1.7414
Oct-96	-0.35	2.74	4.356	8.004	2.2162
Nov-96	-0.21	7.59	1.196	10.097	1.7129
Dec-96	0.30	-1.96	-0.323	-4.887	-0.9299
Jan-97	0.06	6.21	0.443	8.392	0.3058
Feb-97	0.11	0.81	0.235	5.151	0.2485
Mar-97	0.36	-4.16	-4.216	-7.291	-1.1083
Apr-97	-0.18	5.97	-2.698	5.477	1.4980
May-97	-0.07	6.14	4.240	3.067	0.9451
Jun-97	-0.11	4.46	3.795	4.834	1.1873
Jul-97	-0.43	7.94	2.627	12.946	2.6954
Aug-97	0.30	-5.56	-2.423	-6.205	-0.8521
Sep-97	-0.19	5.48	5.010	7.956	1.4752
Oct-97	-0.21	-3.34	1.244	-2.105	1.4506
Nov-97	0.06	4.63	8.323	3.580	0.4603
Dec-97	-0.11	1.72	7.902	3.991	1.0063
Jan-98	-0.25	1.11	-4.273	-4.404	1.2837
Feb-98	0.17	7.21	2.338	9.763	-0.0753

Table 3 (Continued)

Month	Change in Lehman Bros Treasury Yield	S&P500 Return	Monthly Returns for		
			Electric Utility Sector	Commercial Bank Sector	Lehman U.S. Aggregate Bond Index
Mar-98	0.05	5.12	7.850	7.205	0.3441
Apr-98	0.00	1.01	-3.234	2.135	0.5223
May-98	-0.08	-1.72	-0.442	-3.200	0.9481
Jun-98	-0.09	4.06	3.717	2.444	0.8483
Jul-98	0.03	-1.06	-4.566	0.918	0.2122
Aug-98	-0.46	-14.46	7.149	-24.907	1.6277
Sep-98	-0.53	6.41	5.613	2.718	2.3412
Oct-98	0.05	8.13	-2.061	9.999	-0.5276
Nov-98	0.17	6.06	1.631	5.981	0.5664
Dec-98	0.02	5.76	2.608	2.567	0.3007
Jan-99	-0.01	4.18	-6.072	-0.798	0.7143
Feb-99	0.55	-3.11	-5.263	0.524	-1.7460
Mar-99	-0.05	4.00	-2.183	1.370	0.5548
Apr-99	0.05	3.87	6.668	7.407	0.3170
May-99	0.31	-2.36	7.613	-6.782	-0.8763
Jun-99	0.11	5.55	-4.911	5.544	-0.3194
Jul-99	0.11	-3.12	-2.061	-7.351	-0.4248
Aug-99	0.10	-0.50	1.508	-4.507	-0.0508
Sep-99	-0.08	-2.74	-5.267	-6.093	1.1604
Oct-99	0.11	6.33	1.800	15.752	0.3689
Nov-99	0.16	2.03	-8.050	-7.634	-0.0069
Dec-99	0.24	5.89	-0.187	-9.158	-0.4822
Jan-00	0.19	-5.02	5.112	-2.293	-0.3272
Feb-00	-0.13	-1.89	-10.030	-12.114	1.2092
Mar-00	-0.20	9.78	1.671	18.770	1.3166
Apr-00	0.17	-3.01	14.456	-5.885	-0.2854
May-00	0.07	-2.05	2.985	11.064	-0.0459
Jun-00	-0.26	2.47	-5.594	-14.389	2.0803
Jul-00	-0.08	-1.56	6.937	6.953	0.9077
Aug-00	-0.17	6.21	13.842	12.309	1.4497
Sep-00	-0.03	-5.28	12.413	1.812	0.6286
Oct-00	-0.06	-0.42	-3.386	-1.380	0.6608
Nov-00	-0.31	-7.88	3.957	-3.582	1.6355
Dec-00	-0.33	0.49	4.607	12.182	1.8554
Jan-01	-0.22	3.55	-11.234	3.169	1.6346
Feb-01	-0.16	-9.12	6.747	-3.740	0.8713
Mar-01	-0.08	-6.33	1.769	0.017	0.5018
Apr-01	0.22	7.77	5.025	-1.538	-0.4151
May-01	0.00	0.67	0.205	5.934	0.6041
Jun-01	0.01	-2.43	-7.248	0.004	0.3773
Jul-01	-0.40	-0.98	-5.092	2.065	2.2357
Aug-01	-0.14	-6.26	-0.149	-3.940	1.1458
Sep-01	-0.41	-8.08	-10.275	-4.425	1.1647
Oct-01	-0.39	1.91	1.479	-7.773	2.0930
Nov-01	0.41	7.67	-0.833	7.946	-1.3789
Dec-01	0.21	0.88	3.328	3.483	-0.6357
Jan-02	0.00	-1.46	-3.673	1.407	0.8096
Feb-02	-0.08	-1.93	-2.214	-0.096	0.9690
Mar-02	0.56	3.76	10.623	7.374	-1.6632
Apr-02	-0.44	-6.06	1.652	2.035	1.9393

(Continued)

Table 3 (Continued)

Month	Change in Lehman Bros Treasury Yield	S&P500 Return	Monthly Returns for		
			Electric Utility Sector	Commercial Bank Sector	Lehman U.S. Aggregate Bond Index
May-02	-0.06	-0.74	-3.988	1.247	0.8495
Jun-02	-0.23	-7.12	-4.194	-3.767	0.8651
Jul-02	-0.50	-7.80	-10.827	-4.957	1.2062
Aug-02	-0.17	0.66	2.792	3.628	1.6882
Sep-02	-0.45	-10.87	-8.677	-10.142	1.6199
Oct-02	0.11	8.80	-2.802	5.143	-0.4559
Nov-02	0.34	5.89	1.620	0.827	-0.0264
Dec-02	-0.45	-5.88	5.434	-2.454	2.0654
Jan-03	0.11	-2.62	-3.395	-0.111	0.0855
Feb-03	-0.21	-1.50	-2.712	-1.514	1.3843
Mar-03	0.05	0.97	4.150	-3.296	-0.0773
Apr-03	-0.03	8.24	5.438	9.806	0.8254
May-03	-0.33	5.27	10.519	5.271	1.8645
Jun-03	0.08	1.28	1.470	1.988	-0.1986
Jul-03	0.66	1.76	-5.649	3.331	-3.3620
Aug-03	0.05	1.95	1.342	-1.218	0.6637
Sep-03	-0.46	-1.06	4.993	-0.567	2.6469
Oct-03	0.33	5.66	0.620	8.717	-0.9320
Nov-03	0.13	0.88	0.136	1.428	0.2391
Dec-03	-0.14	5.24	NA	NA	NA

by the S&P 500. For the Commercial Bank sector, the coefficient of the changes in Treasury rates is not statistically significant, however the coefficient of the return on the S&P 500 is statistically significant. The opposite is the case for the Lehman U.S. Aggregate Bond Index. It is interesting to note that the duration for the Lehman U.S. Aggregate Bond Index as reported by Lehman Brothers was about 4.55 in November 2003. The empirical duration is 4.1. While the sign of the coefficient that is an estimate of duration is negative (which means the price moves in the opposite direction to the change in interest rates), market participants talk in terms of the positive value of duration for a bond that has this characteristic.

Predicting the 10-Year Treasury Yield²

The U.S. Department of the Treasury issues two types of securities: zero-coupon securities and

coupon securities. Securities issued with one year or less to maturity are called Treasury bills; they are issued as zero-coupon instruments. Treasury securities with more than one year to maturity are issued as coupon-bearing securities. Treasury securities from more than one year up to 10 years of maturity are called Treasury notes; Treasury securities with a maturity in excess of 10 years are called Treasury bonds. The U.S. Treasury auctions securities of specified maturities on a regular calendar basis. The Treasury currently issues 30-year Treasury bonds but had stopped issuance of them from October 2001 to January 2006.

An important Treasury coupon bond is the 10-year Treasury note. In this illustration we will try to forecast this rate based on two independent variables suggested by economic theory. A well-known theory of interest rates is that the interest rate in any economy consists of two components. This relationship is known as Fisher's law. The first is the expected

Table 4 Estimation of Regression Parameters for Empirical Duration

	Electric Utility Sector	Commercial Bank Sector	Lehman U.S. Aggregate Bond Index
a. Simple Linear Regression			
Intercept			
α_i	0.6376	1.1925	0.5308
t -statistic	1.8251	2.3347	21.1592
p -value	0.0698	0.0207	0.0000
Change in the Treasury yield			
β_i	-4.5329	-2.5269	-4.1062
t -statistic	-3.4310	-1.3083	-43.2873
p -value	0.0008	0.1926	0.0000
R^2	0.0655	0.0101	0.9177
F -value	11.7717	1.7116	1873.8000
p -value	0.0007	0.1926	0.0000
b. Multiple Linear Regression			
Intercept			
α_i	0.3937	0.2199	0.5029
t -statistic	1.1365	0.5835	21.3885
p -value	0.2574	0.5604	0.0000
Change in the Treasury yield			
β_{1i}	-4.3780	-1.9096	-4.0885
t -statistic	-3.4143	-1.3686	-46.9711
p -value	0.0008	0.1730	0.0000
Return on the S&P 500			
β_{2i}	0.2664	1.0620	0.0304
t -statistic	3.4020	12.4631	5.7252
p -value	0.0008	0.0000	0.0000
R^2	0.1260	0.4871	0.9312
F -value	12.0430	79.3060	1130.5000
p -value	0.00001	0.00000	0.00000

rate of inflation. The second is the real rate of interest. We use regression analysis to produce a model to forecast the yield on the 10-year Treasury note (simply, the 10-year Treasury yield)—the dependent variable—and the expected rate of inflation (simply, expected inflation) and the real rate of interest (simply, real rate).

The 10-year Treasury yield is observable, but we need a proxy for the two independent variables (i.e., the expected rate of inflation and the real rate of interest at the time) as they are not observable at the time of the forecast. Keep in mind that since we are forecasting, we do not use as our independent variable information that is unavailable at the time of the forecast. Consequently, we need a proxy available at the time of the forecast.

The inflation rate is available from the U.S. Department of Commerce. However, we need a proxy for expected inflation. We can use some type of average of past inflation as a proxy. In our model, we use a 5-year moving average. There are more sophisticated methodologies for calculating expected inflation, but the 5-year moving average is sufficient for our illustration. For example, one can use an exponential smoothing of actual inflation, a methodology used by the OECD. For the real rate, we use the rate on 3-month certificates of deposit (CDs). Again, we use a 5-year moving average.

The monthly data for the three variables from November 1965 to December 2005 (482 observations) are provided in Table 5. The regression results are reported in Table 6. As can be seen, the coefficients of both independent variables

Table 5 Monthly Data for 10-Year Treasury Yield, Expected Inflation, and Real Rate: November 1965–December 2005

Date	10-Yr. Trea. Yield	Exp. Infl.	Real Rate	Date	10-Yr. Trea. Yield	Exp. Infl.	Real Rate	Date	10-Yr. Trea. Yield	Exp. Infl.	Real Rate
1965											
Nov	4.45	1.326	2.739								
Dec	4.62	1.330	2.757								
1966				1970				1974			
Jan	4.61	1.334	2.780	Jan	7.80	3.621	3.061	Jan	6.99	4.652	3.330
Feb	4.83	1.348	2.794	Feb	7.24	3.698	3.064	Feb	6.96	4.653	3.332
Mar	4.87	1.358	2.820	Mar	7.07	3.779	3.046	Mar	7.21	4.656	3.353
Apr	4.75	1.372	2.842	Apr	7.39	3.854	3.035	Apr	7.51	4.657	3.404
May	4.78	1.391	2.861	May	7.91	3.933	3.021	May	7.58	4.678	3.405
June	4.81	1.416	2.883	June	7.84	4.021	3.001	June	7.54	4.713	3.419
July	5.02	1.440	2.910	July	7.46	4.104	2.981	July	7.81	4.763	3.421
Aug	5.22	1.464	2.945	Aug	7.53	4.187	2.956	Aug	8.04	4.827	3.401
Sept	5.18	1.487	2.982	Sept	7.39	4.264	2.938	Sept	8.04	4.898	3.346
Oct	5.01	1.532	2.997	Oct	7.33	4.345	2.901	Oct	7.9	4.975	3.271
Nov	5.16	1.566	3.022	Nov	6.84	4.436	2.843	Nov	7.68	5.063	3.176
Dec	4.84	1.594	3.050	Dec	6.39	4.520	2.780	Dec	7.43	5.154	3.086
1967				1971				1975			
Jan	4.58	1.633	3.047	Jan	6.24	4.605	2.703	Jan	7.5	5.243	2.962
Feb	4.63	1.667	3.050	Feb	6.11	4.680	2.627	Feb	7.39	5.343	2.827
Mar	4.54	1.706	3.039	Mar	5.70	4.741	2.565	Mar	7.73	5.431	2.710
Apr	4.59	1.739	3.027	Apr	5.83	4.793	2.522	Apr	8.23	5.518	2.595
May	4.85	1.767	3.021	May	6.39	4.844	2.501	May	8.06	5.585	2.477
June	5.02	1.801	3.015	June	6.52	4.885	2.467	June	7.86	5.639	2.384
July	5.16	1.834	3.004	July	6.73	4.921	2.436	July	8.06	5.687	2.311
Aug	5.28	1.871	2.987	Aug	6.58	4.947	2.450	Aug	8.4	5.716	2.271
Sept	5.3	1.909	2.980	Sept	6.14	4.964	2.442	Sept	8.43	5.738	2.241
Oct	5.48	1.942	2.975	Oct	5.93	4.968	2.422	Oct	8.15	5.753	2.210
Nov	5.75	1.985	2.974	Nov	5.81	4.968	2.411	Nov	8.05	5.759	2.200
Dec	5.7	2.027	2.972	Dec	5.93	4.964	2.404	Dec	8	5.761	2.186
1968				1972				1976			
Jan	5.53	2.074	2.959	Jan	5.95	4.959	2.401	Jan	7.74	5.771	2.166
Feb	5.56	2.126	2.943	Feb	6.08	4.959	2.389	Feb	7.79	5.777	2.164
Mar	5.74	2.177	2.937	Mar	6.07	4.953	2.397	Mar	7.73	5.800	2.138
Apr	5.64	2.229	2.935	Apr	6.19	4.953	2.403	Apr	7.56	5.824	2.101
May	5.87	2.285	2.934	May	6.13	4.949	2.398	May	7.9	5.847	2.060
June	5.72	2.341	2.928	June	6.11	4.941	2.405	June	7.86	5.870	2.034
July	5.5	2.402	2.906	July	6.11	4.933	2.422	July	7.83	5.900	1.988
Aug	5.42	2.457	2.887	Aug	6.21	4.924	2.439	Aug	7.77	5.937	1.889
Sept	5.46	2.517	2.862	Sept	6.55	4.916	2.450	Sept	7.59	5.981	1.813
Oct	5.58	2.576	2.827	Oct	6.48	4.912	2.458	Oct	7.41	6.029	1.753
Nov	5.7	2.639	2.808	Nov	6.28	4.899	2.461	Nov	7.29	6.079	1.681
Dec	6.03	2.697	2.798	Dec	6.36	4.886	2.468	Dec	6.87	6.130	1.615
1969				1973				1977			
Jan	6.04	2.745	2.811	Jan	6.46	4.865	2.509	Jan	7.21	6.176	1.573
Feb	6.19	2.802	2.826	Feb	6.64	4.838	2.583	Feb	7.39	6.224	1.527
Mar	6.3	2.869	2.830	Mar	6.71	4.818	2.641	Mar	7.46	6.272	1.474
Apr	6.17	2.945	2.827	Apr	6.67	4.795	2.690	Apr	7.37	6.323	1.427
May	6.32	3.016	2.862	May	6.85	4.776	2.734	May	7.46	6.377	1.397
June	6.57	3.086	2.895	June	6.90	4.752	2.795	June	7.28	6.441	1.340
July	6.72	3.156	2.929	July	7.13	4.723	2.909	July	7.33	6.499	1.293
Aug	6.69	3.236	2.967	Aug	7.40	4.699	3.023	Aug	7.4	6.552	1.252
Sept	7.16	3.315	3.001	Sept	7.09	4.682	3.110	Sept	7.34	6.605	1.217
Oct	7.1	3.393	3.014	Oct	6.79	4.668	3.185	Oct	7.52	6.654	1.193
Nov	7.14	3.461	3.045	Nov	6.73	4.657	3.254	Nov	7.58	6.710	1.154
Dec	7.65	3.539	3.059	Dec	6.74	4.651	3.312	Dec	7.69	6.768	1.119

Table 5 (Continued)

Date	10-Yr. Trea. Yield	Exp. Infl.	Real Rate	Date	10-Yr. Trea. Yield	Exp. Infl.	Real Rate	Date	10-Yr. Trea. Yield	Exp. Infl.	Real Rate
1978				1982				1986			
Jan	7.96	6.832	1.068	Jan	14.59	9.285	2.497	Jan	9.19	6.154	5.284
Feb	8.03	6.890	0.995	Feb	14.43	9.334	2.612	Feb	8.7	6.043	5.249
Mar	8.04	6.942	0.923	Mar	13.86	9.375	2.741	Mar	7.78	5.946	5.225
Apr	8.15	7.003	0.854	Apr	13.87	9.417	2.860	Apr	7.3	5.858	5.143
May	8.35	7.063	0.784	May	13.62	9.456	2.958	May	7.71	5.763	5.055
June	8.46	7.124	0.716	June	14.3	9.487	3.095	June	7.8	5.673	4.965
July	8.64	7.191	0.598	July	13.95	9.510	3.183	July	7.3	5.554	4.878
Aug	8.41	7.263	0.482	Aug	13.06	9.524	3.259	Aug	7.17	5.428	4.789
Sept	8.42	7.331	0.397	Sept	12.34	9.519	3.321	Sept	7.45	5.301	4.719
Oct	8.64	7.400	0.365	Oct	10.91	9.517	3.363	Oct	7.43	5.186	4.671
Nov	8.81	7.463	0.322	Nov	10.55	9.502	3.427	Nov	7.25	5.078	4.680
Dec	9.01	7.525	0.284	Dec	10.54	9.469	3.492	Dec	7.11	4.982	4.655
1979				1983				1987			
Jan	9.1	7.582	0.254	Jan	10.46	9.439	3.553	Jan	7.08	4.887	4.607
Feb	9.1	7.645	0.224	Feb	10.72	9.411	3.604	Feb	7.25	4.793	4.558
Mar	9.12	7.706	0.174	Mar	10.51	9.381	3.670	Mar	7.25	4.710	4.493
Apr	9.18	7.758	0.108	Apr	10.4	9.340	3.730	Apr	8.02	4.627	4.445
May	9.25	7.797	0.047	May	10.38	9.288	3.806	May	8.61	4.551	4.404
June	8.91	7.821	-0.025	June	10.85	9.227	3.883	June	8.4	4.476	4.335
July	8.95	7.834	-0.075	July	11.38	9.161	3.981	July	8.45	4.413	4.296
Aug	9.03	7.837	-0.101	Aug	11.85	9.087	4.076	Aug	8.76	4.361	4.273
Sept	9.33	7.831	-0.085	Sept	11.65	9.012	4.152	Sept	9.42	4.330	4.269
Oct	10.3	7.823	0.011	Oct	11.54	8.932	4.204	Oct	9.52	4.302	4.259
Nov	10.65	7.818	0.079	Nov	11.69	8.862	4.243	Nov	8.86	4.285	4.243
Dec	10.39	7.818	0.154	Dec	11.83	8.800	4.276	Dec	8.99	4.279	4.218
1980				1984				1988			
Jan	10.8	7.825	0.261	Jan	11.67	8.741	4.324	Jan	8.67	4.274	4.180
Feb	12.41	7.828	0.418	Feb	11.84	8.670	4.386	Feb	8.21	4.271	4.149
Mar	12.75	7.849	0.615	Mar	12.32	8.598	4.459	Mar	8.37	4.268	4.104
Apr	11.47	7.879	0.701	Apr	12.63	8.529	4.530	Apr	8.72	4.270	4.075
May	10.18	7.926	0.716	May	13.41	8.460	4.620	May	9.09	4.280	4.036
June	9.78	7.989	0.702	June	13.56	8.393	4.713	June	8.92	4.301	3.985
July	10.25	8.044	0.695	July	13.36	8.319	4.793	July	9.06	4.322	3.931
Aug	11.1	8.109	0.716	Aug	12.72	8.241	4.862	Aug	9.26	4.345	3.879
Sept	11.51	8.184	0.740	Sept	12.52	8.164	4.915	Sept	8.98	4.365	3.844
Oct	11.75	8.269	0.795	Oct	12.16	8.081	4.908	Oct	8.8	4.381	3.810
Nov	12.68	8.356	0.895	Nov	11.57	7.984	4.919	Nov	8.96	4.385	3.797
Dec	12.84	8.446	1.004	Dec	12.5	7.877	4.928	Dec	9.11	4.384	3.787
1981				1985				1989			
Jan	12.57	8.520	1.132	Jan	11.38	7.753	4.955	Jan	9.09	4.377	3.786
Feb	13.19	8.594	1.242	Feb	11.51	7.632	4.950	Feb	9.17	4.374	3.792
Mar	13.12	8.649	1.336	Mar	11.86	7.501	4.900	Mar	9.36	4.367	3.791
Apr	13.68	8.700	1.477	Apr	11.43	7.359	4.954	Apr	9.18	4.356	3.784
May	14.1	8.751	1.619	May	10.85	7.215	5.063	May	8.86	4.344	3.758
June	13.47	8.802	1.755	June	10.16	7.062	5.183	June	8.28	4.331	3.723
July	14.28	8.877	1.897	July	10.31	6.925	5.293	July	8.02	4.320	3.679
Aug	14.94	8.956	2.037	Aug	10.33	6.798	5.346	Aug	8.11	4.306	3.644
Sept	15.32	9.039	2.155	Sept	10.37	6.664	5.383	Sept	8.19	4.287	3.623
Oct	15.15	9.110	2.256	Oct	10.24	6.528	5.399	Oct	8.01	4.273	3.614
Nov	13.39	9.175	2.305	Nov	9.78	6.399	5.360	Nov	7.87	4.266	3.609
Dec	13.72	9.232	2.392	Dec	9.26	6.269	5.326	Dec	7.84	4.258	3.611

(Continued)

Table 5 (Continued)

Date	10-Yr. Trea. Yield	Exp. Infl.	Real Rate	Date	10-Yr. Trea. Yield	Exp. Infl.	Real Rate	Date	10-Yr. Trea. Yield	Exp. Infl.
1990				1994				1998		
Jan	8.418	4.257	3.610	Jan	5.642	4.256	1.739	Jan	5.505	2.828
Feb	8.515	4.254	3.595	Feb	6.129	4.224	1.663	Feb	5.622	2.806
Mar	8.628	4.254	3.585	Mar	6.738	4.195	1.586	Mar	5.654	2.787
Apr	9.022	4.260	3.580	Apr	7.042	4.166	1.523	Apr	5.671	2.765
May	8.599	4.264	3.586	May	7.147	4.135	1.473	May	5.552	2.744
June	8.412	4.272	3.589	June	7.32	4.106	1.427	June	5.446	2.725
July	8.341	4.287	3.568	July	7.111	4.079	1.394	July	5.494	2.709
Aug	8.846	4.309	3.546	Aug	7.173	4.052	1.356	Aug	4.976	2.695
Sept	8.795	4.335	3.523	Sept	7.603	4.032	1.315	Sept	4.42	2.680
Oct	8.617	4.357	3.503	Oct	7.807	4.008	1.289	Oct	4.605	2.666
Nov	8.252	4.371	3.493	Nov	7.906	3.982	1.278	Nov	4.714	2.653
Dec	8.067	4.388	3.471	Dec	7.822	3.951	1.278	Dec	4.648	2.641
1991				1995				1999		
Jan	8.007	4.407	3.436	Jan	7.581	3.926	1.269	Jan	4.651	2.631
Feb	8.033	4.431	3.396	Feb	7.201	3.899	1.261	Feb	5.287	2.621
Mar	8.061	4.451	3.360	Mar	7.196	3.869	1.253	Mar	5.242	2.605
Apr	8.013	4.467	3.331	Apr	7.055	3.840	1.240	Apr	5.348	2.596
May	8.059	4.487	3.294	May	6.284	3.812	1.230	May	5.622	2.586
June	8.227	4.504	3.267	June	6.203	3.781	1.222	June	5.78	2.572
July	8.147	4.517	3.247	July	6.426	3.746	1.223	July	5.903	2.558
Aug	7.816	4.527	3.237	Aug	6.284	3.704	1.228	Aug	5.97	2.543
Sept	7.445	4.534	3.223	Sept	6.182	3.662	1.232	Sept	5.877	2.527
Oct	7.46	4.540	3.207	Oct	6.02	3.624	1.234	Oct	6.024	2.515
Nov	7.376	4.552	3.177	Nov	5.741	3.587	1.229	Nov	6.191	2.502
Dec	6.699	4.562	3.133	Dec	5.572	3.549	1.234	Dec	6.442	2.490
1992				1996				2000		
Jan	7.274	4.569	3.092	Jan	5.58	3.505	1.250	Jan	6.665	2.477
Feb	7.25	4.572	3.054	Feb	6.098	3.458	1.270	Feb	6.409	2.464
Mar	7.528	4.575	3.014	Mar	6.327	3.418	1.295	Mar	6.004	2.455
Apr	7.583	4.574	2.965	Apr	6.67	3.376	1.328	Apr	6.212	2.440
May	7.318	4.571	2.913	May	6.852	3.335	1.359	May	6.272	2.429
June	7.121	4.567	2.864	June	6.711	3.297	1.387	June	6.031	2.421
July	6.709	4.563	2.810	July	6.794	3.261	1.417	July	6.031	2.412
Aug	6.604	4.556	2.757	Aug	6.943	3.228	1.449	Aug	5.725	2.406
Sept	6.354	4.544	2.682	Sept	6.703	3.195	1.481	Sept	5.802	2.398
Oct	6.789	4.533	2.624	Oct	6.339	3.163	1.516	Oct	5.751	2.389
Nov	6.937	4.522	2.571	Nov	6.044	3.131	1.558	Nov	5.468	2.382
Dec	6.686	4.509	2.518	Dec	6.418	3.102	1.608	Dec	5.112	2.374
1993				1997				2001		
Jan	6.359	4.495	2.474	Jan	6.494	3.077	1.656	Jan	5.114	2.368
Feb	6.02	4.482	2.427	Feb	6.552	3.057	1.698	Feb	4.896	2.366
Mar	6.024	4.466	2.385	Mar	6.903	3.033	1.746	Mar	4.917	2.364
Apr	6.009	4.453	2.330	Apr	6.718	3.013	1.795	Apr	5.338	2.364
May	6.149	4.439	2.272	May	6.659	2.990	1.847	May	5.381	2.362
June	5.776	4.420	2.214	June	6.5	2.968	1.899	June	5.412	2.363
July	5.807	4.399	2.152	July	6.011	2.947	1.959	July	5.054	2.363
Aug	5.448	4.380	2.084	Aug	6.339	2.926	2.016	Aug	4.832	2.365
Sept	5.382	4.357	2.020	Sept	6.103	2.909	2.078	Sept	4.588	2.365
Oct	5.427	4.333	1.958	Oct	5.831	2.888	2.136	Oct	4.232	2.366
Nov	5.819	4.309	1.885	Nov	5.874	2.866	2.189	Nov	4.752	2.368
Dec	5.794	4.284	1.812	Dec	5.742	2.847	2.247	Dec	5.051	2.370

Table 5 (Continued)

Date	10-Yr. Trea. Yield	Exp. Infl.	Real Rate	Date	10-Yr. Trea. Yield	Exp. Infl.	Real Rate
2002				2004			
Jan	5.033	2.372	2.950	Jan	4.134	2.172	1.492
Feb	4.877	2.372	2.888	Feb	3.973	2.157	1.442
Mar	5.396	2.371	2.827	Mar	3.837	2.149	1.385
Apr	5.087	2.369	2.764	Apr	4.507	2.142	1.329
May	5.045	2.369	2.699	May	4.649	2.136	1.273
June	4.799	2.367	2.636	June	4.583	2.134	1.212
July	4.461	2.363	2.575	July	4.477	2.129	1.156
Aug	4.143	2.364	2.509	Aug	4.119	2.126	1.097
Sept	3.596	2.365	2.441	Sept	4.121	2.124	1.031
Oct	3.894	2.365	2.374	Oct	4.025	2.122	0.966
Nov	4.207	2.362	2.302	Nov	4.351	2.124	0.903
Dec	3.816	2.357	2.234	Dec	4.22	2.129	0.840
2003				2005			
Jan	3.964	2.351	2.168	Jan	4.13	2.131	0.783
Feb	3.692	2.343	2.104	Feb	4.379	2.133	0.727
Mar	3.798	2.334	2.038	Mar	4.483	2.132	0.676
Apr	3.838	2.323	1.976	Apr	4.2	2.131	0.622
May	3.372	2.312	1.913	May	3.983	2.127	0.567
June	3.515	2.300	1.850	June	3.915	2.120	0.520
July	4.408	2.288	1.786	July	4.278	2.114	0.476
Aug	4.466	2.267	1.731	Aug	4.016	2.107	0.436
Sept	3.939	2.248	1.681	Sept	4.326	2.098	0.399
Oct	4.295	2.233	1.629	Oct	4.553	2.089	0.366
Nov	4.334	2.213	1.581	Nov	4.486	2.081	0.336
Dec	4.248	2.191	1.537	Dec	4.393	2.075	0.311

Note:

Expected Infl. (%) = expected rate of inflation as proxied by the 5-year moving average of the actual inflation rate.
 Real Rate (%) = real rate of interest as proxied by the 5-year moving average of the interest rate on 3-month certificates of deposit.

Table 6 Results of Regression for Forecasting 10-Year Treasury Yield

Regression Statistics					
Multiple R^2	0.908318				
R^2	0.825042				
Adjusted R^2	0.824312				
Standard Error	1.033764				
Observations	482				
Analysis of Variance					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	2413.914	1206.957	1129.404	4.8E-182
Residual	479	511.8918	1.068668		
Total	481	2925.806			
	Coefficients	Standard Error	<i>t</i>	Statistics <i>p</i> -value	
Intercept	1.89674	0.147593	12.85118	1.1E-32	
Expected Inflation	0.996937	0.021558	46.24522	9.1E-179	
Real Rate	0.352416	0.039058	9.022903	4.45E-18	

are positive (as would be predicted by economic theory) and highly significant.

NONNORMALITY AND AUTOCORRELATION OF THE RESIDUALS

In the above discussion we assumed that there is no correlation between the residual terms. Let's now relax these assumptions. The correlation of the residuals is critical from the point of view of estimation. *Autocorrelation* of residuals is quite common in financial estimation where we regress quantities that are time series.

A time series is said to be autocorrelated if each term is correlated with its predecessor so that the variance of each term is partially explained by regressing each term on its predecessor.

Recall from the previous section that we organized regressor data in a matrix called the design matrix. Suppose that both regressors and the variable Y are time series data, that is, every row of the design matrix corresponds to a moment in time. The regression equation is written as follows:

$$Y = X\beta + \varepsilon$$

Suppose that residuals are correlated. This means that in general $E[\varepsilon_i \varepsilon_j] = \sigma_{ij} \neq 0$. Thus the variance-covariance matrix of the residuals $\{\sigma_{ij}\}$ will not be a diagonal matrix as in the case of uncorrelated residuals, but will exhibit nonzero off-diagonal terms. We assume that we can write

$$\{\sigma_{ij}\} = \sigma^2 \Omega$$

where Ω is a positive definite symmetric matrix and σ is a parameter to be estimated.

If residuals are correlated, the regression parameters can still be estimated without biases using the formula given by (26). However, this estimate will not be optimal in the sense that there are other estimators with lower variance of the sampling distribution. An optimal linear

unbiased estimator has been derived. It is called Aitken's *generalized least squares* (GLS) estimator and is given by

$$\hat{\beta} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} Y \quad (36)$$

where Ω is the residual correlation matrix.

The GLS estimators vary with the sampling distribution. It can also be demonstrated that the variance of the GLS estimator is also given by the following "sandwich" formula:

$$V(\hat{\beta}) = E((\beta - \hat{\beta})(\beta - \hat{\beta})') = \sigma^2 (X' \Omega^{-1} X)^{-1} \quad (37)$$

This expression is similar to equation (28) with the exception of the sandwiched term Ω^{-1} . Unfortunately, (37) cannot be estimated without first knowing the regression coefficients. For this reason, in the presence of correlation of residuals, it is common practice to replace static regression models with models that explicitly capture autocorrelations and produce uncorrelated residuals.

The key idea here is that autocorrelated residuals signal that the modeling exercise has not been completed. If residuals are autocorrelated, this signifies that the residuals at a generic time t can be predicted from residuals at an earlier time. For example, suppose that we are linearly regressing a time series of returns r_t on N factors:

$$r_t = \alpha_1 f_{1,t-1} + \cdots + \alpha_N f_{N,t-1} + \varepsilon_t$$

Suppose that the residual terms ε_t are autocorrelated and that we can write regressions of the type

$$\varepsilon_t = \varphi \varepsilon_{t-1} + \eta_t$$

where η_t are now uncorrelated variables. If we ignore this autocorrelation, valuable forecasting information is lost. Our initial model has to be replaced with the following model:

$$r_t = \alpha_1 f_{1,t-1} + \cdots + \alpha_N f_{N,t-1} + \varepsilon_t$$

$$\varepsilon_t = \varphi \varepsilon_{t-1} + \eta_t$$

with the initial conditions ε_0 .

Detecting Autocorrelation

How do we detect the autocorrelation of residuals? Suppose that we believe that there is a reasonable linear relationship between two variables, for instance stock returns and some fundamental variable. We then perform a linear regression between the two variables and estimate regression parameters using the OLS method. After estimating the regression parameters, we can compute the sequence of residuals. At this point, we can apply tests such as the Durbin-Watson test or the Dickey-Fuller test to gauge the autocorrelation of residuals. If residuals are auto-correlated, we should modify the model.

PITFALLS OF REGRESSIONS

It is important to understand when regressions are correctly applicable and when they are not. In addition to the autocorrelation of residuals, there are other situations where it would be inappropriate to use regressions. In particular, we analyze the following cases, which represent possible pitfalls of regressions:

- Spurious regressions with integrated variables
- Collinearity
- Increasing the number of regressors

Spurious Regressions

The phenomenon of spurious regressions, observed by Yule in 1927, led to the study of cointegration. We encounter spurious regressions when we perform an apparently meaningful regression between variables that are independent. A typical case is a regression between two independent random walks. Regressing two independent random walks, one might find very high values of R^2 even if the two processes are independent. More in general, one might find high values of R^2 in the regression of two or more integrated variables, even if residuals are highly correlated.

Testing for regressions implies testing for cointegration. Anticipating what will be discussed there, it is always meaningful to perform regressions between stationary variables. When variables are integrated, regressions are possible only if variables are cointegrated. This means that residuals are a stationary (though possibly autocorrelated) process. As a rule of thumb, Granger and Newbold (1974) observe that if the R^2 is greater than the Durbin-Watson statistics, it is appropriate to investigate if correlations are spurious.

Collinearity

Collinearity, also referred to as multicollinearity, occurs when two or more regressors have a linear deterministic relationship. For example, there is collinearity if the design matrix

$$\mathbf{X} = \begin{pmatrix} X_{11} & \cdots & X_{N1} \\ \vdots & \ddots & \vdots \\ X_{1T} & \cdots & X_{NT} \end{pmatrix}$$

exhibits two or more columns that are perfectly proportional. Collinearity is essentially a numerical problem. Intuitively, it is clear that it creates indeterminacy as we are regressing twice on the same variable. In particular, the standard estimators given by (26) and (27) cannot be used because the relative formulas become meaningless.

In principle, collinearity can be easily resolved by eliminating one or more regressors. The problem with collinearity is that some variables might be very close to collinearity, thus leading to numerical problems and indeterminacy of results. In practice, this might happen for many different numerical artifacts. Detecting and analyzing collinearity is a rather delicate problem. In principle one could detect collinearity by computing the determinant of $\mathbf{X}'\mathbf{X}$. The difficulty resides in analyzing situations where this determinant is very small but not zero. One possible strategy for detecting and removing collinearity is to go through a

process of orthogonalization of variables. (See Hendry [1995].)

Increasing the Number of Regressors

Increasing the number of regressors does not always improve regressions. The econometric theorem known as Pyrrho's lemma relates to the number of regressors. (See Dijkstra [1995].) Pyrrho's lemma states that by adding one special regressor to a linear regression, it is possible to arbitrarily change the size and sign of regression coefficients as well as to obtain an arbitrary goodness of fit. This result, rather technical, seems artificial as the regressor is an artificially constructed variable. It is, however, a perfectly rigorous result; it tells us that, if we add regressors without a proper design and testing methodology, we risk obtaining spurious results.

Pyrrho's lemma is the proof that modeling results can be arbitrarily manipulated in-sample even in the simple context of linear regressions. In fact, by adding regressors one might obtain an excellent fit in-sample though these regressors might have no predictive power out-of-sample. In addition, the size and even the sign of the regression relationships can be artificially altered in-sample.

The above observations are especially important for those financial models that seek to forecast prices, returns, or rates based on regressions over economic or fundamental variables. With modern computers, by trial and error, one might find a complex structure of regressions that give very good results in-sample but have no real forecasting power.

KEY POINTS

- In regression analysis, the relationship between a random variable, called the dependent variable, and one or more variables referred to as the independent variables, regressors, or explanatory variables (which can be random variables or deterministic variables) is estimated.
- Factorization, which involves expressing a joint density as a product of a marginal density and a conditional density, is the conceptual basis of financial econometrics.
- An econometric model is a probe that extracts independent samples—the noise terms—from highly dependent variables.
- Regressions have a twofold nature: they can be either (1) the representation of dependence in terms of conditional expectations and conditional distributions or (2) the representation of dependence of random variables on deterministic parameters.
- In many applications in financial modeling, the regressors are deterministic variables. Therefore, on a conceptual level, regressions with deterministic regressors are different from cases where regressors are random variables. In particular, a financial modeler cannot view the regression as a conditional expectation.
- There are two main estimation techniques for estimating the parameters of a regression: maximum likelihood method and ordinary least squares method. The maximum likelihood principle requires maximization of the log-likelihood function. The ordinary least squares method requires minimization of the sum of the squared residuals. The ordinary least squares estimators are the best linear unbiased estimators.
- Because the estimated regression parameters depend on the sample, they are random variables whose distribution is to be determined. The sampling distributions differ depending on whether the regressors are assumed to be fixed deterministic variables or random variables.
- A measure of the quality of approximation offered by a linear regression is given by the variance of the residuals. If residuals are large, the regression model has little explanatory power. However, the size of the average residual in itself is meaningless as it has to be

compared with the range of the variables. A widely used measure of the quality and usefulness of a regression model is given by the coefficient of determination, denoted by R^2 or R -squared, that can attain a value from zero to one. The adjusted R^2 is defined as R^2 corrected by a penalty function that takes into account the number of regressors in the model.

- Stepwise regression is a model-building technique for regression designs. The two methodologies for stepwise regression are the backward stepwise method and the backward removal method.
- A time series is said to be autocorrelated if each term is correlated with its predecessor so that the variance of each term is partially explained by regressing each term on its predecessor. Autocorrelation of residuals, a violation of the regression model assumptions, is quite common in financial estimation where financial modelers regress quantities that are time series. When there is autocorrelation present in a time series, the generalized least squares estimation method is used. The Durbin-Watson test or the Dickey-Fuller test can be utilized to gauge test for the presence of autocorrelation for the residuals.
- Three other situations where there are possible pitfalls of using regressions are spurious regressions with integrated variables, collinearity, and increasing the number of regressors. Spurious regressions occur when an apparently meaningful regression between variables that are independent is estimated. Collinearity occurs when two or more regressors in a regression model have a linear deterministic relationship.
- Pyrrho's lemma, which relates to the number of regressors in a regression model, states that by adding one special regressor to a linear regression, it is possible to arbitrarily change the size and sign of regression coefficients as

well as to obtain an arbitrary goodness of fit. Pyrrho's lemma is the proof that modeling results can be arbitrarily manipulated in-sample even in the simple context of linear regressions.

NOTES

1. The data were supplied by David Wright of Northern Illinois University.
2. We are grateful to Robert Scott of the Bank for International Settlement for suggesting this illustration and for providing the data.

REFERENCES

- Bradley, B., and Taqqu, M. (2003). Financial risk and heavy tails. In S. T. Rachev (ed.), *Handbook of Heavy Tailed Distributions in Finance* (pp. 35–103). Amsterdam: Elsevier/North Holland.
- Dijkstra, T. K. (1995). Pyrrho's lemma, or have it your way. *Metrika* 42: 119–225.
- Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: Properties and pitfalls. In M. Dempster (ed.), *Risk Management: Value at Risk and Beyond* (pp. 176–223). Cambridge: Cambridge University Press.
- Granger, C., and Newbold, P. (1974). Spurious regression in econometrics. *Journal of Econometrics* 2: 111–120.
- Hendry, D. F. (1995). *Dynamic Econometrics*. Oxford: Oxford University Press.
- Markowitz, H. M. (1959). *Portfolio Selection: Efficient Diversification of Investments*. New Haven, CT: Cowles Foundation for Research in Economics.
- Reilly, F. K., Wright, D. J., and Johnson, R. R. (2007). An analysis of the interest rate sensitivity of common stocks. *Journal of Portfolio Management* 33, 2: 85–107.
- Sharpe, W. F. (1963). A simplified model for portfolio analysis. *Management Science* 9, 1: 277–293.

Categorical and Dummy Variables in Regression Models

SERGIO M. FOCARDI, PhD
Partner, The Intertek Group

FRANK J. FABOZZI, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

Abstract: In the application of regression analysis there are many situations where either the dependent variable or one or more of the regressors are categorical variables. When one or more categorical variables are used as regressors, a financial modeler must understand how to code the data, test for the significance of the categorical variables, and, based on the coding, how to interpret the estimated parameters. When the dependent variable is a categorical variable, the model is a probability model.

There are many times in the application of regression analysis when the financial modeler will need to include a categorical variable rather than a continuous variable as a regressor. *Categorical variables* are variables that represent group membership. For example, given a set of bonds, the rating is a categorical variable that indicates to what category—AA, BB, and so on—each bond belongs. A categorical variable does not have a numerical value or a numerical interpretation in itself. Thus the fact that a bond is in category AA or BB does not, in itself, measure any quantitative characteristic of the bond, though quantitative attributes such as a bond's yield spread can be associated with each category.

In this entry, we will discuss how to deal with regressors that are categorical variables in a regression model. There are also applications

where the dependent variable may be a categorical variable. For example, the dependent variable could be bankruptcy or nonbankruptcy of a company over some period of time. In such cases, the product of a regression is a probability. Probability models of this type include linear probability, logit regression, and probit linear models.

INDEPENDENT CATEGORICAL VARIABLES

Categorical input variables are used to cluster input data into different groups. That is, suppose we are given a set of input-output data and a partition of the data set in a number of subsets A_i so that each data point belongs to one and only one set. The A_i represent a categorical input variable. In financial econometrics

categories might represent, for example, different market regimes, economic states, ratings, countries, industries, or sectors.

We cannot, per se, mix quantitative input variables and categorical variables. For example, we cannot sum yield spreads and their ratings. However, we can perform a transformation that allows the mixing of categorical and quantitative variables. Let's see how. Suppose first that there is only one categorical input variable D , one quantitative input variable X , and one quantitative output variable Y . Consider our set of quantitative data, that is, quantitative observations. We organize data in a matrix form as usual:

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_T \end{bmatrix}, X = \begin{bmatrix} 1 & X_{11} \\ \vdots & \vdots \\ 1 & X_{T1} \end{bmatrix}$$

Suppose data belong to two categories. An explanatory variable that distinguishes only two categories is called a dichotomous variable. The key is to represent a dichotomous categorical variable as a numerical variable D , called a *dummy variable*, that can assume the two values 0,1. We can now add the variable D to the input variables to represent membership in one or the other group:

$$X = \begin{bmatrix} D_1 & 1 & X_{11} \\ \vdots & \vdots & \vdots \\ D_T & 1 & X_{T1} \end{bmatrix}$$

If $D_i = 0$, the data X_i belong to the first category; if $D_i = 1$, the data X_i belong to the second category.

Consider now the regression equation

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

In financial econometric applications, the index i will be time or a variable that identifies a cross section of assets, such as bond issues. Consider that we can write three separate regression equations, one for those data that correspond to $D = 1$, one for those data that correspond to $D = 0$, and one for the fully pooled data. Suppose now that the three equations differ by the

intercept term but have the same slope. Let's explicitly write the two equations for those data that correspond to $D = 1$ and for those data that correspond to $D = 0$:

$$y_i = \begin{cases} \beta_{00} + \beta_1 X_i + \varepsilon_i, & \text{if } D_i = 0 \\ \beta_{01} + \beta_1 X_i + \varepsilon_i, & \text{if } D_i = 1 \end{cases}$$

where i defines the observations that belong to the first category when the dummy variable D assumes value 0 and also defines the observations that belong to the second category when the dummy variable D assumes value 1. If the two categories are recession and expansion, the first equation might hold in periods of expansion and the second in periods of recession. If the two categories are investment-grade bonds and noninvestment-grade bonds, the two equations apply to different cross sections of bonds, as will be illustrated in an example later in this entry.

Observe now that, under the assumption that only the intercept term differs in the two equations, the two equations can be combined into a single equation in the following way:

$$Y_i = \beta_{00} + \gamma D(i) + \beta_1 X_i + \varepsilon_i$$

where $\gamma = \beta_{01} - \beta_{00}$ represents the difference of the intercept for the two categories. In this way we have defined a single regression equation with two independent quantitative variables, X , D , to which we can apply all the usual tools of regression analysis, including the ordinary least squares (OLS) estimation method and all the tests. By estimating the coefficients of this regression, we obtain the common slope and two intercepts. Observe that we would obtain the same result if the categories were inverted. However, the interpretation of the estimated parameter for the categorical variable would differ depending on which category is omitted.

Thus far we have assumed that there is no interaction between the categorical and the quantitative variable, that is, the slope of the regression is the same for the two categories. This means that the effects of variables are additive; that is, the effect of one variable is added

regardless of the value taken by the other variable. In many applications, this is an unrealistic assumption.

Using dummy variables, the treatment is the same as that applied to intercepts. Consider the regression equation $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ and write two regression equations for the two categories as we did above:

$$y_i = \begin{cases} \beta_0 + \beta_{10} X_i + \varepsilon_i, & \text{if } D_i = 0 \\ \beta_0 + \beta_{11} X_i + \varepsilon_i, & \text{if } D_i = 1 \end{cases}$$

We can couple these two equations in a single equation as follows:

$$Y_i = \beta_0 + \beta_{10} X_i + \delta(D_i X_i) + \varepsilon_i$$

where $\delta = \beta_{11} - \beta_{10}$. In fact, the above equation is identical to the first equation for $D_i = 0$ and to the second for $D_i = 1$. This regression can be estimated with the usual LS methods.

In practice, it is rarely appropriate to consider only interactions and not the intercept, which is the main effect. We call *marginalization* the fact that the interaction effect is marginal with respect to the main effect. However, we can easily construct a model that combines both effects. In fact we can write the following regression adding two variables, the dummy D and the interaction DX :

$$Y_i = \beta_0 + \gamma D_i + \beta_1 X_i + \delta(D_i X_i) + \varepsilon_i$$

This regression equation, which now includes three regressors, combines both effects.

The above process of introducing dummy variables can be generalized to regressions with multiple variables. Consider the following regression:

$$Y_i = \beta_0 + \sum_{j=1}^N \beta_j X_{ij} + \varepsilon_i$$

where data can be partitioned in two categories with the use of a dummy variable:

$$\mathbf{X} = \begin{bmatrix} D_1 & 1 & X_{11} & \cdots & X_{1N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D_T & 1 & X_{T1} & \cdots & X_{TN} \end{bmatrix}$$

We can introduce the dummy D as well as its interaction with the N quantitative variable and

thus write the following equation:

$$Y_i = \beta_0 + \gamma_i D_i + \sum_{j=1}^N \beta_j X_{ij} + \sum_{j=1}^N \delta_{ij}(D_i X_{ij}) + \varepsilon_i$$

The above discussion depends critically on the fact that there are only two categories, a fact that allows one to use the numerical variable 0,1 to identify the two categories. However, the process can be easily extended to multiple categories by adding dummy variables. Suppose there are $K > 2$ categories. An explanatory variable that distinguishes between more than two categories is called a polytomous variable.

Suppose there are three categories, A , B , and C . Consider a dummy variable $D1$ that assumes a value one on the elements of A and zero on all the others. Let's now add a second dummy variable $D2$ that assumes the value one on the elements of the category B and zero on all the others. The three categories are now completely identified: A is identified by the values 1,0 of the two dummy variables, B by the values 0,1, and C by the values 0,0. Note that the values 1,1 do not identify any category. This process can be extended to any number of categories. If there are K categories, we need $K - 1$ dummy variables.

How can we determine if a given categorization is useful? It is quite obvious that many categorizations will be totally useless for the purpose of any econometric regression. If we categorize bonds in function of the color of the logo of the issuer, it is quite obvious that we obtain meaningless results. In other cases, however, distinctions can be subtle and important. Consider the question of market regime shifts or structural breaks. These are delicate questions that can be addressed only with appropriate statistical tests.

A word of caution about statistical tests is in order. Statistical tests typically work under the assumptions of the model and might be misleading if these assumptions are violated. If we try to fit a linear model to a process that is inherently nonlinear, tests might be misleading. It is good practice to use several tests and to

be particularly attentive to inconsistencies between test results. Inconsistencies signal potential problems in applying tests, typically model misspecification.

The t -statistic applied to the regression coefficients of dummy variables offer a set of important tests to judge which regressors are significant. The t -statistics are the coefficients divided by their respective squared errors. The p -value associated with each coefficient estimate is the probability of the hypothesis that the corresponding coefficient is zero, that is, that the corresponding variable is irrelevant.

We can also use the F -test to test the significance of each specific dummy variable. To do so we can run the regression with and without that variable and form the corresponding F -test. The *Chow test* is the F -test to gauge if all the dummy variables are collectively irrelevant (see Chow, 1960). The Chow test is an F -test of mutual exclusion, written as follows:

$$F = \frac{[SSR - (SSR_1 + SSR_2)] [n - 2(k + 1)]}{SSR_1 + SSR_2} \frac{1}{k + 1}$$

where

SSR_1 = the squared sum of residuals of the regression run with data in the first category without dummy variables

SSR_2 = the squared sum of residuals of the regression run with data in the second category without dummy variables

SSR = the squared sum of residuals of the regression run with fully pooled data without dummy variables

Observe that $SSR_1 + SSR_2$ is equal to the squared sum of residuals of the regression run on fully pooled data but with dummy variables. Thus the Chow test is the F -test of the unrestricted regressions with and without dummy variables.

Illustration: Predicting Corporate Bond Yield Spreads

To illustrate the use of dummy variables, we will estimate a model to predict corporate bond spreads.¹ The regression is relative to a cross section of bonds. The regression equation is the following:

$$\text{Spread}_i = \beta_0 + \beta_1 \text{Coupon}_i + \beta_2 \text{CoverageRatio}_i + \beta_3 \text{LoggedEBIT}_i + \varepsilon_i$$

where

Spread_i = option-adjusted spread (in basis points) for the bond issue of company i

Coupon_i = coupon rate for the bond of company i , expressed without considering percentage sign (i.e., 7.5% = 7.5)

CoverageRatio_i = earnings before interest, taxes, depreciation and amortization (EBITDA) divided by interest expense for company i

LoggedEBIT_i = logarithm of earnings (earnings before interest and taxes, EBIT, in millions of dollars) for company i

The dependent variable, Spread, is not measured by the typically nominal spread but by the option-adjusted spread. This spread measure adjusts for any embedded options in a bond (see Chapter 6 in Fabozzi, 2006).

Theory would suggest the following properties for the estimated coefficients:

- The higher the coupon rate, the greater the issuer's default risk and hence the larger the spread. Therefore, a positive coefficient for the coupon rate is expected.
- A coverage ratio is a measure of a company's ability to satisfy fixed obligations, such as interest, principal repayment, or lease payments. There are various coverage ratios. The one used in this illustration is the ratio of the earnings before interest, taxes,

depreciation, and amortization (EBITDA) divided by interest expense. Since the higher the coverage ratio the lower the default risk, an inverse relationship is expected between the spread and the coverage ratio; that is, the estimated coefficient for the coverage ratio is expected to be negative.

- There are various measures of earnings reported in financial statements. Earnings in this illustration is defined as the trailing 12-months earnings before interest and taxes (EBIT). Holding other factors constant, it is expected that the larger the EBIT, the lower the default risk and therefore an inverse relationship (negative coefficient) is expected.

We used 100 observations at two different dates, 6/6/05 and 11/28/05; thus there are 200 observations in total. This will allow us to test if there is a difference in the spread regression for investment-grade and noninvestment grade bonds using all observations. We will then test to see if there is any structural break between the two dates. We organize the data in matrix form as usual. Data are shown in Table 1. The second column indicates that data belong to two categories and suggests the use of one dummy variable. Another dummy variable is used later to distinguish between the two dates. Let's first estimate the regression equation for the fully pooled data, that is, all data without any distinction in categories. The estimated coefficients for the model and their corresponding *t*-statistics are shown below:

Coefficient	Estimated Coefficient	Standard Error	<i>t</i> -statistic	<i>p</i> -value
β_0	157.01	89.56	1.753	0.081
β_1	61.27	8.03	7.630	9.98E-13
β_2	-13.20	2.27	-5.800	2.61E-08
β_3	-90.88	16.32	-5.568	8.41E-08

Other regression results are:

SSR: 2.3666e + 006
F-statistic: 89.38
p-value: 0
*R*²: 0.57

Given the high value of the *F*-statistic and the *p*-value close to zero, the regression is significant. The coefficient for the three regressors is statistically significant and has the expected sign. However, the intercept term is not statistically significant. The residuals are given in the first column of Table 2.

Let's now analyze if we obtain a better fit if we consider the two categories of investment-grade and below investment-grade bonds. It should be emphasized that this is only an exercise to show the application of regression analysis. The conclusions we reach are not meaningful from an econometric point of view given the small size of the database. The new equation is written as follows:

$$\text{Spread}_i = \beta_0 + \beta_1 D1_i + \beta_2 \text{Coupon}_i + \beta_3 D1_i \text{Coupon}_i + \beta_4 \text{CoverageRatio}_i + \beta_5 D1_i \text{CoverageRatio}_i + \beta_6 \text{LoggedEBIT}_i + \beta_7 D1_i \text{LoggedEBIT}_i + \varepsilon_i$$

There are now seven variables and eight parameters to estimate. The estimated model coefficients and the *t*-statistics are shown below:

Coefficient	Estimated Coefficient	Standard Error	<i>t</i> -statistic	<i>p</i> -value
β_0	284.52	73.63	3.86	0.00
β_1	597.88	478.74	1.25	0.21
β_2	37.12	7.07	5.25	3.96E-07
β_3	-45.54	38.77	-1.17	0.24
β_4	-10.33	1.84	-5.60	7.24E-08
β_5	50.13	40.42	1.24	0.22
β_6	-83.76	13.63	-6.15	4.52E-09
β_7	-0.24	62.50	-0.00	1.00

Other regression results are:

SSR: 1.4744e + 006
F-statistic: 76.83
p-value: 0
*R*²: 0.73

The Chow test has the value 16.60. The *F*-statistic and the Chow test suggest that the use of dummy variables has greatly improved the goodness of fit of the regression, even after compensating for the increase in the number of

Table 1 Regression Data for the Bond Spread Application: 11/28/2005 and 06/06/2005

Issue #	Spread, 11/28/05	CCC+ and Below	Coupon	Coverage Ratio	Logged EBIT	Spread, 6/6/05	CCC+ and Below	Coupon	Coverage Ratio	Logged EBIT
1	509	0	7.400	2.085	2.121	473	0	7.400	2.087	2.111
2	584	0	8.500	2.085	2.121	529	0	8.500	2.087	2.111
3	247	0	8.375	9.603	2.507	377	0	8.375	5.424	2.234
4	73	0	6.650	11.507	3.326	130	0	6.650	9.804	3.263
5	156	0	7.125	11.507	3.326	181	0	7.125	9.804	3.263
6	240	0	7.250	2.819	2.149	312	0	7.250	2.757	2.227
7	866	1	9.000	1.530	2.297	852	1	9.000	1.409	1.716
8	275	0	5.950	8.761	2.250	227	0	5.950	11.031	2.166
9	515	0	8.000	2.694	2.210	480	0	8.000	2.651	2.163
10	251	0	7.875	8.289	1.698	339	0	7.875	8.231	1.951
11	507	0	9.375	2.131	2.113	452	0	9.375	2.039	2.042
12	223	0	7.750	4.040	2.618	237	0	7.750	3.715	2.557
13	71	0	7.250	7.064	2.348	90	0	7.250	7.083	2.296
14	507	0	8.000	2.656	1.753	556	0	8.000	2.681	1.797
15	566	1	9.875	1.030	1.685	634	1	9.875	1.316	1.677
16	213	0	7.500	11.219	3.116	216	0	7.500	10.298	2.996
17	226	0	6.875	11.219	3.116	204	0	6.875	10.298	2.996
18	192	0	7.750	11.219	3.116	201	0	7.750	10.298	2.996
19	266	0	6.250	3.276	2.744	298	0	6.250	3.107	2.653
20	308	0	9.250	3.276	2.744	299	0	9.250	3.107	2.653
21	263	0	7.750	2.096	1.756	266	0	7.750	2.006	3.038
22	215	0	7.190	7.096	3.469	259	0	7.190	6.552	3.453
23	291	0	7.690	7.096	3.469	315	0	7.690	6.552	3.453
24	324	0	8.360	7.096	3.469	331	0	8.360	6.552	3.453
25	272	0	6.875	8.612	1.865	318	0	6.875	9.093	2.074
26	189	0	8.000	4.444	2.790	209	0	8.000	5.002	2.756
27	383	0	7.375	2.366	2.733	417	0	7.375	2.375	2.727
28	207	0	7.000	2.366	2.733	200	0	7.000	2.375	2.727
29	212	0	6.900	4.751	2.847	235	0	6.900	4.528	2.822
30	246	0	7.500	19.454	2.332	307	0	7.500	16.656	2.181
31	327	0	6.625	3.266	2.475	365	0	6.625	2.595	2.510
32	160	0	7.150	3.266	2.475	237	0	7.150	2.595	2.510
33	148	0	6.300	3.266	2.475	253	0	6.300	2.595	2.510
34	231	0	6.625	3.266	2.475	281	0	6.625	2.595	2.510
35	213	0	6.690	3.266	2.475	185	0	6.690	2.595	2.510
36	350	0	7.130	3.266	2.475	379	0	7.130	2.595	2.510
37	334	0	6.875	4.310	2.203	254	0	6.875	5.036	2.155
38	817	1	8.625	1.780	1.965	635	0	8.625	1.851	1.935
39	359	0	7.550	2.951	3.078	410	0	7.550	2.035	3.008
40	189	0	6.500	8.518	2.582	213	0	6.500	13.077	2.479
41	138	0	6.950	25.313	2.520	161	0	6.950	24.388	2.488
42	351	0	9.500	3.242	1.935	424	0	9.500	2.787	1.876
43	439	0	8.250	2.502	1.670	483	0	8.250	2.494	1.697
44	347	0	7.700	4.327	3.165	214	0	7.700	4.276	3.226
45	390	0	7.750	4.327	3.165	260	0	7.750	4.276	3.226
46	149	0	8.000	4.327	3.165	189	0	8.000	4.276	3.226
47	194	0	6.625	4.430	3.077	257	0	6.625	4.285	2.972
48	244	0	8.500	4.430	3.077	263	0	8.500	4.285	2.972
49	566	1	10.375	2.036	1.081	839	1	10.375	2.032	1.014
50	185	0	6.300	7.096	3.469	236	0	6.300	6.552	3.453
51	196	0	6.375	7.096	3.469	221	0	6.375	6.552	3.453
52	317	0	6.625	3.075	2.587	389	0	6.625	2.785	2.551
53	330	0	8.250	3.075	2.587	331	0	8.250	2.785	2.551

Table 1 (Continued)

Issue #	Spread, 11/28/05	CCC+ and Below	Coupon	Coverage Ratio	Logged EBIT	Spread, 6/6/05	CCC+ and Below	Coupon	Coverage Ratio	Logged EBIT
54	159	0	6.875	8.286	3.146	216	0	6.875	7.210	3.098
55	191	0	7.125	8.286	3.146	257	0	7.125	7.210	3.098
56	148	0	7.375	8.286	3.146	117	0	7.375	7.210	3.098
57	112	0	7.600	8.286	3.146	151	0	7.600	7.210	3.098
58	171	0	7.650	8.286	3.146	221	0	7.650	7.210	3.098
59	319	0	7.375	3.847	1.869	273	0	7.375	4.299	1.860
60	250	0	7.375	12.656	2.286	289	0	7.375	8.713	2.364
61	146	0	5.500	5.365	3.175	226	0	5.500	5.147	3.190
62	332	0	6.450	5.365	3.175	345	0	6.450	5.147	3.190
63	354	0	6.500	5.365	3.175	348	0	6.500	5.147	3.190
64	206	0	6.625	7.140	2.266	261	0	6.625	5.596	2.091
65	558	0	7.875	2.050	2.290	455	0	7.875	2.120	2.333
66	190	0	6.000	2.925	3.085	204	0	6.000	3.380	2.986
67	232	0	6.750	2.925	3.085	244	0	6.750	3.380	2.986
68	913	1	11.250	2.174	1.256	733	0	11.250	2.262	1.313
69	380	0	9.750	4.216	1.465	340	0	9.750	4.388	1.554
70	174	0	6.500	4.281	2.566	208	0	6.500	4.122	2.563
71	190	0	7.450	10.547	2.725	173	0	7.450	8.607	2.775
72	208	0	7.125	2.835	3.109	259	0	7.125	2.813	3.122
73	272	0	6.500	5.885	2.695	282	0	6.500	5.927	2.644
74	249	0	6.125	5.133	2.682	235	0	6.125	6.619	2.645
75	278	0	8.750	6.562	2.802	274	0	8.750	7.433	2.785
76	252	0	7.750	2.822	2.905	197	0	7.750	2.691	2.908
77	321	0	7.500	2.822	2.905	226	0	7.500	2.691	2.908
78	379	0	7.750	4.093	2.068	362	0	7.750	4.296	2.030
79	185	0	6.875	6.074	2.657	181	0	6.875	5.294	2.469
80	307	0	7.250	5.996	2.247	272	0	7.250	3.610	2.119
81	533	0	10.625	1.487	1.950	419	0	10.625	1.717	2.081
82	627	0	8.875	1.487	1.950	446	0	8.875	1.717	2.081
83	239	0	8.875	2.994	2.186	241	0	8.875	3.858	2.161
84	240	0	7.375	8.160	2.225	274	0	7.375	8.187	2.075
85	634	0	8.500	2.663	2.337	371	0	8.500	2.674	2.253
86	631	1	7.700	2.389	2.577	654	1	7.700	2.364	2.632
87	679	1	9.250	2.389	2.577	630	1	9.250	2.364	2.632
88	556	1	9.750	1.339	1.850	883	1	9.750	1.422	1.945
89	564	1	9.750	1.861	2.176	775	1	9.750	1.630	1.979
90	209	0	6.750	8.048	2.220	223	0	6.750	7.505	2.092
91	190	0	6.500	4.932	2.524	232	0	6.500	4.626	2.468
92	390	0	6.875	6.366	1.413	403	0	6.875	5.033	1.790
93	377	0	10.250	2.157	2.292	386	0	10.250	2.057	2.262
94	143	0	5.750	11.306	2.580	110	0	5.750	9.777	2.473
95	207	0	7.250	2.835	3.109	250	0	7.250	2.813	3.122
96	253	0	6.500	4.918	2.142	317	0	6.500	2.884	1.733
97	530	1	8.500	0.527	2.807	654	1	8.500	1.327	2.904
98	481	0	6.750	2.677	1.858	439	0	6.750	3.106	1.991
99	270	0	7.625	2.835	3.109	242	0	7.625	2.813	3.122
100	190	0	7.125	9.244	3.021	178	0	7.125	7.583	3.138

Notes:

Spread = option-adjusted spread (in basis points)

Coupon = coupon rate, expressed without considering percentage sign (i.e., 7.5% = 7.5)

Coverage Ratio = EBITDA divided by interest expense for company

Logged EBIT = logarithm of earnings (EBIT in millions of dollars)

Table 2 Illustration of Residuals and Leverage for Corporate Bond Spread

Issue #	Residuals	Residuals Dummy 1	Residuals Dummy 2
1	118.79930	148.931400	162.198700
2	126.39350	183.097400	200.622000
3	-68.57770	-39.278100	-26.716500
4	-37.26080	-60.947500	-71.034400
5	16.63214	4.419645	-3.828890
6	-128.76600	-104.569000	-92.122000
7	386.42330	191.377200	217.840000
8	73.53972	48.516800	56.58778
9	104.15990	146.400600	160.438900
10	-124.78700	-98.020100	-71.374300
11	-4.28874	73.473220	94.555400
12	-117.58200	-88.168700	-82.883100
13	-223.61800	-213.055000	-202.748000
14	54.13075	99.735710	123.153000
15	-29.42160	-132.755000	-179.955000
16	27.74192	26.913670	24.308960
17	79.04072	63.114850	58.091160
18	-8.57759	-3.366800	-5.003930
19	18.62462	13.109110	9.664499
20	-123.21000	-56.256500	-48.090100
21	-181.64800	-140.494000	-118.369000
22	26.43157	27.457990	14.487850
23	71.79254	84.897050	73.862080
24	63.73623	93.025400	84.583560
25	-23.09740	-22.603200	-3.106990
26	-146.00700	-112.938000	-110.018000
27	53.72288	78.075810	78.781050
28	-99.29780	-84.003500	-84.749600
29	-46.31030	-41.105600	-43.489200
30	98.22006	79.285040	96.588250
31	32.05062	37.541930	41.075430
32	-167.12000	-148.947000	-143.382000
33	-127.03400	-129.393000	-127.118000
34	-63.94940	-58.458100	-54.924600
35	-85.93250	-78.871000	-75.085900
36	24.10520	41.795380	47.283410
37	12.86740	23.326060	33.884440
38	333.53890	101.376800	173.584400
39	58.02881	82.472150	77.040360
40	-19.14100	-32.550700	-29.298900
41	118.41190	67.990200	81.986050
42	-169.48100	-90.625700	-64.883800
43	-38.74030	13.936980	39.950520
44	62.91014	86.397490	80.392250
45	102.84620	127.541400	121.729700
46	-153.47300	-122.739000	-127.583000
47	-30.81510	-32.968700	-41.285200
48	-95.711400	-52.572300	-53.631800
49	-101.678000	-219.347000	-237.977000
50	50.969050	30.496460	14.081700
51	57.373200	38.712320	22.587840
52	29.717770	34.958870	36.101100
53	-56.859100	-12.364200	-4.932630
54	-23.959100	-31.659900	-38.650000
55	-7.278620	-8.940330	-14.962800

Table 2 (Continued)

Issue #	Residuals	Residuals Dummy 1	Residuals Dummy 2
56	-65.598100	-61.220800	-66.275700
57	-115.386000	-105.573000	-109.757000
58	-59.449600	-48.429300	-52.419900
59	-69.299000	-43.044000	-23.885700
60	15.946800	13.880220	28.513500
61	11.362190	-21.353800	-35.607900
62	139.148000	129.380400	118.803100
63	158.084100	149.524300	139.140600
64	-56.785300	-60.952000	-51.339900
65	153.651800	194.149900	205.750200
66	-15.653600	-28.630900	-40.227500
67	-19.612200	-14.472300	-23.166100
68	209.488200	144.261600	67.891100
69	-185.659000	-100.217000	-63.396000
70	-91.541800	-92.646100	-91.015000
71	-36.623800	-33.937000	-29.003400
72	-65.586300	-51.301800	-59.080100
73	39.294110	32.661770	32.391920
74	28.197460	14.759650	12.952710
75	-73.910000	-28.902200	-22.353300
76	-78.608000	-47.733800	-48.902600
77	5.711553	30.546620	28.410290
78	-10.926100	22.258560	38.888810
79	-71.611400	-69.462200	-67.416900
80	-10.848000	3.505179	15.383910
81	-78.195700	32.775440	61.748590
82	123.041000	191.738700	213.938800
83	-223.662000	-160.978000	-142.925000
84	-58.977600	-47.671100	-33.850800
85	203.727300	257.223800	270.556600
86	267.904600	-65.208100	89.636310
87	220.923600	-4.162260	42.473790
88	-12.621600	-142.213000	-168.474000
89	31.862060	-127.616000	-134.267000
90	-53.593800	-57.028600	-45.579800
91	-70.794900	-73.470000	-70.669700
92	24.164780	34.342730	62.098550
93	-171.291000	-73.744300	-52.943000
94	17.439710	-22.092800	-20.420000
95	-74.246100	-56.942100	-64.236600
96	-42.690600	-42.602900	-31.958300
97	114.168900	-66.109500	-66.049500
98	114.578500	129.177300	145.600600
99	-34.225400	-7.862790	-13.705900
100	-6.958960	-10.488100	-13.508000
101	81.920940	112.117900	101.420600
102	70.515070	127.283800	120.844000
103	-18.587600	24.683610	20.132390
104	-8.443100	-26.784100	-28.884400
105	13.449820	6.582981	6.321103
106	-50.430600	-26.617000	-36.781100
107	318.056000	133.403000	130.828300
108	47.876010	16.919350	5.068270
109	64.341610	107.038200	99.281600
110	-14.573200	10.557760	3.393970

(Continued)

Table 2 (Continued)

Issue #	Residuals	Residuals Dummy 1	Residuals Dummy 2
111	-66.995600	11.539420	7.987728
112	-113.425000	-82.640800	-88.147800
113	-209.054000	-198.177000	-205.892000
114	107.522000	152.737700	142.464600
115	41.638860	-76.825800	-145.458000
116	7.647833	10.327540	9.887700
117	33.946630	21.528710	18.669900
118	-22.671700	-13.952900	-13.425200
119	40.107630	35.729610	24.798540
120	-142.727000	-74.636000	-73.956000
121	-63.286100	-31.013100	-33.970100
122	61.774140	64.481450	64.302480
123	87.135110	101.920500	103.676700
124	62.078800	93.048860	97.398200
125	48.320900	45.935300	36.150130
126	-121.736000	-90.029000	-92.609500
127	87.253680	111.626800	105.229900
128	-106.767000	-91.452500	-99.300700
129	-28.566900	-22.540100	-29.135400
130	108.560100	98.752280	95.570570
131	64.418690	71.586810	60.886980
132	-95.752300	-75.902200	-84.570100
133	-27.665900	-28.348600	-40.306300
134	-19.581300	-12.413200	-23.113000
135	-119.564000	-110.826000	-121.274000
136	47.473260	66.840260	58.094960
137	-61.953700	-53.237800	-64.316600
138	149.786400	211.505100	204.226300
139	90.609530	118.184700	114.258300
140	55.650810	29.860840	23.239180
141	126.240500	78.712630	79.050720
142	-107.826000	-27.243600	-31.116800
143	7.614932	60.121850	50.036220
144	-65.174500	-41.979400	-42.794500
145	-22.238400	2.164489	1.542950
146	-108.558000	-78.116000	-77.769900
147	20.679750	19.696850	12.963030
148	-88.216600	-43.906700	-43.383600
149	165.253100	48.262590	-23.500200
150	93.311620	74.519920	70.896340
151	73.715770	56.735780	53.402470
152	94.629570	100.961000	90.629950
153	-62.947300	-17.362000	-21.403800
154	14.480140	10.216950	6.659433
155	40.160620	41.936480	39.346550
156	-115.159000	-107.344000	-108.966000
157	-94.946500	-81.696400	-82.447900
158	-28.010400	-13.552500	-14.110500
159	-110.127000	-85.111400	-96.632900
160	9.959282	18.682370	12.662020
161	89.889700	57.689740	48.509480
162	150.675500	141.424000	135.920500
163	150.611600	142.567900	137.258000
164	-38.040900	-36.521000	-48.754100
165	55.443990	95.437610	88.132530

Table 2 (Continued)

Issue #	Residuals	Residuals Dummy 1	Residuals Dummy 2
166	-4.652580	-18.233400	-27.698600
167	-10.611100	-6.074840	-12.637200
168	35.778970	164.163000	162.921500
169	-215.328000	-131.013000	-135.422000
170	-59.986400	-60.605400	-70.729300
171	-74.693600	-66.782400	-69.716200
172	-13.734800	0.523639	-3.905600
173	45.295840	38.898770	30.164940
174	30.476800	13.024800	3.159872
175	-67.888500	-25.271900	-23.635500
176	-135.061000	-103.830000	-107.375000
177	-90.741200	-65.550000	-70.062300
178	-28.683300	4.187387	-4.706060
179	-103.027000	-97.290000	-106.078000
180	-88.975000	-66.845700	-77.367900
181	-177.281000	-67.904100	-66.493200
182	-43.044700	24.059160	18.696920
183	-212.505000	-152.131000	-155.963000
184	-38.210800	-25.916400	-34.173800
185	-66.764700	-12.702000	-17.886300
186	295.611300	-36.578800	106.036400
187	176.630300	-47.533000	-13.126100
188	324.060100	189.413000	136.666400
189	221.951100	76.029960	34.046210
190	-58.422000	-59.380500	-70.254000
191	-37.907200	-39.303500	-49.850800
192	53.841660	65.166450	51.559780
193	-166.323000	-68.275700	-66.904900
194	-45.521100	-79.888400	-90.959200
195	-30.394500	-13.116600	-17.062000
196	-42.709500	-33.855500	-50.285700
197	257.550200	34.224540	70.337910
198	90.307160	102.727000	89.148700
199	-61.373800	-35.037300	-37.531400
200	-30.310400	-29.889500	-32.034600

Notes:

Residuals: residuals from the pooled regression without dummy variables for investment grade.

Residuals Dummy 1: inclusion of dummy variable for investment grade.

Residuals Dummy 2: inclusion of dummy variable to test for regime shift.

parameters. The residuals of the model without and with dummy variable D1 are shown, respectively, in the second and third columns of Table 2.

Now let's use dummy variables to test if there is a regime shift between the two dates. This is a common use for dummy variables in practice. To this end we create a new dummy variable that has the value 0 for the first date 11/28/05

and 1 for the second date 6/6/05. The new equation is written as follows:

$$\begin{aligned}
 \text{Spread}_i = & \beta_0 + \beta_1 D2_i + \beta_2 \text{Coupon}_i \\
 & + \beta_3 D2_i \text{Coupon}_i + \beta_4 \text{CoverageRatio}_i \\
 & + \beta_5 D2_i \text{CoverageRatio}_i + \beta_6 \text{LoggedEBIT}_i \\
 & + \beta_7 D2_i \text{LoggedEBIT}_i + \varepsilon_i
 \end{aligned}$$

as in the previous case but with a different dummy variable. There are seven independent

variables and eight parameters to estimate. The estimated model coefficients and t -statistics are shown below:

Coefficient	Estimated Coefficient	Standard Error	t -statistic	p -value
β_0	257.26	79.71	3.28	0.00
β_1	82.17	61.63	1.33	0.18
β_2	33.25	7.11	4.67	5.53E-06
β_3	28.14	2.78	10.12	1.45E-19
β_4	-10.79	2.50	-4.32	2.49E-05
β_5	0.00	3.58	0.00	1.00
β_6	-63.20	18.04	-3.50	0.00
β_7	-27.48	24.34	-1.13	0.26

Other regression statistics are:

SSR: 1.5399e + 006

F -statistic: 72.39

p -value: 0

R^2 : 0.71

The Chow test has the value 14.73. The F -statistics and the Chow test suggest that there is indeed a regime shift and that the spread regressions at the two different dates are different. Again, the use of dummy variables has greatly improved the goodness of fit of the regression, even after compensating for the increase in the number of parameters. The residuals of the model with dummy variables D2 are shown in the next-to-the-last column of Table 2.

Illustration: Testing the Mutual Fund Characteristic Lines in Different Market Environments

The characteristic line of a mutual fund is the regression of the excess returns of a mutual fund on the market's excess returns:

$$y_{it} = \alpha_i + \beta_i x_t$$

where

y_{it} = mutual fund i 's excess return over the risk-free rate

x_t = market excess return over the risk-free rate

α_i and β_i = the regression parameters to be estimated for mutual fund i

We will first estimate the characteristic line for two large-cap mutual funds. Since we would prefer not to disclose the name of each fund, we simply refer to them as A and B. (Neither mutual fund selected is an index fund.) Because the two mutual funds are large-cap funds, the S&P 500 was used as the benchmark. The risk-free rate used was the 90-day Treasury bill rate. Ten years of monthly data were used from January 1, 1995 to December 31, 2004. The data are reported in Table 3. The first column in the table shows the month. The second and third columns give the return on the market return (r_{Mt}) and risk-free rate (r_{ft}), respectively. The fifth column is the excess market return, which is x_t in the regression equation. The seventh and eighth columns show the returns for mutual funds A and B, respectively. The excess returns for the two mutual funds (y_{it}) are given in the last two columns. The other columns will be explained shortly.

The results of the above regression for both mutual funds are shown in Table 4. The estimated β for both mutual funds is statistically significantly different from zero.

Let's now perform a simple application of the use of dummy variables by determining if the slope (beta) of the two mutual funds is different in a rising stock market ("up market") and a declining stock market ("down market"). To test this, we can write the following multiple regression model:

$$y_{it} = \alpha_i + \beta_{1i}x_t + \beta_{2i}(D_t x_t) + e_{it}$$

where D_t is the dummy variable that can take on a value of 1 or 0. We will let

$D_t = 1$ if period t is classified as an up market

$D_t = 0$ if period t is classified as a down market

The coefficient for the dummy variable is β_{2i} . If that coefficient is statistically significant, then for the mutual fund:

In an up market: $\beta_i = \beta_{1i} + \beta_{2i}$

In a down market: $\beta_i = \beta_{1i}$

Table 3 Data for Estimating Mutual Fund Characteristic Line with a Dummy Variable

Month Ended						Mutual Fund			
	r_M	r_{ft}	Dummy D_t	$r_M - f_{ft}$ x_t	$D_t x_t$	A r_t	B r_t	A y_t	B y_t
01/31/1995	2.60	0.42	0	2.18	0	0.65	1.28	0.23	0.86
02/28/1995	3.88	0.40	0	3.48	0	3.44	3.16	3.04	2.76
03/31/1995	2.96	0.46	1	2.50	2.5	2.89	2.58	2.43	2.12
04/30/1995	2.91	0.44	1	2.47	2.47	1.65	1.81	1.21	1.37
05/31/1995	3.95	0.54	1	3.41	3.41	2.66	2.96	2.12	2.42
06/30/1995	2.35	0.47	1	1.88	1.88	2.12	2.18	1.65	1.71
07/31/1995	3.33	0.45	1	2.88	2.88	3.64	3.28	3.19	2.83
08/31/1995	0.27	0.47	1	-0.20	-0.2	-0.40	0.98	-0.87	0.51
09/30/1995	4.19	0.43	1	3.76	3.76	3.06	3.47	2.63	3.04
10/31/1995	-0.35	0.47	1	-0.82	-0.82	-1.77	-0.63	-2.24	-1.10
11/30/1995	4.40	0.42	1	3.98	3.98	4.01	3.92	3.59	3.50
12/31/1995	1.85	0.49	1	1.36	1.36	1.29	1.73	0.80	1.24
01/31/1996	3.44	0.43	1	3.01	3.01	3.36	2.14	2.93	1.71
02/29/1996	0.96	0.39	1	0.57	0.57	1.53	1.88	1.14	1.49
03/31/1996	0.96	0.39	1	0.57	0.57	0.59	1.65	0.20	1.26
04/30/1996	1.47	0.46	1	1.01	1.01	1.46	1.83	1.00	1.37
05/31/1996	2.58	0.42	1	2.16	2.16	2.17	2.20	1.75	1.78
06/30/1996	0.41	0.40	1	0.01	0.01	-0.63	0.00	-1.03	-0.40
07/31/1996	-4.45	0.45	1	-4.90	-4.9	-4.30	-3.73	-4.75	-4.18
08/31/1996	2.12	0.41	0	1.71	0	2.73	2.24	2.32	1.83
09/30/1996	5.62	0.44	0	5.18	0	5.31	4.49	4.87	4.05
10/31/1996	2.74	0.42	1	2.32	2.32	1.42	1.34	1.00	0.92
11/30/1996	7.59	0.41	1	7.18	7.18	6.09	5.30	5.68	4.89
12/31/1996	-1.96	0.46	1	-2.42	-2.42	-1.38	-0.90	-1.84	-1.36
01/31/1997	6.21	0.45	1	5.76	5.76	4.15	5.73	3.70	5.28
02/28/1997	0.81	0.39	1	0.42	0.42	1.65	-1.36	1.26	-1.75
03/31/1997	-4.16	0.43	1	-4.59	-4.59	-4.56	-3.75	-4.99	-4.18
04/30/1997	5.97	0.43	1	5.54	5.54	4.63	3.38	4.20	2.95
05/31/1997	6.14	0.49	1	5.65	5.65	5.25	6.05	4.76	5.56
06/30/1997	4.46	0.37	1	4.09	4.09	2.98	2.90	2.61	2.53
07/31/1997	7.94	0.43	1	7.51	7.51	6.00	7.92	5.57	7.49
08/31/1997	-5.56	0.41	1	-5.97	-5.97	-4.40	-3.29	-4.81	-3.70
09/30/1997	5.48	0.44	1	5.04	5.04	5.70	4.97	5.26	4.53
10/31/1997	-3.34	0.42	1	-3.76	-3.76	-2.76	-2.58	-3.18	-3.00
11/30/1997	4.63	0.39	0	4.24	0	3.20	2.91	2.81	2.52
12/31/1997	1.72	0.48	1	1.24	1.24	1.71	2.41	1.23	1.93
01/31/1998	1.11	0.43	1	0.68	0.68	-0.01	-0.27	-0.44	-0.70
02/28/1998	7.21	0.39	1	6.82	6.82	5.50	6.84	5.11	6.45
03/31/1998	5.12	0.39	1	4.73	4.73	5.45	3.84	5.06	3.45
04/30/1998	1.01	0.43	1	0.58	0.58	-0.52	1.07	-0.95	0.64
05/31/1998	-1.72	0.40	1	-2.12	-2.12	-1.25	-1.30	-1.65	-1.70
06/30/1998	4.06	0.41	1	3.65	3.65	3.37	4.06	2.96	3.65
07/31/1998	-1.06	0.40	1	-1.46	-1.46	0.10	-1.75	-0.30	-2.15
08/31/1998	-14.46	0.43	1	-14.89	-14.89	-15.79	-13.44	-16.22	-13.87
09/30/1998	6.41	0.46	0	5.95	0	5.00	4.86	4.54	4.40
10/31/1998	8.13	0.32	0	7.81	0	5.41	4.56	5.09	4.24
11/30/1998	6.06	0.31	0	5.75	0	5.19	5.56	4.88	5.25
12/31/1998	5.76	0.38	1	5.38	5.38	7.59	7.18	7.21	6.80
01/31/1999	4.18	0.35	1	3.83	3.83	2.60	3.11	2.25	2.76
02/28/1999	-3.11	0.35	1	-3.46	-3.46	-4.13	-3.01	-4.48	-3.36
03/31/1999	4.00	0.43	1	3.57	3.57	3.09	3.27	2.66	2.84
04/30/1999	3.87	0.37	1	3.50	3.5	2.26	2.22	1.89	1.85

(Continued)

Table 3 (Continued)

Month Ended	r_M	r_{ft}	Dummy D_t	$r_M - f_{ft}$ x_t $D_t x_t$		Mutual Fund			
						A r_t	B r_t	A y_t	B y_t
05/31/1999	-2.36	0.34	1	-2.70	-2.7	-2.12	-1.32	-2.46	-1.66
06/30/1999	5.55	0.40	1	5.15	5.15	4.43	5.36	4.03	4.96
07/31/1999	-3.12	0.38	1	-3.50	-3.5	-3.15	-1.72	-3.53	-2.10
08/31/1999	-0.50	0.39	0	-0.89	0	-1.05	-2.06	-1.44	-2.45
09/30/1999	-2.74	0.39	1	-3.13	-3.13	-2.86	-1.33	-3.25	-1.72
10/31/1999	6.33	0.39	0	5.94	0	5.55	2.29	5.16	1.90
11/30/1999	2.03	0.36	1	1.67	1.67	3.23	3.63	2.87	3.27
12/31/1999	5.89	0.44	1	5.45	5.45	8.48	7.09	8.04	6.65
01/31/2000	-5.02	0.41	1	-5.43	-5.43	-4.09	-0.83	-4.50	-1.24
02/29/2000	-1.89	0.43	1	-2.32	-2.32	1.43	2.97	1.00	2.54
03/31/2000	9.78	0.47	0	9.31	0	6.84	5.86	6.37	5.39
04/30/2000	-3.01	0.46	1	-3.47	-3.47	-4.04	-4.55	-4.50	-5.01
05/31/2000	-2.05	0.50	1	-2.55	-2.55	-2.87	-4.47	-3.37	-4.97
06/30/2000	2.46	0.40	1	2.06	2.06	0.54	6.06	0.14	5.66
07/31/2000	-1.56	0.48	0	-2.04	0	-0.93	1.89	-1.41	1.41
08/31/2000	6.21	0.50	0	5.71	0	7.30	6.01	6.80	5.51
09/30/2000	-5.28	0.51	1	-5.79	-5.79	-4.73	-4.81	-5.24	-5.32
10/31/2000	-0.42	0.56	0	-0.98	0	-1.92	-4.84	-2.48	-5.40
11/30/2000	-7.88	0.51	0	-8.39	0	-6.73	-11.00	-7.24	-11.51
12/31/2000	0.49	0.50	0	-0.01	0	2.61	3.69	2.11	3.19
01/31/2001	3.55	0.54	0	3.01	0	0.36	5.01	-0.18	4.47
02/28/2001	-9.12	0.38	0	-9.50	0	-5.41	-8.16	-5.79	-8.54
03/31/2001	-6.33	0.42	0	-6.75	0	-5.14	-5.81	-5.56	-6.23
04/30/2001	7.77	0.39	0	7.38	0	5.25	4.67	4.86	4.28
05/31/2001	0.67	0.32	0	0.35	0	0.47	0.45	0.15	0.13
06/30/2001	-2.43	0.28	1	-2.71	-2.71	-3.48	-1.33	-3.76	-1.61
07/31/2001	-0.98	0.30	1	-1.28	-1.28	-2.24	-1.80	-2.54	-2.10
08/31/2001	-6.26	0.31	0	-6.57	0	-4.78	-5.41	-5.09	-5.72
09/30/2001	-8.08	0.28	0	-8.36	0	-6.46	-7.27	-6.74	-7.55
10/31/2001	1.91	0.22	0	1.69	0	1.01	2.30	0.79	2.08
11/30/2001	7.67	0.17	0	7.50	0	4.49	5.62	4.32	5.45
12/31/2001	0.88	0.15	1	0.73	0.73	1.93	2.14	1.78	1.99
01/31/2002	-1.46	0.14	1	-1.60	-1.6	-0.99	-3.27	-1.13	-3.41
02/28/2002	-1.93	0.13	1	-2.06	-2.06	-0.84	-2.68	-0.97	-2.81
03/31/2002	3.76	0.13	0	3.63	0	3.38	4.70	3.25	4.57
04/30/2002	-6.06	0.15	0	-6.21	0	-4.38	-3.32	-4.53	-3.47
05/31/2002	-0.74	0.14	0	-0.88	0	-1.78	-0.81	-1.92	-0.95
06/30/2002	-7.12	0.13	0	-7.25	0	-5.92	-5.29	-6.05	-5.42
07/31/2002	-7.80	0.15	0	-7.95	0	-6.37	-7.52	-6.52	-7.67
08/31/2002	0.66	0.14	0	0.52	0	-0.06	1.86	-0.20	1.72
09/30/2002	-10.87	0.14	0	-11.01	0	-9.38	-6.04	-9.52	-6.18
10/31/2002	8.80	0.14	0	8.66	0	3.46	5.10	3.32	4.96
11/30/2002	5.89	0.12	0	5.77	0	3.81	1.73	3.69	1.61
12/31/2002	-5.88	0.11	1	-5.99	-5.99	-4.77	-2.96	-4.88	-3.07
01/31/2003	-2.62	0.10	1	-2.72	-2.72	-1.63	-2.34	-1.73	-2.44
02/28/2003	-1.50	0.09	0	-1.59	0	-0.48	-2.28	-0.57	-2.37
03/31/2003	0.97	0.10	0	0.87	0	1.11	1.60	1.01	1.50
04/30/2003	8.24	0.10	0	8.14	0	6.67	5.44	6.57	5.34
05/31/2003	5.27	0.09	1	5.18	5.18	4.96	6.65	4.87	6.56
06/30/2003	1.28	0.10	1	1.18	1.18	0.69	1.18	0.59	1.08
07/31/2003	1.76	0.07	1	1.69	1.69	1.71	3.61	1.64	3.54
08/31/2003	1.95	0.07	1	1.88	1.88	1.32	1.13	1.25	1.06

Table 3 (Continued)

Month Ended	r_M	r_{ft}	Dummy D_t	$r_M - r_{ft}$ x_t	$D_t x_t$	Mutual Fund			
						A r_t	B r_t	A y_t	B y_t
09/30/2003	-1.06	0.08	1	-1.14	-1.14	-1.34	-1.12	-1.42	-1.20
10/31/2003	5.66	0.07	1	5.59	5.59	5.30	4.21	5.23	4.14
11/30/2003	0.88	0.07	1	0.81	0.81	0.74	1.18	0.67	1.11
12/31/2003	5.24	0.08	1	5.16	5.16	4.87	4.77	4.79	4.69
01/31/2004	1.84	0.07	1	1.77	1.77	0.87	2.51	0.80	2.44
02/29/2004	1.39	0.06	1	1.33	1.33	0.97	1.18	0.91	1.12
03/31/2004	-1.51	0.09	1	-1.60	-1.6	-0.89	-1.79	-0.98	-1.88
04/30/2004	-1.57	0.08	1	-1.65	-1.65	-2.59	-1.73	-2.67	-1.81
05/31/2004	1.37	0.06	0	1.31	0	0.66	0.83	0.60	0.77
06/30/2004	1.94	0.08	0	1.86	0	1.66	1.56	1.58	1.48
07/31/2004	-3.31	0.10	1	-3.41	-3.41	-2.82	-4.26	-2.92	-4.36
08/31/2004	0.40	0.11	0	0.29	0	-0.33	0.00	-0.44	-0.11
09/30/2004	1.08	0.11	0	0.97	0	1.20	1.99	1.09	1.88
10/31/2004	1.53	0.11	0	1.42	0	0.33	1.21	0.22	1.10
11/30/2004	4.05	0.15	1	3.90	3.9	4.87	5.68	4.72	5.53
12/31/2004	3.40	0.16	1	3.24	3.24	2.62	3.43	2.46	3.27

Notes:

1. The following information is used for determining the value of the dummy variable for the first three months:

	r_m	r_f	$r_m - r_f$
Sep-94	-2.41	0.37	-2.78
Oct-94	2.29	0.38	1.91
Nov-94	-3.67	0.37	-4.04
Dec-94	1.46	0.44	1.02

2. The dummy variable is defined as follows:

$$D_t x_t = x_t \text{ if } (r_M - r_{ft}) \text{ for the prior three months } > 0$$

$$D_t x_t = 0 \text{ otherwise}$$

If β_{2i} is not statistically significant, then there is no difference in β_i for up and down markets.

In our illustration, we have to define what we mean by an up and a down market. We will

define an up market precisely as one where the average excess return (market return over the risk-free rate or $(r_{Mt} - r_{ft})$) for the prior three months is greater than zero. Then

$$D_t = 1 \text{ if the average } (r_{Mt} - r_{ft}) \text{ for the prior three months } > 0$$

$$D_t = 0 \text{ otherwise}$$

The regressor will then be

$$D_t x_t = x_t \text{ if } (r_{Mt} - r_{ft}) \text{ for the prior three months } > 0$$

$$D_t x_t = 0 \text{ otherwise}$$

The data are presented in Table 3. The fourth column provides the coding for the dummy variable, D_t , and the sixth column shows the

Table 4 Characteristic Line for Mutual Funds A and B

	Coefficient Estimate	Standard Error	t -statistic ^a	p -value
Mutual Fund A				
α	0.206	0.102	-2.014	0.046
β	0.836	0.022	37.176	0.000
R_2	0.92			
p -value	0.000			
Mutual Fund B				
α	0.010	0.140	0.073	0.942
β	0.816	0.031	26.569	0.000
R_2	0.86			
p -value	0.000			

^aNull hypothesis is that β is equal to zero.

Table 5 Regression Results for Dummy Variable Regression for Mutual Funds A and B

	Coefficient	Standard		
Coefficient	Estimate	Error	<i>t</i> -statistic	<i>p</i> -value
Fund A				
α	-0.23	0.10	-2.36	0.0198
β_1	0.75	0.03	25.83	4E-50
β_2	0.18	0.04	4.29	4E-05
Fund B				
α	0.00	0.14	-0.03	0.9762
β_1	0.75	0.04	18.02	2E-35
β_2	0.13	0.06	2.14	0.0344

product of D_t and x_t . The regression results for the two mutual funds are shown in Table 5. The adjusted R^2 is 0.93 and 0.83 for mutual funds A and B, respectively.

For both funds, β_{2i} is statistically significantly different from zero. Hence, for these two mutual funds, there is a difference in the β_i for up and down markets. From the results reported previously, we would find that:

	Mutual Fund A	Mutual Fund B
Down market $\beta_i (= \beta_{1i})$	0.75	0.75
Up market $\beta_i (= \beta_{1i} + \beta_{2i})$	0.93 (= 0.75 + 0.18)	0.88 (= 0.75 + 0.13)

DEPENDENT CATEGORICAL VARIABLES

Thus far we have discussed models where the independent variables can be either quantitative or categorical while the dependent variable is quantitative. Let's now discuss models where the *dependent variable* is categorical. Recall that a regression model can be interpreted as a conditional probability distribution. Suppose that the dependent variable is a categorical variable Y that can assume two values, which we represent conventionally as 0 and 1. The probability distribution of the dependent variable is then a discrete function:

$$\begin{cases} P(Y = 1) = p \\ P(Y = 0) = q = 1 - p \end{cases}$$

A regression model where the dependent variable is a categorical variable is therefore a probability model; that is, it is a model of the probability p given the values of the independent variables \mathbf{X} :

$$P(Y = 1|\mathbf{X}) = f(\mathbf{X})$$

In the following sections we will discuss three probability models: the linear probability model, the probit regression model, and the logit regression model.

Linear Probability Model

The *linear probability model* assumes that the function $f(\mathbf{X})$ is linear. For example, a linear probability model of default assumes that there is a linear relationship between the probability of default and the factors that determine default.

$$P(Y = 1|\mathbf{X}) = f(\mathbf{X})$$

The parameters of the model can be obtained by using ordinary least squares applying the estimation methods of multiple regression models entry. Once the parameters of the model are estimated, the predicted value for $P(Y)$ can be interpreted as the event probability such as the probability of default in our previous example. Note, however, that when using a linear probability model, in this entry the R^2 is used only if all the independent variables are also binary variables.

A major drawback of the linear probability model is that the predicted value may be negative. In the probit regression and logit regression models described below, the predicted probability is forced to be between 0 and 1.

Probit Regression Model

The *probit regression model* is a nonlinear regression model where the dependent variable is a binary variable. Due to its nonlinearity, one cannot estimate this model with least squares methods. We have to use maximum likelihood (ML) methods as described below.

Because what is being predicted is the standard normal cumulative probability distribution, the predicted values are between 0 and 1.

The general form for the probit regression model is

$$P(Y = 1 | X_1, X_2, \dots, X_K) = N(a + b_1 X_1 + b_2 X_2 + \dots + b_K X_K)$$

where N is the cumulative standard normal distribution function.

To see how ML methods work, consider a model of the probability of corporate bond defaults. Suppose that there are three factors that have been found to historically explain corporate bond defaults. The probit regression model is then

$$\begin{cases} P(Y = 1 | X_1, X_2, X_3) \\ = N(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3) \\ P(Y = 0 | X_1, X_2, X_3) \\ = 1 - N(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3) \end{cases}$$

The likelihood function is formed from the products

$$\prod_i N(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i})^{Y_i} (1 - N(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}))^{1 - Y_i}$$

extended to all the samples, where the variable Y assumes a value of 0 for defaulted companies and 1 for nondefaulted companies. Parameters are estimated by maximizing the likelihood.

Suppose that the following parameters are estimated:

$$\beta = -2.1 \quad \beta_1 = 1.9 \quad \beta_2 = 0.3 \quad \beta_3 = 0.8$$

Then

$$N(a + b_1 X_1 + b_2 X_2 + b_3 X_3) = N(-2.1 + 1.9X_1 + 0.3X_2 + 0.8X_3)$$

Now suppose that the probability of default of a company with the following values for the

independent variables is sought:

$$X_1 = 0.2 \quad X_2 = 0.9 \quad X_3 = 1.0$$

Substituting these values we get

$$N(-2.1 + 1.9(0.2) + 0.3(0.9) + 0.8(1.0)) = N(-0.65)$$

The standard normal cumulative probability for $N(-0.65)$ is 25.8%. Therefore, the probability of default for a company with this characteristic is 25.8%.

Application to Hedge Fund Survival

An illustration of the probit regression model is provided by Malkiel and Saha (2005) who use it to calculate the probability of the demise of a hedge fund. The dependent variable in the regression is 1 if a fund is defunct (did not survive) and 0 if it survived. The explanatory variables, their estimated coefficient, and the standard error of the coefficient using hedge fund data from 1994 to 2003 are given below:

Explanatory Variable	Coefficient	Standard Deviation
1. Return for the first quarter before the end of fund performance	-1.47	0.36
2. Return for the second quarter before the end of fund performance	-4.93	0.32
3. Return for the third quarter before the end of fund performance	-2.74	0.33
4. Return for the fourth quarter before the end of fund performance	-3.71	0.35
5. Standard deviation for the year prior to the end of fund performance	17.76	0.92
6. Number of times in the final three months the fund's monthly return fell below the monthly median of all funds in the same primary category	0.00	0.33
7. Assets of the fund (in billions of dollars) estimated at the end of performance	-1.30	-7.76
Constant term	-0.37	0.07

For only one explanatory variable, the sixth one, the coefficient is not statistically significant from zero. That explanatory variable is a proxy for peer comparison of the hedge fund versus similar hedge funds. The results suggest that there is a lower probability of the demise of a hedge fund if there is good recent performance (the negative coefficient of the first four variables above) and the more assets under management (the negative coefficient for the last variable above). The greater the hedge fund performance return variability, the higher the probability of demise (the positive coefficient for the fifth variable above).

Logit Regression Model

As with the probit regression model, the *logit regression model* is a nonlinear regression model where the dependent variable is a binary variable and the predicted values are between 0 and 1. The predicted value is also a cumulative probability distribution. However, rather than being a standard normal cumulative probability distribution, it is a standard cumulative probability distribution of a distribution called the *logistic distribution*.

The general formula for the logit regression model is

$$\begin{aligned} P(Y = 1 | X_1, X_2, \dots, X_N) \\ &= F(a + b_1 X_1 + b_2 X_2 + \dots + b_N X_N) \\ &= -1/1 + e^{-W} \end{aligned}$$

where $W = a + b_1 X_1 + b_2 X_2 + \dots + b_N X_N$.

As with the probit regression model, the logit regression model is estimated with ML methods.

Using our previous illustration, $W = -0.65$. Therefore

$$1/[1 + e^{-W}] = 1/[1 + e^{-(-0.65)}] = 34.3\%$$

The probability of default for the company with these characteristics is 34.3%.

KEY POINTS

- Categorical variables are variables that represent group membership and can appear in a regression equation as a regressor or as an independent variable.
- A dichotomous variable is an explanatory variable that distinguishes only two categories; the key is to represent a dichotomous categorical variable as a numerical variable, referred to as a dummy variable, that can assume the two values 0,1.
- When a dummy variable is a regressor, the *t*-statistic can be used to determine if that variable is statistically significant. The Chow test can also be used to test if all the dummy variables in a regression model are collectively relevant.
- A regression model where the dependent variable is a categorical variable is a probability model, and there are three types of such models: the probability model, the probit regression model, and the logit regression model.
- The linear probability model assumes that the probability model to be estimated is linear and can be estimated using least squares.
- The probit regression model is a nonlinear regression model where the dependent variable is a binary variable. The model cannot be estimated using least squares because it is a nonlinear model and is instead estimated using maximum likelihood methods.
- The logit regression model is a nonlinear regression model where the dependent variable is a binary variable and the predicted values are between 0 and 1 and represent a cumulative probability distribution. Rather than being a standard normal cumulative probability distribution, it is a standard cumulative probability of a logit.

NOTE

1. The model presented in this illustration was developed by FridsonVision and is described in "Focus Issues Methodology," *Leverage World* (May 30, 2003). The data for this illustration were provided by Greg Braylovskiy of FridsonVision. The firm uses about 650 companies in its analysis. Only 100 observations were used in this illustration.

REFERENCES

- Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica* 28: 591–605.
- Fabozzi, F. J. (2006). *Bond Markets, Analysis, and Strategies*, 6th ed. Upper Saddle River, NJ: Prentice Hall.
- Malkiel, B., and Saha, A. (2005). Hedge funds: Risk and return. *Financial Analysts Journal* 22: 80–88.

Quantile Regression

CHRIS GOWLLAND, CFA

Senior Quantitative Analyst, Delaware Investments

Abstract: Many of the statistical methods that are most commonly used by researchers and practitioners in finance are mainly focused on identifying the central tendency within a data set. However, there are numerous situations where it may be equally or more important to understand the dispersion between outcomes that are higher or lower than the central tendency. One statistical method that can be useful in such investigations is quantile regression, which conceptually can be viewed as a logical extension of ordinary least squares methods.

Many investors use regression methods to gauge the relative attractiveness of different firms, the risks inherent in active or passive portfolios, the historical performance of investment factors, and similar topics. Such research often focuses on understanding the “central tendency” within a data set, and for this purpose perhaps the most commonly used tool is regression based on *ordinary least squares* (OLS) approaches. OLS methods are designed to find the “line of best fit” by minimizing the sum of squared errors from individual data points. OLS analysis generally does a good job of describing the *central tendency* within a data set, but typically will be much less effective at describing the behavior of data points that are distant from the line of best fit. *Quantile regressions*, however, can be useful in such investigations. This statistical approach can be viewed conceptually as a logical extension of ordinary least squares methods. We present a brief

overview of quantile regression approaches, together with some examples of how such methods can be applied in practical situations.

COMPARING QUANTILE AND OLS APPROACHES

Conceptually, OLS *statistical analysis* can be summarized by the following equation, as expressed in a univariate context where a single independent variable is being used to explain or predict a single dependent variable:

$$Y_i = \alpha + \sum_{i=1}^N \beta X_i + \varepsilon$$

where Y represents the dependent variable, X represents the observed value of an independent variable, $i = 1, \dots, N$ data points, α represents the intercept (in other words, the value on the vertical axis when the horizontal axis is

The material discussed here does not necessarily represent the opinions, methods, or views of Delaware Investments.

zero), β represents the slope of the relationship between X and Y , and ε is an error term with an expected mean value of zero.

As a hypothetical example, suppose that X reflects the expected dividend in dollars for a universe of firms, and Y represents the stock price for each of those firms. Then the value of β will reflect the value that the market is assigning to each \$1 of dividend payment, while the value of α will reflect the expected price of a stock that does not pay a dividend. (Please note that we are not proposing that such an equation would provide a usable investment thesis.) It is possible to adapt this simple OLS equation to a multivariate context, in which several different independent variables are being used together to explain or predict the value of the dependent variable.

Similarly, quantile regression approaches can be summarized by the following equation, again in a univariate context:

$$Y_i = \alpha^p + \sum_{i=1}^N \beta^p X_i + \varepsilon^p$$

where α^p represents the intercept for a specified quantile, β^p represents the slope of the relationship between X and Y for a specified quantile, and ε^p similarly represents the error term for that specified quantile. (The specific form for these two equations has been adapted from Meligkotsidou, Vrontos, and Vrontos, 2007; other authors might use different terminology, but the underlying concepts are the same.) And just as OLS methods can be used in both univariate and multivariate contexts, the same is true for quantile regression approaches.

In this context, what is a quantile? It is a generalized form of a percentile, in other words a measure of spread between the highest and lowest values in a particular range. A quantile can conveniently be expressed in terms of percentages, so that the median will be the 50th quantile. But the same method can be used for any quantile, not just the 50th quantile. In this sense, quantile methods are somewhat similar

to value-at-risk (VaR) approaches, which seek to measure the “95th percentile” or “99th percentile” of potential losses in a portfolio.

REASONS FOR USING QUANTILE METHODS

If a data set is distributed in an approximately normal fashion, and if the analysis focuses specifically on the 50th quantile, then the results will often be quite similar to those derived from conventional OLS analysis. However, OLS methods tend to provide unreliable results if a data set is skewed, has “fat tails,” or has some extreme *outliers*—any or all of which can exist when the relevant data are drawn from economics or finance (Koenker and Hallock, 2001). In such circumstances, quantile regression focusing on the 50th quantile will often provide a more robust estimate of the central tendency than would be available from OLS approaches. Figure 1 provides a hypothetical example of a situation where quantile regression might be useful.

Figure 1 shows a scatter plot of a hypothetical relationship that has three main traits: (1) positive slope, (2) higher dispersion of results when the independent variable is small, and (3) a single outlier toward the top end of the range. The graph shows that the outlier exerts considerable influence on the OLS analysis by tending to skew the relationship upward. A conventional OLS approach might decide to exclude the outlier, but this would effectively mean throwing away the information contained in that data point. By contrast, the quantile analysis includes the outlier, but is less affected by its presence. As a consequence, quantile regression does a better job of identifying the “central tendency” within this data set—in exactly the same way as an analyst might choose to use the *median* rather than the mean when describing a distribution that has a heavy weight in the left or right tail.

The above analysis shows that quantile regression is more robust than OLS methods in

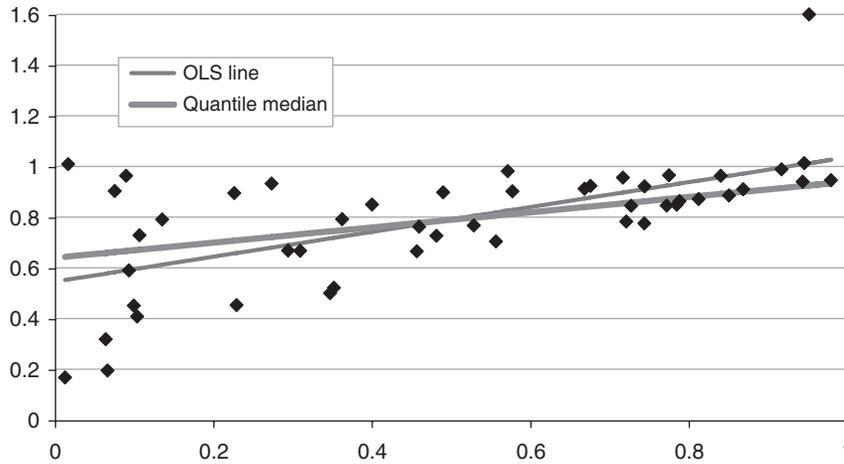


Figure 1 Effect of Outliers on OLS and Quantile Analysis
Note: Data are hypothetical and based on a simulated relationship.

the presence of outliers and other potentially distorting influences. Another useful feature of quantile approaches is that they allow analysis of areas away from the middle of the distribution. Conventional regression techniques focus on the “central tendency” of the data, and thus tend to prioritize describing the relationship that is most representative of the average. However, from the perspective of an active investor or a risk manager, the most interesting information may well be in the tails of the dis-

tribution, where the standard OLS approaches are not generally very informative, but where quantile methods can be readily applied.

Figure 2 shows the same scatter plot as Figure 1, but instead of showing the quantile median, it shows estimated lines for the 10th and 90th quantiles. The lines form a funnel-like shape, indicating that there is greater variation on the left of the distribution than the right. From the perspective of an investor, this suggests that the range of possible outcomes from

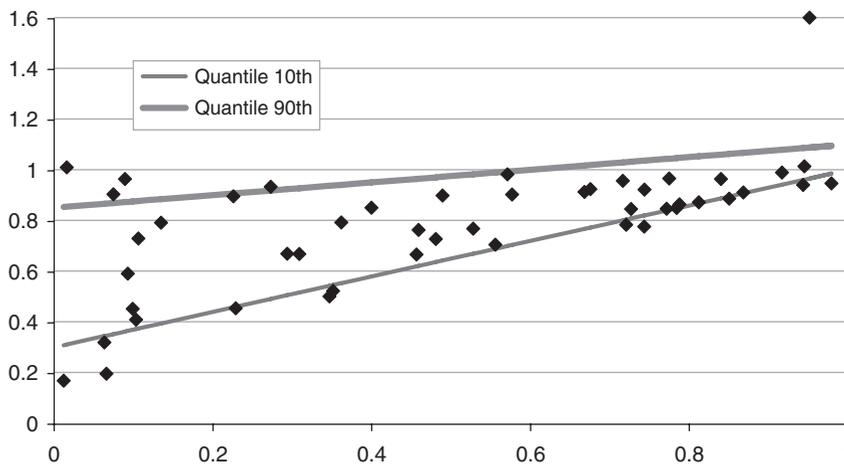


Figure 2 Effect of Outliers on OLS and Quantile Analysis: Estimated Lines for the 10th and 90th Quantiles
Note: Data are hypothetical and based on a simulated relationship.

investing in companies on the left of the distribution may be larger, and thus require more careful analysis. From the perspective of a risk manager, the difference in slope between the 10th and 90th percentiles might suggest that greater provisioning would be appropriate if a portfolio tends to have greater weight in the left of the distribution. Once again, the outlier is included in the analysis, but its impact on the estimated intercept and slope for the 10th and 90th percentiles is considerably muted by comparison with what would be expected using OLS-like methods.

BACKGROUND AND FURTHER EXAMPLES

Quantile regression methods were first developed in the 1970s in the discipline of statistics (Koenker and Bassett, 1978). Koenker (2005) provides a comprehensive overview of quantile regression in general, with numerous examples drawn from finance and from other subject areas. The statistical packages R, S-Plus, Stata, SAS, and SPSS all have quantile regression capabilities, either as part of their base distribution or as separate modules. These packages typically focus on linear quantile regression, but extensions to nonlinear applications are also feasible (Koenker and Hallock, 2001).

In recent years, quantile regression methods have become increasingly popular in finance and economics. Chernozhukov and Umantsev (2000) applied quantile methods to estimate VaR, noting that the basic structure could be applied to various possible modeling approaches. Wu and Xiao (2002) also used quantile methods to estimate VaR and provided an example of how such approaches could be used in the context of an index fund. Engle and Manganelli (2004) provided an example of how to use quantile regression approaches in calculating a conditional VaR measure. Kuester, Mittnik, and Paolella (2006) proposed extending the conditional VaR approach by incorporating some additional autoregressive elements.

An important area of research for academics and practitioners has been the influence of investment style on portfolio returns. One way to perform such analysis is through the analysis of portfolio holdings, but these are typically only available periodically and with a considerable lag. Another approach has been to focus on portfolio returns, which may be available at higher frequency and with a smaller delay. Early work in this area, such as Sharpe (1992) and Carhart (1997), generally relied on OLS approaches, which led to a focus on a portfolio's "central tendency" relative to its benchmark. Bassett and Chen (2001) extended this earlier work by applying quantile methods, and showed that this permits examination of active performance during periods when the portfolio and/or its benchmark are far away from their central tendency.

As shown above, quantile regression provides a more complete picture than OLS approaches of the conditional relationship among financial variables. Landajo, de Andrés, and Lorca (2008) used quantile methods to gauge the relationship between size and profitability for publishing firms in Spain, and showed that the patterns for small firms were rather different from those for their larger peers. Similarly, Lee, Chan, Yeh, and Chan (2010) used quantile methods on a sample of firms from Taiwan in order to assess how increasing internationalization affects relative valuation.

Quantile methods can also be used to test whether the quantile-specific parameters are stable over different quantiles and over time, as noted by Koenker and Xiao (2006). Quantile models can thus demonstrate how different variables affect the location, scale, and shape of the conditional distribution of the response. Such methods therefore constitute a significant extension of classical constant coefficient time series models, in which the effect of conditioning is typically confined to a shift of the intercept and/or the slope of the central tendency. Fatouh, Scaramozzino, and Harris (2005) used quantile methods to analyze how the capital structure of firms in Korea had changed

over time. Billett and Xue (2008) used quantile approaches to analyze the motivations behind open market share repurchases, and found that firms are generally more likely to repurchase shares when they are at higher risk of being taken over. Pires, Pereira, and Martins (2010) use quantile methods to analyze the determinants of credit default swap spreads over time, and report that some previously reported anomalous results may have occurred due to the emphasis on the conditional mean of the distribution, rather than on the upper and lower tails.

KEY POINTS

- Quantile regression methods are well established in the statistical literature, and are increasingly being used in finance.
- Quantile regression methods are more robust than conventional OLS approaches to skewed distributions, fat tails, and the presence of outliers—all of which are frequently encountered in real-world financial data.
- Quantile regression approaches can be used to assess the central tendency of a data set, and in this sense can be viewed as a regression-based analogue of the median of a distribution. The same approaches can also be used to examine the upper or lower reaches of a data set, which is not possible using conventional OLS methods.
- For active investors and risk managers, the upper or lower tails of a distribution may well be more interesting than the central tendency, and quantile regression is an appropriate tool for such work.
- Quantile regression methods can be applied to data from a single period, but can also be applied in a time-series context. Such methods can help in analyzing how relationships may have changed over time.

REFERENCES

Bassett, G. W., Jr., and Chen, H.-L. (2001). Portfolio style: Return-based attribution using quan-

- tile regression. *Empirical Economics* 26, 1: 293–305.
- Billett, M. T., and Xue, H. (2008). The takeover deterrent effect of open market share repurchases. *Journal of Finance* 62, 4: 1827–1850.
- Carhart, M. M. (1997). On persistence in mutual fund performance. *Journal of Finance* 52, 1: 57–82.
- Chernozhukov, V., and Umantsev, L. (2001). Conditional value-at-risk: Aspects of modeling and estimation. *Empirical Economics* 26, 1: 271–292.
- Engle, R. F., and Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics* 22, 4: 367–381.
- Fattouh, B., Scaramozzino, P., and Harris, L. Capital structure in South Korea: A quantile regression approach. *Journal of Development Economics* 76, 1: 231–250.
- Koenker, R. (2005). *Quantile Regression*. New York: Cambridge University Press.
- Koenker, R., and Bassett, G. W., Jr. (1978). Regression quantiles. *Econometrica* 46, 1: 33–50.
- Koenker, R., and Hallock, K. (2001). Quantile regression. *Journal of Economic Perspectives* 15, 4: 143–156.
- Koenker, R., and Xiao, Z. (2006). Quantile autoregression. *Journal of the American Statistical Association* 101, 475: 980–990.
- Kuester, K., Mittnik, S., and Paolella, M. (2006). Value-at-Risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics* 4, 1: 53–89.
- Landajo, M., de Andres, J., and Lorca, P. (2008). Measuring firm performance by using linear and non-parametric quantile regressions. *Applied Statistics* 57, 2: 227–250.
- Lee, T., Chan, K., Yeh, J.-H., and Chan, H.-Y. (2010). The impact of internationalization on firm performance: A quantile regression analysis. *International Review of Accounting, Banking and Finance* 2, 4: 39–59.
- Meligkotsidou, L., Vrontos, I. D., and Vrontos, S. D. (2009). Quantile analysis of hedge fund strategies. *Journal of Empirical Finance* 16, 2: 264–279.
- Pires, P., Pereira, J. P., and Martins, L. F. (2010). The complete picture of credit default swap spreads: A quantile regression approach. ISCTE Business School working paper.
- Sharpe, W.F. (1992). Asset allocation: Management style and performance measurement. *Journal of Portfolio Management* 18, 2: 7–19.
- Wu, G., and Xiao, Z. (2002). An analysis of risk measures. *Journal of Risk* 4, 4: 53–76.

ARCH/GARCH Models in Applied Financial Econometrics

ROBERT F. ENGLE, PhD

Michael Armellino Professorship in the Management of Financial Services and Director of the Volatility Institute, Leonard N. Stern School of Business, New York University

SERGIO M. FOCARDI, PhD

Partner, The Intertek Group

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Volatility is a key parameter used in many financial applications, from derivatives valuation to asset management and risk management. Volatility measures the size of the errors made in modeling returns and other financial variables. It was discovered that, for vast classes of models, the average size of volatility is not constant but changes with time and is predictable. Autoregressive conditional heteroskedasticity (ARCH), generalized autoregressive conditional heteroskedasticity (GARCH) models, and stochastic volatility models are the main tools used to model and forecast volatility. Moving from single assets to portfolios made of multiple assets, not only are there idiosyncratic volatilities but also correlations and covariances between assets that are time varying and predictable. Multivariate ARCH/GARCH models and dynamic factor models, eventually in a Bayesian framework, are the basic tools used to forecast correlations and covariances.

In this entry we discuss the modeling of the time behavior of the uncertainty related to many econometric models when applied to financial data. Finance practitioners know that errors made in predicting markets are not of a constant magnitude. There are periods when unpredictable market fluctuations are larger and periods when they are smaller. This behavior, known as *heteroskedasticity*, refers to the fact that the size of market *volatility* tends to cluster in periods of high volatility and periods of

low volatility. The discovery that it is possible to formalize and generalize this observation was a major breakthrough in econometrics. In fact, we can describe many economic and financial data with models that predict, simultaneously, the economic variables and the average magnitude of the squared prediction error.

In this entry, we show how the average error size can be modeled as an autoregressive process. Given their autoregressive nature, these

models are called *autoregressive conditional heteroskedasticity (ARCH)* or *generalized autoregressive conditional heteroskedasticity (GARCH)*. This discovery is particularly important in financial econometrics, where the error size is, in itself, a variable of great interest.

REVIEW OF LINEAR REGRESSION AND AUTOREGRESSIVE MODELS

Let's first discuss two examples of basic econometric models, the *linear regression model* and the *autoregressive model*, and illustrate the meaning of homoskedasticity or heteroskedasticity in each case.

The linear regression model is the workhorse of economic modeling. A univariate linear regression represents a proportionality relationship between two variables:

$$y = \alpha + \beta x + \varepsilon$$

The preceding linear regression model states that the expectation of the variable y is β times the expectation of the variable x plus a constant α . The proportionality relationship between y and x is not exact but subject to an error ε .

In standard regression theory, the error ε is assumed to have a zero mean and a constant standard deviation σ . The standard deviation is the square root of the variance, which is the expectation of the squared error: $\sigma^2 = E(\varepsilon^2)$. It is a positive number that measures the size of the error. We call *homoskedasticity* the assumption that the expected size of the error is constant and does not depend on the size of the variable x . We call *heteroskedasticity* the assumption that the expected size of the error term is not constant.

The assumption of homoskedasticity is convenient from a mathematical point of view and is standard in regression theory. However, it is an assumption that must be verified empirically. In many cases, especially if the range of variables is large, the assumption of homo-

skedasticity might be unreasonable. For example, assuming a linear relationship between consumption and household income, we can expect that the size of the error depends on the size of household income. In fact, high-income households have more freedom in the allocation of their income.

In the preceding household-income example, the linear regression represents a cross-sectional model without any time dimension. However, in finance and economics in general, we deal primarily with time series, that is, sequences of observations at different moments of time. Let's call X_t the value of an economic time series at time t . Since the groundbreaking work of Haavelmo (1944), economic time series are considered to be realizations of stochastic processes. That is, each point of an economic time series is considered to be an observation of a random variable.

We can look at a stochastic process as a sequence of variables characterized by joint-probability distributions for every finite set of different time points. In particular, we can consider the distribution f_t of each variable X_t at each moment. Intuitively, we can visualize a stochastic process as a very large (infinite) number of paths. A process is called weakly stationary if all of its second moments are constant. In particular this means that the mean and variance are constants $\mu_t = \mu$ and $\sigma_t^2 = \sigma^2$ that do not depend on the time t . A process is called strictly stationary if none of its finite distributions depends on time. A strictly stationary process is not necessarily weakly stationary as its finite distributions, though time-independent, might have infinite moments.

The terms μ_t and σ_t^2 are the unconditional mean and variance of a process. In finance and economics, however, we are typically interested in making forecasts based on past and present information. Therefore, we consider the distribution $f_{t_2}(x | I_{t_1})$ of the variable X_{t_2} at time t_2 conditional on the information I_{t_1} known at time t_1 . Based on information available at time $t - 1$, I_{t-1} , we can also define the

conditional mean and the conditional variance $(\mu_t | I_{t-1})$, $(\sigma_t^2 | I_{t-1})$.

A process can be weakly stationary but have time-varying conditional variance. If the conditional mean is constant, then the unconditional variance is the unconditional expectation of the conditional variance. If the conditional mean is not constant, the unconditional variance is not equal to the unconditional expectation of the conditional variance; this is due to the dynamics of the conditional mean.

In describing ARCH/GARCH behavior, we focus on the error process. In particular, we assume that the errors are an innovation process, that is, we assume that the conditional mean of the errors is zero. We write the error process as: $\varepsilon_t = \sigma_t z_t$ where σ_t is the conditional standard deviation and the z terms are a sequence of independent, zero-mean, unit-variance, normally distributed variables. Under this assumption, the unconditional variance of the error process is the unconditional mean of the conditional variance. Note, however, that the unconditional variance of the process variable does not, in general, coincide with the unconditional variance of the error terms.

In financial and economic models, conditioning is often stated as regressions of the future values of the variables on the present and past values of the same variable. For example, if we assume that time is discrete, we can express conditioning as an autoregressive model:

$$X_{t+1} = \alpha_0 + \beta_0 X_t + \dots + \beta_n X_{t-n} + \varepsilon_{t+1}$$

The error term ε_i is conditional on the information I_i that, in this example, is represented by the present and the past n values of the variable X . The simplest autoregressive model is the random walk model of the logarithms of prices p_i :

$$p_{t+1} = \mu t + p_t + \varepsilon_t$$

In terms of returns, the random walk model is simply:

$$r_t = \Delta p_t = \mu + \varepsilon_t$$

A major breakthrough in econometric modeling was the discovery that, for many families of econometric models, linear and nonlinear alike, it is possible to specify a stochastic process for the error terms and predict the average size of the error terms when models are fitted to empirical data. This is the essence of ARCH modeling introduced by Engle (1982).

Two observations are in order. First, we have introduced two different types of heteroskedasticity. In the first example, regression errors are heteroskedastic because they depend on the value of the independent variables: The average error is larger when the independent variable is larger. In the second example, however, error terms are conditionally heteroskedastic because they vary with time and do not necessarily depend on the value of the process variables. Later in this entry we will describe a variant of the ARCH model where the size of volatility is correlated with the level of the variable. However, in the basic specification of ARCH models, the level of the variables and the size of volatility are independent.

Second, let's observe that the volatility (or the variance) of the error term is a hidden, nonobservable variable. Later in this entry, we will describe realized volatility models that treat volatility as an observed variable. Theoretically, however, time-varying volatility can be only inferred, not observed. As a consequence, the error term cannot be separated from the rest of the model. This occurs both because we have only one realization of the relevant time series and because the volatility term depends on the model used to forecast expected returns. The ARCH/GARCH behavior of the error term depends on the model chosen to represent the data. We might use different models to represent data with different levels of accuracy. Each model will be characterized by a different specification of heteroskedasticity.

Consider, for example, the following model for returns:

$$r_t = m + \varepsilon_t$$

In this simple model, the clustering of volatility is equivalent to the clustering of the squared returns (minus their constant mean). Now suppose that we discover that returns are predictable through a regression on some predictor f :

$$r_t = m + f_{t-1} + \varepsilon_t$$

As a result of our discovery, we can expect that the model will be more accurate, the size of the errors will decrease, and the heteroskedastic behavior will change.

Note that in the model $r_t = m + \varepsilon_t$, the errors coincide with the fluctuations of returns around their unconditional mean. If errors are an innovation process, that is, if the conditional mean of the errors is zero, then the variance of returns coincides with the variance of errors, and ARCH behavior describes the fluctuations of returns. However, if we were able to make conditional forecasts of returns, then the ARCH model describes the behavior of the errors and it is no longer true that the unconditional variance of errors coincides with the unconditional variance of returns. Thus, the statement that ARCH models describe the time evolution of the variance of returns is true only if returns have a constant expectation.

ARCH/GARCH effects are important because they are very general. It has been found empirically that most model families presently in use in econometrics and financial econometrics exhibit conditionally heteroskedastic errors when applied to empirical economic and financial data. The heteroskedasticity of errors has not disappeared with the adoption of more sophisticated models of financial variables. The ARCH/GARCH specification of errors allows one to estimate models more accurately and to forecast volatility.

ARCH/GARCH MODELS

In this section, we discuss univariate ARCH and GARCH models. Because in this entry we focus on financial applications, we will use finan-

cial notation. Let the dependent variable, which might be the return on an asset or a portfolio, be labeled r_t . The mean value m and the variance h will be defined relative to a past information set. Then the return r in the present will be equal to the conditional mean value of r (that is, the expected value of r based on past information) plus the conditional standard deviation of r (that is, the square root of the variance) times the error term for the present period:

$$r_t = m_t + \sqrt{h_t}z_t$$

The econometric challenge is to specify how the information is used to forecast the mean and variance of the return conditional on the past information. While many specifications have been considered for the mean return and used in efforts to forecast future returns, rather simple specifications have proven surprisingly successful in predicting conditional variances.

First, note that if the error terms were strict white noise (that is, zero-mean, independent variables with the same variance), the conditional variance of the error terms would be constant and equal to the unconditional variance of errors. We would be able to estimate the error variance with the empirical variance:

$$h = \frac{\sum_{i=1}^n \varepsilon_i^2}{n}$$

using the largest possible available sample. However, it was discovered that the residuals of most models used in financial econometrics exhibit a structure that includes heteroskedasticity and autocorrelation of their absolute values or of their squared values.

The simplest strategy to capture the time dependency of the variance is to use a short rolling window for estimates. In fact, before ARCH, the primary descriptive tool to capture time-varying conditional standard deviation and conditional variance was the rolling standard deviation or the rolling variance. This is the standard deviation or variance calculated using a fixed number of the most recent

observations. For example, a rolling standard deviation or variance could be calculated every day using the most recent month (22 business days) of data. It is convenient to think of this formulation as the first ARCH model; it assumes that the variance of tomorrow's return is an equally weighted average of the squared residuals of the last 22 days.

The idea behind the use of a rolling window is that the variance changes slowly over time, and it is therefore approximately constant on a short rolling-time window. However, given that the variance changes over time, the assumption of equal weights seems unattractive: It is reasonable to consider that more recent events are more relevant and should therefore have higher weights. The assumption of zero weights for observations more than one month old is also unappealing.

In the ARCH model proposed by Engle (1982), these weights are parameters to be estimated. Engle's ARCH model thereby allows the data to determine the best weights to use in forecasting the variance. In the original formulation of the ARCH model, the variance is forecasted as a moving average of past error terms:

$$h_t = \omega + \sum_{i=1}^p \alpha_i \varepsilon_{t-i}^2$$

where the coefficients α_i must be estimated from empirical data. The errors themselves will have the form

$$\varepsilon_t = \sqrt{h_t} z_t$$

where the z terms are independent, standard normal variables (that is, zero-mean, unit-variance, normal variables). In order to ensure that the variance is nonnegative, the constants (ω, α_i) must be nonnegative. If $\sum_{i=1}^p \alpha_i < 1$, the ARCH process is weakly stationary with constant unconditional variance:

$$\sigma^2 = \frac{\omega}{1 - \sum_{i=1}^p \alpha_i}$$

Two remarks should be made. First, ARCH is a forecasting model insofar as it forecasts the error variance at time t on the basis of information known at time $t - 1$. Second, forecasting is conditionally deterministic, that is, the ARCH model does not leave any uncertainty on the expectation of the squared error at time t knowing past errors. This must always be true of a forecast, but, of course, the squared error that occurs can deviate widely from this forecast value.

A useful generalization of this model is the GARCH parameterization introduced by Bollerslev (1986). This model is also a weighted average of past squared residuals, but it has declining weights that never go completely to zero. In its most general form, it is not a Markovian model, as all past errors contribute to forecast volatility. It gives parsimonious models that are easy to estimate and, even in its simplest form, has proven surprisingly successful in predicting conditional variances.

The most widely used GARCH specification asserts that the best predictor of the variance in the next period is a weighted average of the long-run average variance, the variance predicted for this period, and the new information in this period that is captured by the most recent squared residual. Such an updating rule is a simple description of adaptive or learning behavior and can be thought of as Bayesian updating. Consider the trader who knows that the long-run average daily standard deviation of the Standard and Poor's 500 is 1%, that the forecast he made yesterday was 2%, and the unexpected return observed today is 3%. Obviously, this is a high-volatility period, and today is especially volatile, suggesting that the volatility forecast for tomorrow could be even higher. However, the fact that the long-term average is only 1% might lead the forecaster to lower his forecast. The best strategy depends on the dependence between days. If these three numbers are each squared and weighted equally, then the new forecast would be $2.16 = \sqrt{(1 + 4 + 9)}/3$. However, rather than weighting these equally, for daily data it is

generally found that weights such as those in the empirical example of (0.02, 0.9, 0.08) are much more accurate. Hence, the forecast is $2.08 = \sqrt{0.02 \times 1 + 0.9 \times 4 + 0.08 \times 9}$. To be precise, we can use h_t to define the variance of the residuals of a regression $r_t = m_t + \sqrt{h_t}\varepsilon_t$. In this definition, the variance of ε_t is one. Therefore, a GARCH(1,1) model for variance looks like this:

$$h_{t+1} = \omega + \alpha (r_t - m_t)^2 + \beta h_t = \omega + \alpha h_t \varepsilon_t^2 + \beta h_t$$

This model forecasts the variance of date t return as a weighted average of a constant, yesterday's forecast, and yesterday's squared error. If we apply the previous formula recursively, we obtain an infinite weighted moving average. Note that the weighting coefficients are different from those of a standard exponentially weighted moving average (EWMA). The econometrician must estimate the constants ω, α, β ; updating simply requires knowing the previous forecast h and the residual.

The weights are $(1 - \alpha - \beta, \beta, \alpha)$ and the long-run average variance is $\sqrt{\omega / (1 - \alpha - \beta)}$. It should be noted that this works only if $\alpha + \beta < 1$ and it really makes sense only if the weights are positive, requiring $\alpha > 0, \beta > 0, \omega > 0$. In fact, the GARCH(1,1) process is weakly stationary if $\alpha + \beta < 1$. If $E[\log(\beta + \alpha z^2)] < 0$, the process is strictly stationary. The GARCH model with $\alpha + \beta = 1$ is called an *integrated GARCH* or IGARCH. It is a strictly stationary process with infinite variance.

The GARCH model described above and typically referred to as the GARCH(1,1) model derives its name from the fact that the 1,1 in parentheses is a standard notation in which the first number refers to the number of autoregressive lags (or ARCH terms) that appear in the equation and the second number refers to the number of moving average lags specified (often called the number of GARCH terms). Models with more than one lag are sometimes needed to find good variance forecasts. Although this model is directly set up to forecast for just one

period, it turns out that, based on the one-period forecast, a two-period forecast can be made. Ultimately, by repeating this step, long-horizon forecasts can be constructed. For the GARCH(1,1), the two-step forecast is a little closer to the long-run average variance than is the one-step forecast, and, ultimately, the distant-horizon forecast is the same for all time periods as long as $\alpha + \beta < 1$. This is just the unconditional variance. Thus, GARCH models are mean reverting and conditionally heteroskedastic but have a constant unconditional variance.

Let's now address the question of how the econometrician can estimate an equation like the GARCH(1,1) when the only variable on which there are data is r_t . One possibility is to use maximum likelihood by substituting h_t for σ^2 in the normal likelihood and then maximizing with respect to the parameters. GARCH estimation is implemented in commercially available software such as EViews, GAUSS, Matlab, RATS, SAS, or TSP. The process is quite straightforward: For any set of parameters ω, α, β and a starting estimate for the variance of the first observation, which is often taken to be the observed variance of the residuals, it is easy to calculate the variance forecast for the second observation. The GARCH updating formula takes the weighted average of the unconditional variance, the squared residual for the first observation, and the starting variance and estimates the variance of the second observation. This is input into the forecast of the third variance, and so forth. Eventually, an entire time series of variance forecasts is constructed.

Ideally, this series is large when the residuals are large and small when the residuals are small. The likelihood function provides a systematic way to adjust the parameters ω, α, β to give the best fit. Of course, it is possible that the true variance process is different from the one specified by the econometrician. In order to check this, a variety of diagnostic tests are available. The simplest is to construct the series of $\{\varepsilon_t\}$, which are supposed to have constant mean

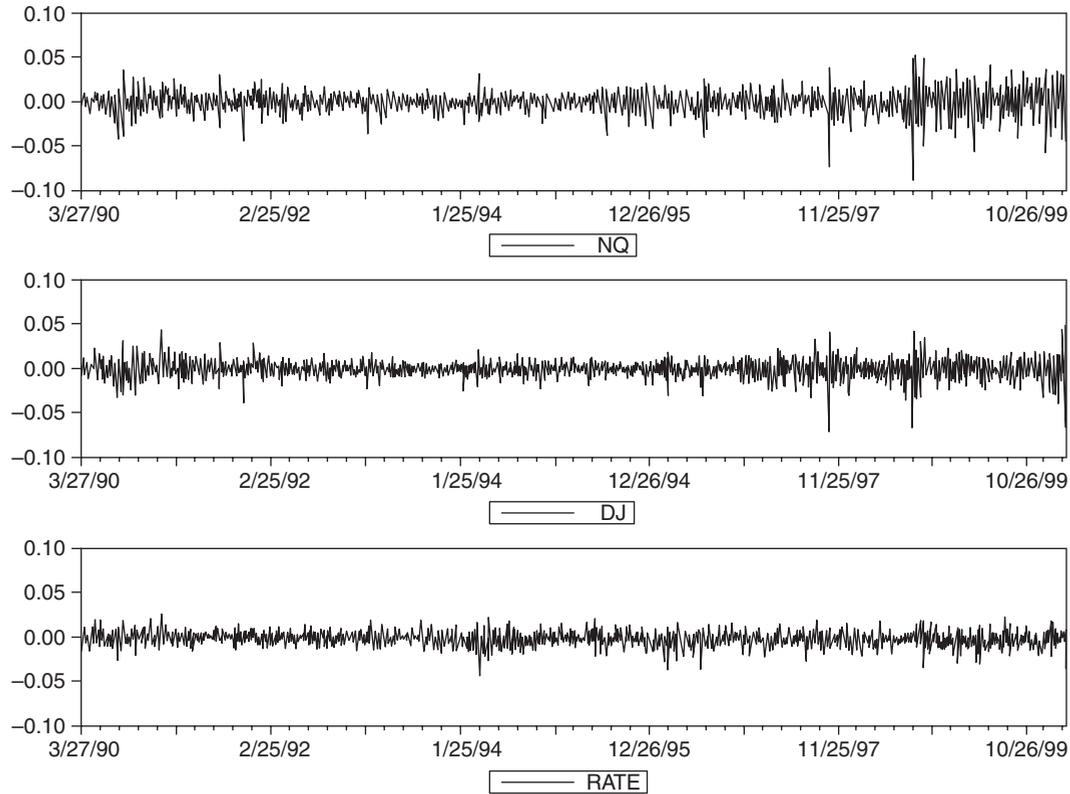


Figure 1 Nasdaq, Dow Jones, and Bond Returns

and variance if the model is correctly specified. Various tests, such as tests for autocorrelation in the squares, can detect model failures. The Ljung-Box test with 15 lagged autocorrelations is often used.

Application to Value at Risk

Applications of the ARCH/GARCH approach are widespread in situations where the volatility of returns is a central issue. Many banks and other financial institutions use the idea of *value at risk* (VaR) as a way to measure the risks in their portfolios. The 1% VaR is defined as the number of dollars that one can be 99% certain exceeds any losses for the next day. Let's use the GARCH(1,1) tools to estimate the 1% VaR of a \$1 million portfolio on March 23, 2000. This portfolio consists of 50% Nasdaq, 30% Dow

Jones, and 20% long bonds. We chose this date because, with the fall of equity markets in the spring of 2000, it was a period of high volatility. First, we construct the hypothetical historical portfolio. (All calculations in this example were done with the EViews software program.) Figure 1 shows the pattern of the Nasdaq, Dow Jones, and long Treasury bonds. In Table 1, we present some illustrative statistics for each of these three investments separately and, in

Table 1 Portfolio Data

	NQ	DJ	RATE	PORT
Mean	0.0009	0.0005	0.0001	0.0007
Std. Dev.	0.0115	0.0090	0.0073	0.0083
Skewness	-0.5310	-0.3593	-0.2031	-0.4738
Kurtosis	7.4936	8.3288	4.9579	7.0026

Table 2 GARCH(1,1)

Dependent Variable: PORT				
Sample (adjusted): 3/26/1990 3/23/2000				
Convergence achieved after 16 iterations				
Bollerslev-Wooldrige robust standard errors and covariance				
Variance Equation				
C	0.0000	0.0000	3.1210	0.0018
ARCH(1)	0.0772	0.0179	4.3046	0.0000
GARCH(1)	0.9046	0.0196	46.1474	0.0000
S.E. of regression	0.0083	Akaike info criterion		-6.9186
Sum squared resid	0.1791	Schwarz criterion		-6.9118
Log likelihood	9028.2809	Durbin-Watson stat		1.8413

the final column, for the portfolio as a whole. Then we forecast the standard deviation of the portfolio and its 1% quantile. We carry out this calculation over several different time frames: the entire 10 years of the sample up to March 23, 2000, the year before March 23, 2000, and from January 1, 2000 to March 23, 2000.

Consider first the quantiles of the historical portfolio at these three different time horizons. Over the full 10-year sample, the 1% quantile times \$1 million produces a VaR of \$22,477. Over the last year, the calculation produces a VaR of \$24,653—somewhat higher, but not significantly so. However, if the first quantile is calculated based on the data from January 1, 2000, to March 23, 2000, the VaR is \$35,159. Thus, the level of risk has increased significantly over the last quarter.

The basic GARCH(1,1) results are given in Table 2. Notice that the coefficients sum up to a number slightly less than one. The forecasted standard deviation for the next day is 0.014605, which is almost double the average standard deviation of 0.0083 presented in the last column of Table 1. If the residuals were normally distributed, then this would be multiplied by 2.326348, giving a VaR equal to \$33,977. As it turns out, the standardized residuals, which are the estimated values of $\{\varepsilon_t\}$, have a 1% quantile of 2.8437, which is well above the normal quantile. The estimated 1% VaR is \$39,996. Notice that this VaR has risen to reflect the increased risk in 2000.

Finally, the VaR can be computed based solely on estimation of the quantile of the forecast distribution. This has been proposed by Engle and Manganelli (2001), adapting the quantile regression methods of Koenker and Basset (1978). Application of their method to this dataset delivers a VaR of \$38,228. Instead of assuming the distribution of return series, Engle and Manganelli (2004) propose a new VaR modeling approach, *conditional autoregressive value at risk (CAViaR)*, to directly compute the quantile of an individual financial asset. On a theoretical level, due to structural changes of the return series, the constant-parameter CAViaR model can be extended. Huang et al. (2010) formulate a time-varying CAViaR model, which they call an index-exciting time-varying CAViaR model. The model incorporates the market index information to deal with the unobservable structural break points for the individual risky asset.

WHY ARCH/GARCH?

The ARCH/GARCH framework proved to be very successful in predicting volatility changes. Empirically, a wide range of financial and economic phenomena exhibit the clustering of volatilities. As we have seen, ARCH/GARCH models describe the time evolution of the average size of squared errors, that is, the evolution of the magnitude of uncertainty. Despite the empirical success of ARCH/GARCH models, there is no real consensus on the economic

reasons why uncertainty tends to cluster. That is why models tend to perform better in some periods and worse in other periods.

It is relatively easy to induce ARCH behavior in simulated systems by making appropriate assumptions on agent behavior. For example, one can reproduce ARCH behavior in artificial markets with simple assumptions on agent decision-making processes. The real economic challenge, however, is to explain ARCH/GARCH behavior in terms of features of agents behavior and/or economic variables that could be empirically ascertained.

In classical physics, the amount of uncertainty inherent in models and predictions can be made arbitrarily low by increasing the precision of initial data. This view, however, has been challenged in at least two ways. First, quantum mechanics has introduced the notion that there is a fundamental uncertainty in any measurement process. The amount of uncertainty is prescribed by the theory at a fundamental level. Second, the theory of complex systems has shown that nonlinear complex systems are so sensitive to changes in initial conditions that, in practice, there are limits to the accuracy of any model. ARCH/GARCH models describe the time evolution of uncertainty in a complex system.

In financial and economic models, the future is always uncertain but over time we learn new information that helps us forecast this future. As asset prices reflect our best forecasts of the future profitability of companies and countries, these change whenever there is news. ARCH/GARCH models can be interpreted as measuring the intensity of the news process. Volatility clustering is most easily understood as news clustering. Of course, many things influence the arrival process of news and its impact on prices. Trades convey news to the market and the macroeconomy can moderate the importance of the news. These can all be thought of as important determinants of the volatility that is picked up by ARCH/GARCH.

GENERALIZATIONS OF THE ARCH/GARCH MODELS

Thus far, we have described the fundamental ARCH and GARCH models and their application to VaR calculations. The ARCH/GARCH framework proved to be a rich framework and many different extensions and generalizations of the initial ARCH/GARCH models have been proposed. We will now describe some of these generalizations and extensions. We will focus on applications in finance and will continue to use financial notation assuming that our variables represent returns of assets or of portfolios.

Let's first discuss why we need to generalize the ARCH/GARCH models. There are three major extensions and generalizations:

1. Integration of first, second, and higher moments
2. Generalization to high-frequency data
3. Multivariate extensions

Integration of First, Second, and Higher Moments

In the ARCH/GARCH models considered thus far, returns are assumed to be normally distributed and the forecasts of the first and second moments independent. These assumptions can be generalized in different ways, either allowing the conditional distribution of the error terms to be non-normal and/or integrating the first and second moments.

Let's first consider asymmetries in volatility forecasts. There is convincing evidence that the direction does affect volatility. Particularly for broad-based equity indexes and bond market indexes, it appears that market declines forecast higher volatility than do comparable market increases. There are now a variety of asymmetric GARCH models, including the *exponential GARCH (EGARCH)* model of Nelson (1991), the *threshold ARCH (TARCH)* model attributed to Rabemananjara and Zakoian (1993) and Glosten, Jagannathan, and Runkle (1993),

and a collection and comparison by Engle and Ng (1993).

In order to illustrate asymmetric GARCH, consider, for example, the asymmetric GARCH(1,1) model of Glosten, Jagannathan, and Runkle (1993). In this model, we add a term $\gamma (I_{\{\varepsilon_t < 0\}}) \varepsilon_t^2$ to the basic GARCH:

$$h_{t+1} = \omega + \alpha h_t \varepsilon_t^2 + \gamma (I_{\{\varepsilon_t < 0\}}) \varepsilon_t^2 + \beta h_t.$$

The term $(I_{\{\varepsilon_t < 0\}})$ is an indicator function that is zero when the error is positive and 1 when it is negative. If γ is positive, negative errors are leveraged. The parameters of the model are assumed to be positive. The relationship $\alpha + \beta + \gamma/2 < 1$ is assumed to hold.

In addition to asymmetries, it has been empirically found that residuals of ARCH/GARCH models fitted to empirical financial data exhibit excess kurtosis. One way to handle this problem is to consider non-normal distributions of errors. Non-normal distributions of errors were considered by Bollerslev (1987), who introduced a GARCH model where the variable z follows a Student- t distribution.

Let's now discuss the integration of first and second moments through the *GARCH-M* model. ARCH/GARCH models imply that the risk inherent in financial markets varies over time. Given that financial markets implement a risk-return trade-off, it is reasonable to ask whether changing risk entails changing returns. Note that, in principle, predictability of returns in function of predictability of risk is not a violation of market efficiency. To correlate changes in volatility with changes in returns, Engle, Lilien, and Robins (1987) proposed the GARCH-M model (not to be confused with the multivariate *MGARCH* model that will be described shortly). The GARCH-M model, or GARCH in mean model, is a complete nonlinear model of asset returns and not only a specification of the error behavior. In the GARCH-M model, returns are assumed to be a constant plus a term proportional to the conditional variance:

$$r_{t+1} = \mu_t + \sigma_t z_t, \quad \mu_t = \mu_0 + \mu_1 \sigma_t^2$$

where σ_t^2 follows a GARCH process and the z terms are independent and identically distributed (IID) normal variables. Alternatively, the GARCH-M process can be specified making the mean linear in the standard deviation but not in the variance.

The integration of volatilities and expected returns, that is the integration of risk and returns, is a difficult task. The reason is that not only volatilities but also correlations should play a role. The GARCH-M model was extended by Bollerslev (1986) in a multivariate context. The key challenge of these extensions is the explosion in the number of parameters to estimate; we will see this when discussing multivariate extensions in the following sections.

Generalizations to High-Frequency Data

With the advent of electronic trading, a growing amount of data has become available to practitioners and researchers. In many markets, data at transaction level, called tick-by-tick data or *ultra-high-frequency data*, are now available. The increase of data points in moving from daily data to transaction data is significant. For example, the average number of daily transactions for U.S. stocks in the Russell 1000 is in the order of 2,000. Thus, we have a 2,000-fold increase in data going from daily data to tick-by-tick data.

The interest in high-frequency data is twofold. First, researchers and practitioners want to find events of interest. For example, the measurement of intraday risk and the discovery of trading profit opportunities at short time horizons are of interest to many financial institutions. Second, researchers and practitioners would like to exploit high-frequency data to obtain more precise forecasts at the usual forecasting horizon. Let's focus on the latter objective.

As observed by Merton (1980), while in diffusive processes the estimation of trends requires long stretches of data, the estimation of volatility can be done with arbitrary precision using data extracted from arbitrarily short time

periods provided that the sampling rate is arbitrarily high. In other words, in diffusive models, the estimation of volatility greatly profits from high-frequency data. It therefore seems tempting to use data at the highest possible frequency, for example spaced at a few minutes, to obtain better estimates of volatility at the frequency of practical interest, say daily or weekly. As we will see, the question is not so straightforward and the answer is still being researched.

We will now give a brief account of the main modeling strategies and the main obstacles in using high-frequency data for volatility estimates. We will first assume that the return series are sampled at a high but fixed frequency. In other words, we initially assume that data are taken at fixed intervals of time. Later, we will drop this assumption and consider irregularly spaced tick-by-tick data, what Engle (2000) refers to as “ultra-high-frequency data.”

Let's begin by reviewing some facts about the *temporal aggregation* of models. The question of temporal aggregation is the question of whether models maintain the same form when used at different time scales. This question has two sides: empirical and theoretical. From the empirical point of view, it is far from being obvious that econometric models maintain the same form under temporal aggregation. In fact, patterns found at some time scales might disappear at another time scale. For example, at very short time horizons, returns exhibit autocorrelations that disappear at longer time horizons. Note that it is not a question of the precision and accuracy of models. Given the uncertainty associated with financial modeling, there are phenomena that exist at some time horizon and disappear at other time horizons.

Time aggregation can also be explored from a purely theoretical point of view. Suppose that a time series is characterized by a given data-generating process (DGP). We want to investigate what DGPs are closed under temporal aggregation; that is, we want to investigate

what DGPs, eventually with different parameters, can represent the same series sampled at different time intervals.

The question of time aggregation for GARCH processes was explored by Drost and Nijman (1993). Consider an infinite series $\{x_t\}$ with given fixed-time intervals $\Delta x_t = x_{t+1} - x_t$. Suppose that the series $\{x_t\}$ follows a GARCH(p, q) process. Suppose also that we sample this series at intervals that are multiples of the basic intervals: $\Delta y_t = h \Delta x_t = x_{t+h} - x_t$. We obtain a new series $\{y_t\}$. Drost and Nijman found that the new series $\{y_t\}$ does not, in general, follow another GARCH(p', q') process. The reason is that, in the standard GARCH definition presented in the previous sections, the series $\{x_t = \sigma_t z_t\}$ is supposed to be a martingale difference sequence (that is, a process with zero conditional mean). This property is not conserved at longer time horizons.

To solve this problem, Drost and Nijman introduced weak GARCH processes, processes that do not assume the martingale difference condition. They were able to show that weak GARCH(p, q) models are closed under temporal aggregation and established the formulas to obtain the parameters of the new process after aggregation. One consequence of their formulas is that the fluctuations of volatility tend to disappear when the time interval becomes very large. This conclusion is quite intuitive given that conditional volatility is a mean-reverting process.

Christoffersen, Diebold, and Schuerman (1998) use the Drost and Nijman formula to show that the usual scaling of volatility, which assumes that volatility scales with the square root of time as in the random walk, can be seriously misleading. In fact, the usual scaling magnifies the GARCH effects when the time horizon increases while the Drost and Nijman analysis shows that the GARCH effect tends to disappear with growing time horizons. If, for example, we fit a GARCH model to daily returns and then scale to monthly volatility multiplying by the square root of the number of days in

a month, we obtain a seriously biased estimate of monthly volatility.

Various proposals to exploit high-frequency data to estimate volatility have been made. Meddahi and Renault (2004) proposed a class of autoregressive stochastic volatility models—the SR-SARV model class—that are closed under temporal aggregation; they thereby avoid the limitations of the weak GARCH models. Andersen and Bollerslev (1998) proposed realized volatility as a virtually error-free measure of instantaneous volatility. To compute realized volatility using their model, one simply sums intraperiod high-frequency squared returns.

Thus far, we have briefly described models based on regularly spaced data. However, the ultimate objective in financial modeling is using all the available information. The maximum possible level of information on returns is contained in tick-by-tick data. Engle and Russell (1998) proposed the *autoregressive conditional duration* (ACD) model to represent sequences of random times subject to clustering phenomena. In particular, the ACD model can be used to represent the random arrival of orders or the random time of trade execution.

The arrival of orders and the execution of trades are subject to clustering phenomena insofar as there are periods of intense trading activity with frequent trading followed by periods of calm. The ACD model is a point process. The simplest point process is likely the Poisson process, where the time between point events is distributed as an exponential variable independent of the past distribution of points. The ACD model is more complex than a Poisson process because it includes an autoregressive effect that induces the point process equivalent of ARCH effects. As it turns out, the ACD model can be estimated using standard ARCH/GARCH software. Different extensions of the ACD model have been proposed. In particular, Bauwens and Giot (1997) introduced the logarithmic ACD model to represent the bid-ask prices in the Nasdaq stock market.

Ghysel and Jasiak (1997) introduced a class of approximate ARCH models of returns series sampled at the time of trade arrivals. This model class, called *ACD-GARCH*, uses the ACD model to represent the arrival times of trades. The GARCH parameters are set as a function of the duration between transactions using insight from the Drost and Nijman weak GARCH. The model is bivariate and can be regarded as a random coefficient GARCH model.

Multivariate Extensions

The models described thus far are models of single assets. However, in finance, we are also interested in the behavior of portfolios of assets. If we want to forecast the returns of portfolios of assets, we need to estimate the correlations and covariances between individual assets. We are interested in modeling correlations not only to forecast the returns of portfolios but also to respond to important theoretical questions. For example, we are interested in understanding if there is a link between the magnitude of correlations and the magnitude of variances and how correlations propagate between different markets. Questions like these have an important bearing on investment and risk management strategies.

Conceptually, we can address covariances in the same way as we addressed variances. Consider a vector of N return processes: $r_t = \{r_{i,t}\}$, $i = 1, \dots, N$, $t = 1, \dots, T$. At every moment t , the vector r_t can be represented as: $r_t = m_t(\vartheta) + \varepsilon_t$, where $m_t(\vartheta)$ is the vector of conditional means that depend on a finite vector of parameters ϑ and the term ε_t is written as:

$$\varepsilon_t = H_t^{1/2}(\vartheta) z_t$$

where $H_t^{1/2}(\vartheta)$ is a positive definite matrix that depends on the finite vector of parameters ϑ . We also assume that the N -vector z_t has the following moments: $E(z_t) = 0$, $\text{Var}(z_t) = I_N$ where I_N is the $N \times N$ identity matrix.

To explain the nature of the matrix $H_t^{1/2}(\vartheta)$, consider that we can write:

$$\begin{aligned}\text{Var}(r_t|I_{t-1}) &= \text{Var}_{t-1}(r_t) = \text{Var}_{t-1}(\varepsilon_t) \\ &= H_t^{1/2} \text{Var}_{t-1}(z_t) H_t^{1/2'} = H_t\end{aligned}$$

where I_{t-1} is the information set at time $t - 1$. For simplicity, we left out in the notation the dependence on the parameters ϑ . Thus $H_t^{1/2}$ is any positive definite $N \times N$ matrix such that H_t is the conditional covariance matrix of the process r_t . The matrix $H_t^{1/2}$ could be obtained by Cholesky factorization of H_t . Note the formal analogy with the definition of the univariate process.

Consider that both the vector $m_t(\vartheta)$ and the matrix $H_t^{1/2}(\vartheta)$ depend on the vector of parameters ϑ . If the vector ϑ can be partitioned into two subvectors, one for the mean and one for the variance, then the mean and the variance are independent. Otherwise, there will be an integration of mean and variance as was the case in the univariate GARCH-M model. Let's abstract from the mean, which we assume can be modeled through some autoregressive process, and focus on the process $\varepsilon_t = H_t^{1/2}(\vartheta) z_t$.

We will now define a number of specifications for the variance matrix H_t . In principle, we might consider the covariance matrix heteroskedastic and simply extend the ARCH/GARCH modeling to the entire covariance matrix. There are three major challenges in MGARCH models:

1. Determining the conditions that ensure that the matrix H_t is positive definite for every t .
2. Making estimates feasible by reducing the number of parameters to be estimated.
3. Stating conditions for the weak stationarity of the process.

In a multivariate setting, the number of parameters involved makes the (conditional) covariance matrix very noisy and virtually impossible to estimate without appropriate restrictions. Consider, for example, a large aggregate such as the S&P 500. Due to symme-

tries, there are approximately 125,000 entries in the conditional covariance matrix of the S&P 500. If we consider each entry as a separate GARCH(1,1) process, we would need to estimate a minimum of three GARCH parameters per entry. Suppose we use three years of data for estimation, that is, approximately 750 data points for each stock's daily returns. In total, there are then $500 \times 750 = 375,000$ data points to estimate $3 \times 125,000 = 375,000$ parameters. Clearly, data are insufficient and estimation is therefore very noisy. To solve this problem, the number of independent entries in the covariance matrix has to be reduced.

Consider that the problem of estimating large covariance matrices is already severe if we want to estimate the unconditional covariance matrix of returns. Using the theory of random matrices, Potter, Bouchaud, Laloux, and Cizeau (1999) show that only a small number of the eigenvalues of the covariance matrix of a large aggregate carry information, while the vast majority of the eigenvalues cannot be distinguished from the eigenvalues of a random matrix. Techniques that impose constraints on the matrix entries, such as factor analysis or principal components analysis, are typically employed to make less noisy the estimation of large covariance matrices.

Assuming that the conditional covariance matrix is time varying, the simplest estimation technique is using a rolling window. Estimating the covariance matrix on a rolling window suffers from the same problems already discussed in the univariate case. Nevertheless, it is one of the two methods used in RiskMetrics. The second method is the EWMA method. EWMA estimates the covariance matrix using the following equation:

$$H_t = \alpha \varepsilon_t \varepsilon_t' + (1 - \alpha) H_{t-1}$$

where α is a small constant.

Let's now turn to multivariate GARCH specifications, or MGARCH, and begin by introducing the vech notation. The vech operator stacks the lower triangular portion of an $N \times N$ matrix

as an $N(N+1)/2 \times 1$ vector. In the vech notation, the MGARCH(1,1) model, called the VEC model, is written as follows:

$$h_t = \omega + A\eta_{t-1} + Bh_{t-1}$$

where $h_t = \text{vech}(H_t)$, ω is an $N(N+1)/2 \times 1$ vector, and A, B are $N(N+1)/2 \times N(N+1)/2$ matrices.

The number of parameters in this model makes its estimation impossible except in the bivariate case. In fact, for $N = 3$ we should already estimate 78 parameters. In order to reduce the number of parameters, Bollerslev, Engle, and Wooldridge (1988) proposed the diagonal VEC model (DVEC), imposing the restriction that the matrices A, B be diagonal matrices. In the DVEC model, each entry of the covariance matrix is treated as an individual GARCH process. Conditions to ensure that the covariance matrix H_t is positive definite are derived in Attanasio (1991). The number of parameters of the DVEC model, though much smaller than the number of parameters in the full VEC formulation, is still very high: $3N(N+1)/2$.

To simplify conditions to ensure that H_t is positive definite, Engle and Kroner (1995) proposed the BEKK model (the acronym BEKK stands for Baba, Engle, Kraft, and Kroner). In its most general formulation, the BEKK(1,1,K) model is written as follows:

$$H_t = CC' + \sum_{k=1}^K A'_k \varepsilon_{t-1} \varepsilon'_{t-1} A_k + \sum_{k=1}^K B'_k H_{t-1} B_k$$

where C, A_k, B_k are $N \times N$ matrices and C is upper triangular. The BEKK(1,1,1) model simplifies as follows:

$$H_t = CC' + A' \varepsilon_{t-1} \varepsilon'_{t-1} A + B' H_{t-1} B$$

which is a multivariate equivalent of the GARCH(1,1) model. The number of parameters in this model is very large; the diagonal BEKK was proposed to reduce the number of parameters.

The VEC model can be weakly (covariance) stationary but exhibit a time-varying conditional covariance matrix. The stationarity con-

ditions require that the eigenvalues of the matrix $A + B$ are less than one in modulus. Similar conditions can be established for the BEKK model. The unconditional covariance matrix H is the unconditional expectation of the conditional covariance matrix. We can write:

$$H = [I_{N^*} - A - B]^{-1}, \quad N^* = N(N+1)/2 \times$$

MGARCH based on factor models offers a different modeling strategy. Standard (strict) factor models represent returns as linear regressions on a small number of common variables called factors:

$$r_t = m + Bf_t + \varepsilon_t$$

where r_t is a vector of returns, f_t is a vector of K factors, B is a matrix of factor loadings, ε_t is noise with diagonal covariance, so that the covariance between returns is accounted for only by the covariance between the factors. In this formulation, factors are static factors without a specific time dependence. The unconditional covariance matrix of returns Ω can be written as:

$$\Omega = B\Omega_K B' + \Sigma$$

where Ω_K is the covariance matrix of the factors.

We can introduce a dynamics in the expectations of returns of factor models by making some or all of the factors dynamic, for example, assuming an autoregressive relationship:

$$\begin{aligned} r_t &= m + Bf_t + \varepsilon_t \\ f_{t+1} &= a + bf_t + \eta_t \end{aligned}$$

We can also introduce a dynamic of volatilities assuming a GARCH structure for factors. Engle, Ng, and Rothschild (1990) used the notion of *factors* in a dynamic conditional covariance context assuming that one factor, the market factor, is *dynamic*. Various GARCH factor models have been proposed: the *F-GARCH* model of Lin (1992); the full factor FF-GARCH model of Vrontos, Dellaportas, and Politis (2003); the orthogonal O-GARCH model of Kariya (1988); and Alexander and Chibumba (1997).

Another strategy is followed by Bollerslev (1990) who proposed a class of GARCH models in which the conditional correlations are constant and only the idiosyncratic variances are time varying (CCC model). Engle (2002) proposed a generalization of Bollerslev's CCC model called the dynamic conditional correlation (DCC) model.

KEY POINTS

- Volatility, a key parameter used in many financial applications, measures the size of the errors made in modeling returns and other financial variables. For vast classes of models, the average size of volatility is not constant but changes with time and is predictable.
- In standard regression theory, the assumption of homoskedasticity is convenient from a mathematical point of view. The homoskedasticity assumption means that the expected size of the error is constant and does not depend on the size of the explanatory variable. When it is assumed in regression analysis that the expected size of the error term is not constant, this means the error terms are assumed to be heteroskedastic.
- A major breakthrough in econometric modeling was the discovery that for many families of econometric models it is possible to specify a stochastic process for the error terms and predict the average size of the error terms when models are fitted to empirical data. This is the essence of ARCH modeling. This original modeling of conditional heteroskedasticity has developed into a full-fledged econometric theory of the time behavior of the errors of a large class of univariate and multivariate models.
- The availability of more and better data and the availability of low-cost, high-performance computers allowed the development of a vast family of ARCH/GARCH models. Among these are the EGARCH, IGARCH, GARCH-M, MGARCH, and ACD models.
- While the forecasting of expected returns remains a rather elusive task, predicting the level of uncertainty and the strength of comovements between asset returns has become a fundamental pillar of financial econometrics.

REFERENCES

- Alexander, C. O., and Chibumba, A. M. (1997). Multivariate orthogonal factor GARCH. University of Sussex.
- Attanasio, O. (1991). Risk, time-varying second moments and market efficiency. *Review of Economic Studies* 58: 479–494.
- Andersen, T. G., and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review* 39, 4: 885–905.
- Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica* 71: 579–625.
- Bauwens, L., and Giot, P. (1997). The logarithmic ACD model: An application to market microstructure and NASDAQ. Université Catholique de Louvain—CORE discussion paper 9789.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31: 307–327.
- Bollerslev, T. (1990). Modeling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH approach. *Review of Economics and Statistics* 72: 498–505.
- Bollerslev, T., Engle, R. F., and Wooldridge, J. M. (1988). A capital asset pricing model with time-varying covariance. *Journal of Political Economy* 96, 1: 116–131.
- Drost, C. D., and Nijman, T. (1993). Temporal aggregation of GARCH processes. *Econometrica* 61: 909–927.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 4: 987–1007.
- Engle, R. F. (2000). The econometrics of ultra high frequency data. *Econometrica* 68, 1: 1–22.
- Engle, R. F. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business and Economic Statistics* 20: 339–350.

- Engle, R. F., Lilien, D., and Robins, R. (1987). Estimating time varying risk premia in the term structure: The ARCH-M model. *Econometrica* 55: 391–407.
- Engle, R. F., and Ng, V. (1993). Measuring and testing the impact of news on volatility. *Journal of Finance* 48, 5: 1749–1778.
- Engle, R. F., and Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business and Economic Statistics* 22: 367–381.
- Engle, R. F., Ng, V., and Rothschild, M. (1990). Asset pricing with a factor-ARCH covariance structure: Empirical estimates for Treasury bills. *Journal of Econometrics* 45: 213–238.
- Engle, R. F., and Russell, J. R. (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica* 66: 1127–1162.
- Ghysels, E., and Jasiak, J. (1997). GARCH for irregularly spaced financial data: The ACD-GARCH model. DP 97s-06. CIRANO, Montréal.
- Glosten, L. R., Jagannathan, R., and Runkle, D. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance* 48, 5: 1779–1801.
- Haavelmo, M. T. (1944). The probability approach in econometrics. *Econometrica* 12 (Supplement): 1–115.
- Huang, D., Yu, B., Lu, Z., Focardi, S., Fabozzi, F. J., and Fukushima, M. (2010). Index-exciting CAViaR: A new empirical time-varying risk model. *Studies in Nonlinear Dynamics and Econometrics* 14, 2: Article 1.
- Kariya, T. (1988). MTV model and its application to the prediction of stock prices. In T. Pullila and S. Puntanen (eds.), *Proceedings of the Second International Tampere Conference in Statistics*. University of Tampere, Finland.
- Lin, W. L. (1992). Alternative estimators for factor GARCH models—a Monte Carlo comparison. *Journal of Applied Econometrics* 7: 259–279.
- Meddahi, N., and Renault, E. (2004). Temporal aggregation of volatility models. *Journal of Econometrics* 119: 355–379.
- Merton, R. C. (1980). On estimating the expected return on the market: An exploratory investigation. *Journal of Financial Economics* 8: 323–361.
- Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* 59, 2: 347–370.
- Potters, M., Bouchaud, J.-P., Laloux, L., and Cizeau, P. (1999). Noise dressing of financial correlation matrices. *Physical Review Letters* 83, 7: 1467–1489.
- Rabemananjara, R., and Zakoian, J. M. (1993). Threshold ARCH models and asymmetries in volatility. *Journal of Applied Econometrics* 8, 1: 31–49.
- Vrontos, I. D., Dellaportas, P., and Politis, D. N. (2003). A full-factor multivariate GARCH model. *Econometrics Journal* 6: 311–333.

Classification and Regression Trees and Their Use in Financial Modeling

MIN ZHU

Business School, Queensland University of Technology, Australia

DAVID PHILPOTTS

QEP Global Equities, Schroder Investment Management, Sydney, Australia

MAXWELL J. STEVENSON, PhD

Discipline of Finance, Business School, University of Sydney, Australia

Abstract: Classification and regression trees (CART) are nonparametric and nonlinear modeling techniques that do not rely upon the many stringent assumptions required by classical parametric models. Despite the fact that researchers in many fields have regularly found trees to be an attractive way to express underlying relationships, they are relatively unfamiliar to financial modelers where the historical focus of financial modeling has been on parametric regression. Although the linear type of regression analysis is convenient and sometimes intuitive, it may not fully capture the complexity of financial markets. As the quantity and variety of financial information available to data exploration have increased over time, there has been a commensurate need for a more robust and versatile process to analyze these data. CART offers a valuable alternative to traditional methods for modeling financial data.

Classification and regression trees (CART) are nonparametric and nonlinear modeling techniques that essentially use recursive partitioning techniques to separate observations in a binary and sequential fashion. There are two varieties: (1) classification trees when the dependent variable is categorical, and (2) regression trees when the dependent variable is continuous.

Although the approach is not widely utilized within the investment community, the applications of CART to financial markets

nevertheless include the classification of financially distressed firms by Frydman, Altman, and Kao (1985), asset allocation by Sorensen, Mezrich, and Miller (1998), equity style timing by Kao and Shumaker (1999), and stock selection by Sorensen, Miller, and Ooi (2000). In this entry we provide an introduction to the CART framework and contrast it to more traditional modeling methods. We then illustrate the technique by applying it to stock selection across the North American equity markets.

TECHNICAL DETAILS

We begin by introducing the standard tree terminology. The root is the top node, which includes all observations in the learning sample. The splitting condition at each node is expressed as an “if-then-else” rule that is determined by a specific splitting criterion. The splitting node is also called the parent and the two descendant subnodes are called children. A node keeps splitting until a terminal node or leaf is reached.

The fundamental idea behind CART is to recursively partition the space until all the subspaces are sufficiently homogenous in order to apply simple models to them. This is in contrast to linear and logistic regressions, the linear parametric counterparts of regression and classification trees, respectively, which are global models where a single predictive formula is imposed over the entire data space. When the dataset has multiple features that interact in complicated and nonlinear ways, as is often the case with financial data, a single global model may inadequately capture the underlying relationships.

There are two major steps in the CART analysis: (1) Build a tree using a recursive splitting of nodes, and (2) prune the tree in order to obtain the optimal tree size so as to prevent overfitting. Each of these two steps will be discussed in more detail below. Breiman et al. (1984) provide a detailed overview of the theory and methodology of CART, including a number of examples from many disciplinary areas. There are also many software packages that implement the CART algorithm. Popular ones include R packages such as *rpart* and *tree* and the Matlab function *classregtree*.

Binary Recursive Partitioning

Let \mathcal{L} be a learning sample, $\mathcal{L} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, where x_i is a vector of attributes; y_i is the response, which can be categorical or continuous; and n is the number of observations. The attribute vector x_i belongs to X , the

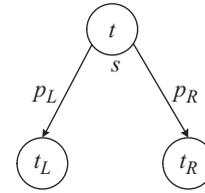


Figure 1 A split generates two children of the node t , denoted by t_L and t_R . A proportion p_L of the initial data go into the left child and a proportion of p_R go into the right child.

attribute space. The tree-building algorithm involves repeatedly splitting subsets of \mathcal{L} into two descendant subsets, beginning with \mathcal{L} itself. For a continuous variable x_i , the allowed splits are of the form $x_i < c$ versus $x_i \geq c$. For categorical variables the levels are divided into two classes. Therefore, for a categorical variable with K levels, there are $2^{K-1} - 1$ possible splits, disallowing the empty split and ignoring the order.

In choosing the best splitting rule, CART seeks to maximize the average purity of the two child nodes. Hence, some criterion measuring data homogeneity or, alternatively, impurity should be introduced. These impurity measures are loosely classed splitting criteria. Let us introduce, for any node t , a measure $i(t)$ that signifies the impurity of the node. Suppose that a candidate split s divides the node into t_L and t_R such that a proportion p_L of the cases in t go into t_L and a proportion p_R go into t_R (see Figure 1). Then the goodness of the split is defined to be the decrease of impurity

$$\Delta i(s, t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

For an arbitrary node t and a set of splitting candidates S , the optimal split is chosen to be the one

$$s^* = \max_{s \in S} \Delta i(s, t)$$

In other words, the optimal split is the one that reduces impurity by the greatest amount.

The idea for classification and regression trees is quite similar in terms of partitioning methods, which is based on impurity reducing.

However, they use different measures of impurity to decode the split.

In a classification problem, suppose that we want to classify data into K classes. At each node t of a classification tree we have a probability distribution p_{tk} , $k = 1, \dots, K$, over all K categories. The probabilities are conventionally estimated from the node proportions, such that $p_{tk} = n_{tk}/n_t$, where n_{tk} is the number of observations in the k -th class, and n_t is the sample size at node t .

The two most common measures of impurity are the Gini index

$$i(t) = \sum_{j \neq k} p_{tj} p_{tk} = 1 - \sum_k p_{tk}^2$$

and entropy or information

$$i(t) = - \sum_k p_{tk} \log(p_{tk})$$

where $0 \log(0) = 0$.

As for regression trees, the most popular impurity measure is

$$i(t) = \sum_{j=1}^{n_t} (y_{tj} - \mu_t)^2$$

where the constant μ_t for node t is estimated by the mean of the values of the training data falling into node t .

TREE PRUNING

However, the use of partitioning rules alone cannot guarantee a useful tree model. If reducing impurity is the only goal in tree induction, we will eventually end up with a maximal tree, which has one observation or one class in each leaf, whichever reaches first. This kind of tree adapts too well to the features of the learning sample and has a very high risk of being overfitted. Tree pruning is a way to improve the robustness of the model by trading off in-sample fitting against out-of-sample accuracy. This is particularly important if the model is being used to make predictions.

The best-known procedure for tree pruning is the cost-complexity pruning proposed by Breiman et al. (1984). Let T be a subtree of the maximal tree grown without pruning. Let the

size of a tree be the number of terminal nodes. The optimal tree is the one that minimizes the following cost-complexity measure

$$R_\alpha(T) = R(T) + \alpha \text{size}(T)$$

where α is a complexity parameter to penalize the tree size, and R is the cost, which is commonly taken as misclassification errors in classification cases and deviance in regression cases. For a given value of the complexity parameter α , an optimal tree can be determined. In general, finding the optimal value for α would require an independent set of data (i.e., a testing sample). This requirement is often avoided in practice by using a cross validation procedure.

STRENGTHS AND WEAKNESSES OF CART

Compared to classical parametric models, CART offers a number of benefits in data exploration. In particular, it has a very high degree of interpretability. CART efficiently compresses a large volume of data into an easy-to-understand graphical form that identifies the essential characteristics. It is also very flexible in terms of the structure of the input variables, as either categorical or continuous factors or a combination can be used as inputs. Furthermore, CART is quite robust in the presence of outliers and well suited to noisy datasets, both of which tend to be features of financial data.

Being nonparametric it does not require any assumptions to be made about the underlying distribution of the variables being modeled. The high incidences of extreme events in the financial markets suggest that the supposition of distributional normality is questionable in many areas of finance. While the assumption may in many cases be a fair approximation to the underlying structural relationship, it is quite rare that tests for non-normality or nonlinearity are explicitly assessed in advance even though this information would help to inform the appropriate choice of modeling technique.

The CART approach also departs from traditional modeling methods by determining a hierarchy of input variables that may be closer to the human decision-making processes. Indeed, a key strength of CART over classical modeling methods is that it allows one to represent various types of *interactions* between variables, particularly conditional relevance. Conditional relevance occurs if a factor is relevant only when it is conditioned upon some other factor. For example, only if a certain condition is met by the first high-level attribute is a second attribute taken into consideration. The same holds for the next attribute in the tree hierarchy, and so on.

Another possible benefit for financial modelers using CART is the diversification of model risk as argued by Philpotts et al. (2011). The widespread use of linear modeling methodologies among quantitative asset managers, taken together with the similarity in data sources and risk models, may in turn have contributed to model risk in financial markets leading to a high degree of commonality in investment decisions. As a less used technique, CART is appealing in the context of potentially offering a degree of model diversification. Philpotts et al. (2011) present empirical evidence highlighting the favorable performance of tree-based models compared to more traditional modeling techniques.

One potential weakness of the recursive partitioning tree construction process is local optimization instead of global optimization. That is, the sequential node-splitting process chooses the next split without attempting to optimize the performance of the whole tree. The resulting tree structure therefore does not guarantee global optimization. Instability is another possible problem in CART solutions. This refers to perturbing a small proportion of the learning sample or resampling the learning sample, which often results in a solution with a very different tree structure. Several alternatives to CART have been developed to address these

problems, such as random forests (see Brieman, 2001) and a hybrid approach that combines CART with logistic regression (see Zhu et al., 2011).

APPLICATION OF CART IN STOCK SELECTION

In this section, we provide a detailed example of the CART algorithm as applied to the problem of identifying profitable stocks. This example was specifically chosen so as to provide a contrast with the vast majority of the linear modeling techniques used by financial practitioners. The model was built with monthly stock data from December 1986 to August 2010 covering all liquid stocks listed on the North American equity markets but excluding financial stocks because they would require their own specific model.¹ The number of total observations is 279,188 (or 980 stocks per month on average).

At the end of each month, forward total stock returns (price return plus dividends) were calculated. Using the median return of all sample companies in the same period as a proxy of the market return, the excess returns were then computed as the total returns minus the market returns.

A broad spectrum of company valuation and quality-based characteristics, as well as measures of investor sentiment such as price momentum and earnings revisions were selected as reported in Table 1. Instead of using raw values, we use rank orders in order to improve the robustness of the analyses. At each month, the rank order for each variable was computed by first ranking n stocks according to the corresponding variable value, and then dividing the rank by n to scale it between 0 and 1. Furthermore, in order to overcome the high correlation among some of the explanatory variables, nine composite factors were promoted as potential explanatory variables, which were constructed as an equally weighted average of multiple variables as described in Table 1.

Table 1 Input Variables

Composite factor	Description
Value (VAL)	An equally weighted average of value metrics including dividends to price, cash flow to price, sales to price, and book to price.
Profitability (PROF)	An equally weighted average of profitability terms including return on equity, cash return on equity, pretax margins, and asset turnover.
Leverage (LEVERAGE)	An equally weighted average of financial strength terms including debt to equity and debt to market cap.
Debt Service (DEBT.SERVICE)	An equally weighted average of debt sustainability measures including interest cover and free cash flow to debt.
Momentum (MOM)	An equally weighted average of momentum terms measured over various time horizons including 6 months and 12 months.
Stability (STAB)	A composite term that captures the volatility in earnings, sales, and cash flows over the previous 5 years.
Historic Growth (HIST.GROWTH)	An equally weighted average of 3-year historic growth in earnings, sales, and cash flow.
Forward Growth (FWD.GROWTH)	An equally weighted average of I/B/E/S forecasted earnings growth expectation for FY1 and FY2.
Earnings Revisions (EREV)	An equally weighted average of the 3-month change in I/B/E/S forecasted earnings expectations for FY1 and FY2.

We built a classification tree with the purpose of predicting subsequent stock performance. Stocks were sorted into two groups, “outperformers” for those with positive excess returns and “underperformers” for the remainder. The induced categorical variable was then used as the dependent variable in the subsequent modeling process. One of the benefits of working with categorical responses instead of raw returns lies in the fact that it alleviates the impact of extreme returns, which may have multiple causes. The tree model was built with the data up to and including April 2007 while the data between May 2007 and August 2010 were reserved for out-of-sample testing. Figure 2 graphically illustrates the *hierarchical structure* of the stock selection tree.

The first observation to note is that the primary split is valuation based. More specifically, the tree makes a distinction between those stocks that are relatively expensive (the right-hand branch) and those that are not expensive. One of the most attractive nodes splits again on high value and therefore identifies cheap

stocks as having a 59.2% probability of outperforming the universe (Node 1). In contrast, the worst performing stocks are characterized by being expensive and exhibiting low profitability (Node 14). Companies with these attributes only have a 42% chance of outperforming.

The tree is able to identify the exception to the rule. For example, while identifying that value was the most important driver of stock returns, the tree also suggests that more expensive stocks still have a good chance of outperforming the market providing that they are blessed with profitability, stability in earnings, strong momentum, and are also associated with strong earnings revisions (Node 10).

Similarly, the decision tree framework also highlights the nonlinear behavior of the stock returns to the underlying predictor variables. For example, stocks in Nodes 3 and 5 have similar outperforming probabilities but are of opposite preference with regard to leverage. Conditional on above-average debt cover, Node 3 actually prefers some degree of leverage and more significantly penalizes overly

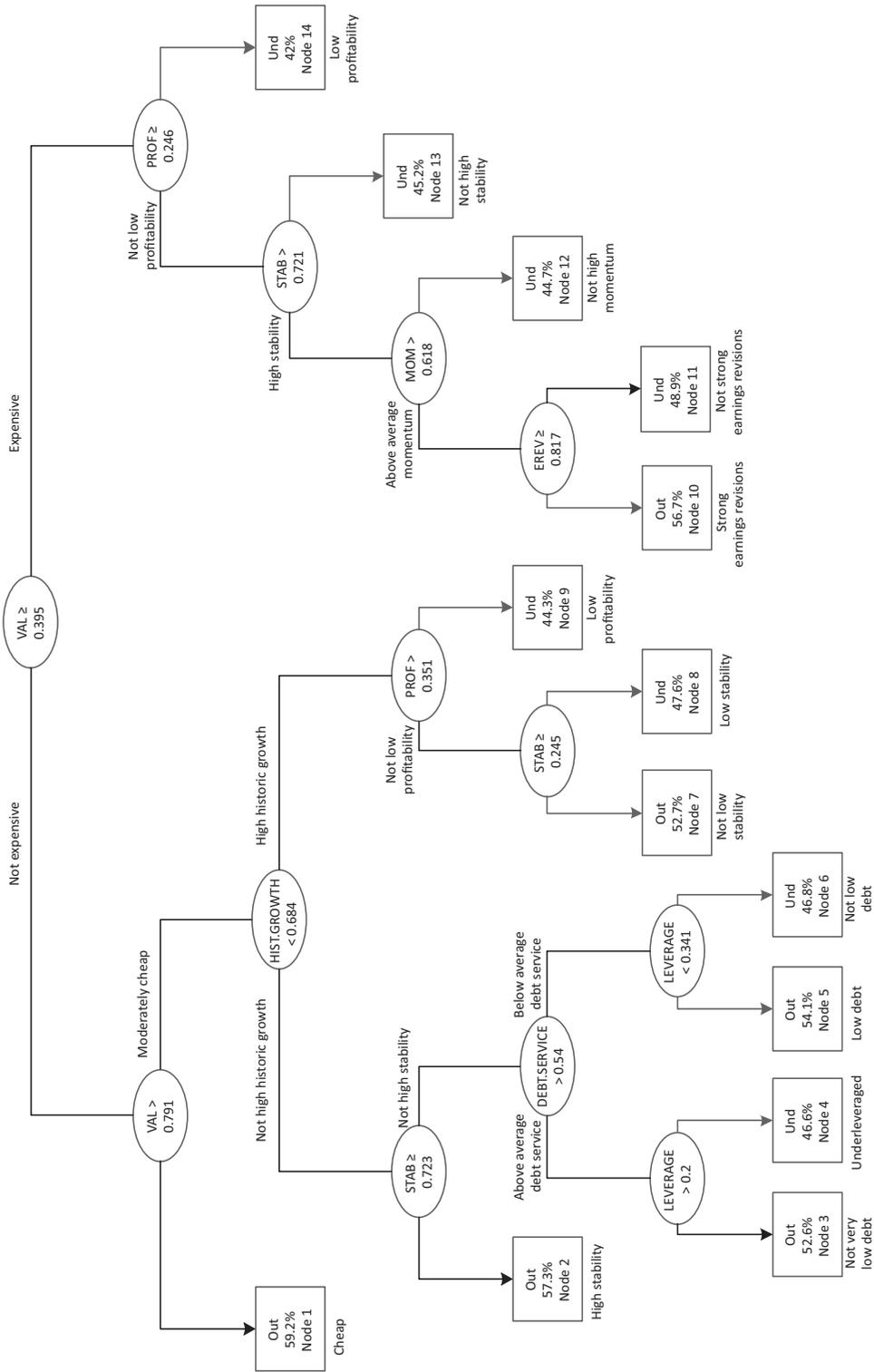


Figure 2 Decision Tree for North American stocks built using data from December 1986 to April 2007 to model the chance of a stock outperforming the overall market. The dependent variable is set as an “outperformer” (Out) if a stock subsequently achieves a higher return than the market, and “underperformer” (Und) otherwise. The outperforming probabilities are reported in percentages at each terminal node along with the splitting criteria.

Table 2 Out-of-Sample Performance (May 2007–August 2010). Portfolios were rebalanced monthly and transaction costs were not taken into account.

Portfolio	Excess Return (%)	Tracking Error (%)	IR	Monthly Win Rate
Long	2.6	2.9	0.89	0.57
Short	−2.8	3.4	−0.82	0.43

conservative firms (with too low leverage). In contrast, leverage is a characteristic to be avoided among firms that cannot service their debts (Node 5).

Table 2 is an out-of-sample test of the model. Each month from May 2007 until August 2010, we ranked all stocks based upon the predicted outperforming probabilities by the tree model and formed two portfolios. One portfolio is an equal weighting of stocks with the highest half of outperforming probabilities (long), and the second is an equal weighting of the rest expected to underperform (short). Table 2 reports the annualized excess return, the tracking error, the information ratio, and the monthly win rate of the two portfolios. The long portfolio outperformed the benchmark by 2.6% with a similar relative risk. The short portfolio underperformed by 2.8% with a slightly higher tracking error. The monthly win rate is the proportion of months that a portfolio outperformed the benchmark out-of-sample. The tree model achieved a monthly win rate of 57%.

KEY POINTS

- CART is a flexible modeling technique that offers significant potential to assist in financial decision making.
- CART is a nonparametric modeling technique that does not impose the stringent assumptions required by classical regression analysis, and therefore sidesteps many of the known issues associated with traditional parametric models.
- CART is well suited to identifying nonlinearities and complex interactions in the data. It

is minimally affected by missing values, outliers, or multicollinearity.

- Unlike many other methods, CART can be easily visualized, which helps financial decision makers to assess the theoretical support behind the resulting investment insights.
- The hierarchical structure of a tree model may more closely resemble how the human brain makes decisions. In particular, the “if-then-else” nature of the rules in the model emulates an expert system that is able to incorporate the exception to the rule.
- CART also embraces the important feature of conditional relevance, which is widespread in financial data. In the CART framework, input variables are allowed to interact and have different influences under varying conditions.
- As with any other quantitative model development process, care must be taken to ensure the integrity of the input data and that the intended application makes intuitive sense.

NOTE

1. Financial stocks were excluded due to their different accounting structure, which makes comparisons with nonfinancials troublesome, although similarly structured stock selection models can also be applied within the sector.

ACKNOWLEDGMENT

Min Zhu gratefully acknowledges financial support from the Capital Market Cooperative Research Centre (CMCRC).

REFERENCES

- Breiman, L. (2001). Random forests. *Machine Learning* 45: 5–32.
- Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.
- Frydman, H., Altman, E. I., and Kao, D. L. (1985). Introducing recursive partitioning for financial

- classification: The case of financial distress. *Journal of Finance* 40: 269–291.
- Kao, D. L., and Shumaker, D. L. (1999). Equity style timing. *Financial Analysts Journal* 55: 37–48.
- Philpotts, D., Zhu, M., and Stevenson, M. J. (2011). The benefits of tree-based models for stock selection. Working paper.
- Sorensen, E. H., Mezrich, J. J., and Miller, K. L. (1998). A new technique for tactical asset allocation. In F. J. Fabozzi, ed., *Active Equity Portfolio Management*, Chapter 12. New Hope, PA: Frank J. Fabozzi Associates.
- Sorensen, E. H., Miller, K. L., and Ooi, C. K. (2000). The decision tree approach to stock selection. *Journal of Portfolio Management*, Fall: 42–52.
- Zhu, M., Philpotts, D., and Stevenson, M. J. (2011). A hybrid approach to combining CART and logistic regression for stock ranking. *Journal of Portfolio Management*, Fall: 100–109.

Applying Cointegration to Problems in Finance

BALA ARSHANAPALLI, PhD

Professor of Finance, Indiana University Northwest

WILLIAM NELSON, PhD

Professor of Finance, Indiana University Northwest

Abstract: Financial time series data tend to exhibit stochastic trends. To uncover relationships among financial variables it is important to model changes in stochastic trends over time. Cointegration can be used to identify common stochastic trends among different financial variables. If financial variables are cointegrated, it can also be shown that the variables exhibit a long-run relationship. If this long-run relationship is severed, this may indicate the presence of a financial bubble.

The long-term relationships among economic variables, such as short-term versus long-term interest rates, or stock prices versus dividends, have long interested finance practitioners. For certain types of trends, multiple regression analysis needs modification to uncover these relationships. A trend represents a long-term movement in the variable. One type of trend, a *deterministic trend*, has a straightforward solution. Since a deterministic trend is a function of time, we merely include this time function in the regression. For example, if the variables are increasing or decreasing as a linear function of time, we may simply include time as a variable in the regression equation. The issue becomes more complex when the trend is stochastic. A stochastic trend is defined (Stock and Watson, 2003) as “a persistent but random long-term movement of the variable over time.” Thus a variable with a stochastic trend may

exhibit prolonged long-run increases followed by prolonged long-run declines and perhaps another period of long-term increases.

Most financial theorists believe stochastic trends better describe the behavior of financial variables than deterministic trends. For example, if stock prices are rising, there is no reason to believe they will continue to do so in the future. Or, even if they continue to increase in the future, they may not do so at the same rate as in the past. This is because stock prices are driven by a variety of economic factors and the impact of these factors may change over time. One way of capturing these common stochastic trends is by using an econometric technique usually referred to as *cointegration*.

In this entry, we explain the concept of cointegration. There are two major ways of testing for cointegration. We outline both econometric methods and the underlying theory for each

method. We illustrate the first technique with an example of the first type of cointegration problem, testing market efficiency. Specifically, we examine the present value model of stock prices. We illustrate the second technique with an example of the second type of cointegration problem, examining market linkages. In particular, we test the linkage and the dynamic interactions among stock market indexes of three European countries. Finally, we also use cointegration to test for the presence of an asset price bubble. Specifically, we test for the possibility of bubbles in the real estate markets.

STATIONARY AND NONSTATIONARY VARIABLES AND COINTEGRATION

The presence of stochastic trends may lead a researcher to conclude that two economic variables are related over time when in fact they are not. This problem is referred to as spurious regression. For example, during the 1980s the U.S. stock market and the Japanese stock market were both rising. An ordinary least squares (OLS) regression of the U.S. Morgan Stanley Stock Index on the Japanese Morgan Stanley Stock Index (USD) for the time period 1980–1990 using monthly data yields

$$\text{Japanese Stock Index} = 76.74 + 19 \text{ U.S.}$$

Stock Index

$$t\text{-statistic } (-13.95) \quad (26.51) \quad R\text{-square} = 0.86$$

The t -statistic on the slope coefficient (26.51) is quite large, indicating a strong positive relationship between the two stock markets. This strong relationship is reinforced with a very high R-square value. However, estimating the same regression over a different time period, 1990–2007, reveals

$$\text{Japanese Stock Index} = 2905.67 - 0.29 \text{ U.S.}$$

Stock Index

$$t\text{-statistic } (30.54) \quad (-2.80) \quad R\text{-square} = 0.04$$

This regression equation suggests there is a strong negative relationship between the two stock market indexes. Although the t -statistic on the slope coefficient (2.80) is large, the low R-square value suggests that the relationship is rather weak.

The reason behind these contradictory results is the presence of stochastic trends in both series. During the first time span, these stochastic trends were aligned, but not during the latter time span. Since different economic forces influence the stochastic trends and these forces change over time, during some periods they will line up and in some periods they will not. In summary, when the variables have stochastic trends, the OLS technique may provide misleading results—the spurious regression problem.

Another problem is that when the variables contain a stochastic trend, the t -values of the regressors no longer follow a normal distribution, even for large samples. Standard hypothesis tests are no longer valid for these nonnormal distributions.

At first, researchers attempted to deal with these problems by removing the trend through differencing these variables. That is, they focused on the change in these variables, $X_t - X_{t-1}$, rather than the level of these variables, X_t . Although this technique was successful for univariate Box-Jenkins analysis, there are two problems with this approach in a multivariate scenario. First, we can only make statements about the changes in the variables rather than the level of the variables. This will be particularly troubling if our major interest is the level of the variable. Second, if the variables are subject to a stochastic trend, then focusing on the changes in the variables will lead to a specification error in our regressions.

The cointegration technique allows researchers to investigate variables that share the same stochastic trend and at the same time avoid the spurious regression problem. Cointegration analysis uses regression analysis to study the long-run linkages among economic

variables and allows us to consider the short-run adjustments to deviations from the long-run equilibrium.

The use of cointegration in finance has grown significantly. Surveying this vast literature would take us beyond the scope of this entry. To narrow our focus, we note that cointegration analysis has been used mainly for two types of problems in finance. First, it has been used to evaluate the efficiency of financial markets in a wide variety of contexts. For example, it was used to evaluate the purchasing power parity theory (see Enders, 1988), the rational expectations theory of the term structure, the present value model of stock prices (Campbell and Shiller, 1987), and the relationship between the forward and spot exchange rates (Liu and Maddala, 1992). The second type of cointegration study investigates market linkages. For example, Hendry and Juselius (2000) examine how gasoline prices at different stations are linked to the world price of oil. Arshanapalli and Doukas (1993) investigate the linkages and dynamic interactions among stock market indexes of several countries.

Before explaining cointegration it is first necessary to distinguish between stationary and nonstationary variables. A variable is said to be stationary (more formally, weakly stationary) if its mean and variance are constant and its autocorrelation depends on the lag length, that is

$$E(X_t) = \mu, \quad \text{Var}(X_t) = \sigma^2, \quad \text{and} \\ \text{Cov}(X_t, X_{t-1}) = \gamma(l)$$

Stationary means that the variable fluctuates about its mean with constant variation. Another way to put it is that the variable exhibits mean reversion and so displays no stochastic trend. In contrast, nonstationary variables may wander arbitrarily far from the mean. Thus, only nonstationary variables exhibit a stochastic trend.

The simplest example of a nonstationary variable is a random walk. A variable is a random walk if $X_t = X_{t-1} + e_t$ where e_t is a

random error term with mean 0 and standard deviation σ . It can be shown that the standard deviation $\sigma(X_t) = t\sigma$ (see Stock and Watson, 1993), where t is time. Since the standard deviation depends on time, a random walk is nonstationary.

Nonstationary time series are often referred to as a *unit root* series. The *unit root* reflects the coefficient of the X_{t-1} term in an autoregressive relationship of order one. In higher-order autoregressive models, the condition of nonstationarity is more complex. Consider the p order autoregressive model where the a_i terms are coefficients and the L^i is the lag operator. If the sum of polynomial coefficients equals 1, then the X_t series are nonstationary and again are referred to as a unit root process.

$$(1 - a_1L^1 - \dots - a_pL^p)X_t = e_t + a_0 \quad (1)$$

If all the variables under consideration are stationary, then there is no spurious regression problem and standard OLS applies. If some of the variables are stationary, and some are nonstationary, then no economically significant relationships exist. Since nonstationary variables contain a stochastic trend, they will not exhibit any relationship with the stationary variables that lack this trend. The spurious regression problem occurs only when all the variables in the system are nonstationary.

If the variables share a common stochastic trend, we may overcome the spurious regression problem. In this case, cointegration analysis may be used to uncover the long-term relationship and the short-term dynamics. Two or more nonstationary variables are cointegrated if there exists a linear combination of the variables that is stationary. This suggests cointegrated variables share long-run links. They may deviate in the short run but are likely to get back to some sort of equilibrium in the long run. The term "equilibrium" is not the same as used in economics. To economists equilibrium means the desired amount equals the actual amount, and there is no inherent tendency to change. In contrast, equilibrium in cointegration analysis

means that if variables are apart, they show a greater likelihood to move closer together than further apart.

More formally, consider two-time series x_t and y_t . Assume that both series are nonstationary and integrated order one. (Integrated order one means that if we difference the variable one time, the resultant series is stationary.) These series are cointegrated if $z_t = x_t - ay_t$, z_t is stationary for some value of a . In the multivariate case, the definition is similar with vector notation. Let A and Y be vectors (a_1, a_2, \dots, a_n) and $(y_{1t}, y_{2t}, \dots, y_{nt})'$. Then the variables in Y are cointegrated if each of the $y_{1t} \dots y_{nt}$ are nonstationary and $Z = AY$, Z is stationary. A represents a cointegrating vector.

Cointegration represents a special case. We should not expect most nonstationary variables to be cointegrated. If two variables lack cointegration, then they do not share a long-run relationship or a common stochastic trend because they can move arbitrarily far away from each other. In terms of the present value model of stock prices, suppose stock prices and dividends lack cointegration. Then stock prices could rise arbitrarily far above the level of their dividends. This would be consistent with a stock market bubble (see Gurkaynak, 2005, for a more rigorous discussion of cointegration tests of financial market bubbles) and even if it is not a bubble, it is still inconsistent with the efficient market theory. In terms of stock market linkages, if the stock price indexes of different countries lack cointegration, then stock prices can wander arbitrarily far apart from each other. This possibility should encourage international portfolio diversification.

TESTING FOR COINTEGRATION

There are two popular methods of testing for cointegration: the Engle-Granger tests and the Johansen-Juselius tests. We discuss and illustrate both in the remainder of this entry.

Engle-Granger Cointegration Tests

The Engle-Granger cointegration test, developed by Engle and Granger (1987), involves the following four-step process:

Step 1

First determine whether the time series variables under investigation are stationary. We may consider both informal and formal methods. Informal methods entail an examination of a graph of the variable over time and an examination of the autocorrelation function. The autocorrelation function describes the autocorrelation of the series for various lags. The correlation coefficient between x_t and x_{t-i} is called the lag- i autocorrelation. For nonstationary variables, the lag one autocorrelation coefficient should be very close to one and decay slowly as the lag length increases. Thus, examining the autocorrelation function allows us to determine the *stationarity* of a variable. This method is not perfect. For stationary series that are very close to unit root processes, the autocorrelation function may exhibit the slow-fading behavior described above. If more formal methods are desired, the researcher may employ the Dickey-Fuller statistic, the augmented Dickey-Fuller statistic, or the Phillips-Perron statistic. These statistics test the hypothesis that the variables have a unit root, against the alternative that they do not (Dickey and Fuller, 1979, 1981; Phillips and Perron, 1988). The Phillips-Perron test makes weaker assumptions than both Dickey-Fuller statistics and is generally considered more reliable (Phillips and Perron, 1988). If it is determined that the variable is nonstationary and the differenced variable is stationary, proceed to step 2.

Step 2

Estimate the following regression:

$$y_t = c + dx_t + z_t \quad (2)$$

To make this concrete, let y_t represent some U.S. stock market index, x_t represents stock dividends on that stock market index, and z_t is the error term. c and d are regression parameters. For cointegration tests, the null hypothesis states that the variables lack cointegration, and the alternative claims that they are cointegrated.

Step 3

To test for cointegration, we test for stationarity in z_t . The Dickey-Fuller test is the most obvious candidate. That is, we should consider the following *autoregression* of the error term:

$$\Delta z_t = pz_{t-1} + u_t \tag{3}$$

where z_t is the estimated residual from equation (2). The test focuses on the significance of the estimated p . If the estimate of p is statistically negative, we conclude that the residuals, z_t , are stationary and reject the hypothesis of no cointegration.

The residuals of equation (3) should be checked to ensure they are white noise. If they are not, we should employ the augmented Dickey-Fuller test (ADF). The augmented Dickey-Fuller test is analogous to the Dickey-Fuller test but includes additional lags of Δz_t as shown in equation (4). The ADF test for stationarity, like the Dickey-Fuller test, tests the hypothesis of $p = 0$ against the alternative hypothesis of $p < 0$ for the equation (4):

$$\Delta z_t = pz_{t-1} + a_1\Delta z_{t-1} + \dots + a_n\Delta z_{t-n} + u_t \tag{4}$$

Generally, the OLS-produced residuals tend to have as small a sample variance as possible, thereby making residuals look as stationary as possible. Thus, the standard t -statistic or ADF statistic may reject the null hypothesis of nonstationarity too often. Hence, it is important to have correct statistics; fortunately, Engle and Yoo (1987) provide the correct statistics. Furthermore, if it is believed that the variable under investigation has a long-run growth

component, it is appropriate to test the series for stationarity around a deterministic time trend for both the DF and ADF tests. This is accomplished by adding a time trend to equations (3) or (4).

Step 4

The final step involves estimating the error-correction model. Engle and Granger (1987) showed that if two variables are cointegrated, then these variables can be described in an *error-correction* format described in the following two equations:

$$\Delta y_t = b_{10} + \sum_{i=1}^n b_{1i} \Delta y_{t-i} + \sum_{j=1}^n c_{1j} \Delta x_{t-j} + d_1(y_{t-1} - ax_{t-1}) + e_{1t} \tag{5}$$

$$\Delta x_t = b_{20} + \sum_{i=1}^n b_{2i} \Delta y_{t-i} + \sum_{j=1}^n c_{2j} \Delta x_{t-j} + d_2(y_{t-1} - ax_{t-1}) + e_{2t} \tag{6}$$

Equation (5) tells us that the changes in y_t depend on its own past changes, the past changes in x_t , and the disequilibrium between x_{t-1} and y_{t-1} ($y_{t-1} - ax_{t-1}$). The size of the error-correction term, d_1 , captures the speed of adjustment of x_t and y_t to the previous period's disequilibrium. Equation (6) has a corresponding interpretation.

The appropriate lag length is found by experimenting with different lag lengths. For each lag the Akaike information criterion (AIC), the Bayes information criterion, or the Schwarz information criterion is calculated and the lag with the lowest value of the criteria is employed.¹

The value of $(y_{t-1} - ax_{t-1})$ is estimated with the residuals from the cointegrating equation (3), z_{t-1} . This procedure is only legitimate if the variables are cointegrated. The error-correction term, z_{t-1} , will be stationary by definition if and only if they are cointegrated. The remaining terms in the equation, the lag difference of each variable, are also stationary because the levels

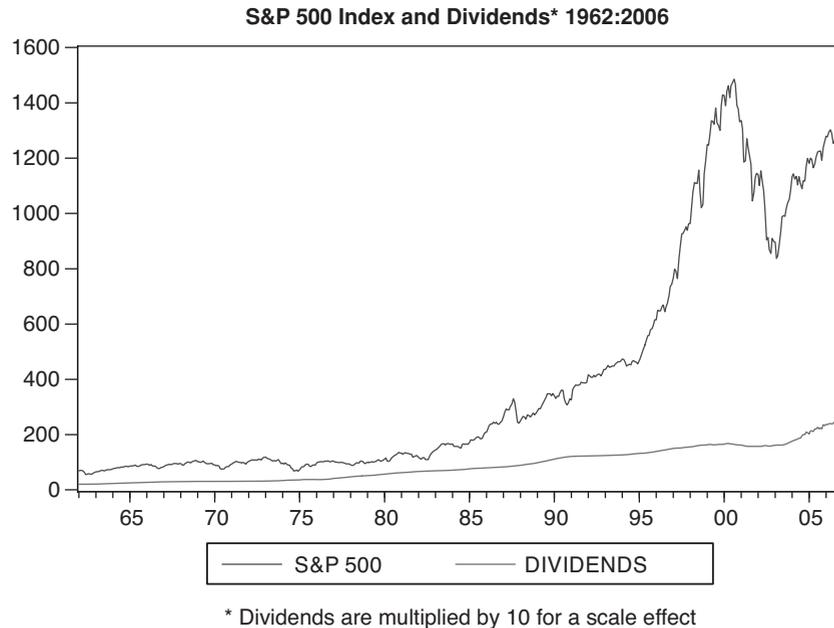


Figure 1

were assumed nonstationary. This guarantees the stationarity of all the variables in equations (5) and (6) and justifies the use of OLS.

Empirical Illustration Using the Dividend Growth Model

The dividend growth model of stock price valuation claims the fundamental value of a stock is determined by the present value of its future dividend stream. This model may be represented as:

$$P_0 = \sum d_i / (1 + r)$$

where

P_0 is the current stock price

d_i is a dividend in period i

r is the discount rate

If the discount rate exceeds the growth rate of dividends and the discount rate remains constant over time, then one can test for cointegration between stock prices and dividends. In brief, if the present value relationship holds, one

does not expect stock prices and dividends to meander arbitrarily far from each other.

Before starting any analysis it is useful to examine the plot of the underlying time series variables. Figure 1 presents a plot of stock prices and dividends for the years 1962 through 2006. The stock prices are represented by the S&P 500 index and the dividends represent the dividend received by the owner of \$1,000 worth of the S&P 500 index. The plot shows that the variables move together until the early 1980s. As a result of this visual analysis, we will entertain the possibility that the variables were cointegrated until the 1980s. After that, the common stochastic trend may have dissipated. We will first test for cointegration in the 1962–1982 period and then for the whole 1962–2006 period.

In accordance with the first step of the cointegration protocol, we must first establish the nonstationarity of the variables. To identify nonstationarity, we will use both formal and informal methods. The first informal test consists of analyzing the plot of the series shown in Figure 1. Neither series exhibits mean reversion. The dividend series wanders less

Table 1 Auto Correlation Functions of the S&P 500 Index and Dividends

Lag Auto Correlation	1	2	3	4	5	6	7	8	9	10	11	12
S&P 500	.993	.986	.979	.973	.967	.961	.954	.948	.940	.933	.926	.918
Dividend	.991	.983	.974	.966	.958	.979	.941	.933	.925	.916	.908	.900
Lag Auto Correlation	13	14	15	16	17	18	19	20	21	22	23	24
S&P 500	.911	.903	.896	.889	.881	.874	.866	.858	.851	.843	.835	.827
Dividend	.891	.883	.876	.868	.860	.852	.845	.837	.830	.822	.815	.808
Lag Auto Correlation	25	26	27	28	29	30	31	32	33	34	35	36
S&P 500	.819	.811	.804	.796	.789	.782	.775	.768	.761	.754	.748	.741
Dividend	.801	.794	.788	.781	.775	.769	.763	.758	.753	.747	.743	.738

from its mean than the stock prices. Nevertheless, neither series appears stationary.

The second informal method involves examining the autocorrelation function. We present in Table 1 the autocorrelation function for 36 lags of the S&P 500 index and the dividends for the 1962–2006 period using monthly data. The autocorrelations for the early lags are quite close to one. Furthermore, the autocorrelation function exhibits a slow decay at higher lags. This provides sufficient evidence to conclude that stock prices and dividends are nonstationary. When we inspect the autocorrelation function of their first differences (not shown), the autocorrelation of the first lag is not close to one. We may conclude the series are stationary in the first differences.

In Table 2, we present the results of formal tests of nonstationarity. The lag length for the ADF test was determined by the Schwarz criterion. The null hypothesis is that the S&P 500

stock index (dividends) contains a unit root; the alternative is that it does not. For both statistics, the ADF and the Phillips-Perron, the results indicate that the S&P 500 index is nonstationary and the changes in that index are stationary. The results for dividends are mixed. The ADF statistic supports the presence of a unit root in dividends, while the Phillips-Peron statistic does not. Since both the autocorrelation function and the ADF statistic conclude there is a unit root process, we shall presume that the dividend series is nonstationary. In sum, our analysis suggests that the S&P 500 index and dividends series each contain a stochastic trend in the levels, but not in their first differences.

In the next step of the protocol we examine whether the S&P 500 index and dividends are cointegrated. This is accomplished by estimating the long-run equilibrium relation by regressing the logarithm (log) of the S&P 500 index on the log of the dividends. We use the

Table 2 Stationarity Test for the S&P 500 Index and Dividends 1962–2006

Variable	Augmented Dickey Fuller (ADF)	Phillips-Perron	Critical Value of Test Statistics at 1%, 5%, 10% Significance
S&P 500	1.22	1.12	-3.44 (1%)
Δ S&P 500	-19.07	-19.35	-2.87 (5%)
Dividends	1.52	4.64	-2.56 (10%)
Δ Dividends	-2.13	-31.68	

Null hypothesis: Variable is nonstationary.

The lag length for the ADF test was determined by the Schwarz Criterion. For the S&P 500 index and its first difference, the lag length was 1. For the dividends and its first difference, the lag lengths were 12 and 11, respectively.

Table 3 Cointegration Regression: S&P 500 and Dividends
 $\text{Log S\&P 500} = a + b \log \text{dividends} + z_t$

Period	Constant	Coefficient of Dividends	<i>t</i> -Stat Dividends
1962–1982	4.035	.404	17.85
1962–2006	2.871	1.336	68.54

logarithms of both variables to help smooth the series. The results using monthly data are reported in Table 3 for both the 1962–1982 period and the 1962–2006 period. We pay little attention to the high *t*-statistic on the dividends variable because the *t*-test is not appropriate unless the variables are cointegrated. This is, of course, the issue.

Once we estimate the regression in step 2, the next step involves testing the residuals of the regression, z_t , for stationarity. By definition, the residuals have a zero mean and lack a time trend. This simplifies the test for stationarity. This is accomplished by estimating equation (4). The null hypothesis is that the variables lack cointegration. If we conclude that p in equation (4) is significantly negative, then we reject the null hypothesis and conclude that the evidence is consistent with the presence of cointegration between the stock index and dividends.

The appropriate lag lengths may be determined by the Akaike information criterion or theoretical and practical considerations. We decided to use a lag length of three periods representing one quarter. The results are presented in Table 4. For the 1962–1982 period, we may reject the null hypothesis of no cointegration at the 10% level of statistical significance. For the entire period (1962–2006), we cannot reject the null hypothesis ($p = 0$) of no cointegration. Apparently, the relationship between stock prices and dividends unraveled in the 1980s and the 1990s. This evidence is consistent with the existence of an Internet stock bubble in the 1990s.

Having established that the S&P 500 index and dividends are cointegrated from 1962–1982, in the final step of the protocol we examine the interaction between stock prices and dividends by estimating the error-correction model, equations (5) and (6). It is useful at this point to

Table 4 Augmented Dickey Fuller Tests of Residuals for Cointegration

Variable	Coefficient	<i>t</i> -Stat	<i>p</i> -Value
Panel A 1962–1982 $n = 248$			
z_t	-.063	-3.23	.001
Δz_{t-1}	.272	4.32	.000
Δz_{t-2}	-.030	-.46	.642
Δz_{t-3}	.090	1.40	.162
<i>t</i> -statistic of $p = -3.23$; critical values (5%) -3.36 (10%) -3.06			
Panel B 1962–2006 $n = 536$			
z_t	-.008	-1.81	.070
Δz_{t-1}	.265	6.13	.000
Δz_{t-2}	-.048	-1.08	.280
Δz_{t-3}	.031	.71	.477
<i>t</i> -statistic of $p = -1.81$; critical values (5%) 3.35 (10%) 3.05			

The critical values of the Augmented Dickey Fuller (ADF) statistic are from Engle and Yoo (1987). The cointegration equation errors used to perform the ADF test is based on the following regression:

$$\Delta z_t = -p z_{t-1} + a \Delta z_{t-1} + b \Delta z_{t-2} + c \Delta z_{t-3} + e_t$$

where Δz_t is the change in the error term from the co-integration regression and e_t is a random error. If p is positive and significantly different from zero, the z residuals from the equilibrium equation are stationary so we may accept the null hypothesis of cointegration. In both equations the error terms are white noise, so no further stationarity tests were performed.

review our interpretation of equations (5) and (6). Equation (5) claims that changes in the S&P 500 Index depend upon past changes in the S&P 500 Index and past changes in dividends and the extent of disequilibrium between the S&P 500 index and dividends. Equation (6) has a similar statistical interpretation. However, from a theoretical point of view, equation (6) is meaningless. Financial theory does not claim that changes in dividends are impacted either by past changes in stock prices or the extent of the disequilibrium between stock prices and dividends. As such, equation (6) degenerates into an autoregressive model of dividends.

We estimated the error-correction equations using three lags. The error term, z_{t-1} , used in these error-correction regressions was obtained from OLS estimation of the cointegration equation reported in Table 3. Estimates of the error-correction equations are reported in Table 5. By construction, the error-correction term represents the degree to which the stock prices and dividends deviate from long-run equilibrium. The error-correction term is included in both equations to guarantee that the variables do not drift too far apart. If the variables are cointegrated, Engle and Granger (1987) showed that the coefficient on the error-correction term ($y_{t-1} - ax_{t-1}$) in at least one of the equations must be nonzero. The t value of the error-correction term in equation (5) is statistically different

from zero. The coefficient of -0.07 is known as the speed of adjustment coefficient. It suggests that 7% of the previous month's disequilibrium between the stock index and dividends is eliminated in the current month. In general, the higher the speed of adjustment coefficient, the faster the long-run equilibrium is restored. Since the speed of adjustment coefficient for the dividend equation is statistically indistinguishable from zero, all of the adjustment falls on the stock price.

An interesting observation from Table 5 relates to the lag structure of equation (5). The first lag on past stock price changes is statistically significant. This means that the change in the stock index this month depends upon the change during the last month. This is inconsistent with the efficient market hypothesis. On the other hand, the change in dividend lags is not statistically different from zero. The efficient market theory suggests, and the estimated equation confirms, that past changes in dividends do not affect the current changes in stock prices.

Johansen-Juselius Cointegration Tests

The Engle-Granger method does have some problems (see Enders, 1995). These problems are magnified in a multivariate (three or more variables) context. In principle, when the

Table 5 Error Correction Model: S&P 500 Index and Dividends 1962-1982

$$\Delta Y_t = b_{01} + b_{11}\Delta Y_{t-2} + b_{12}\Delta Y_{t-2} + b_{13}\Delta Y_{t-3} + c_{11}\Delta X_{t-1} + c_{12}X_{t-2} + c_{13}\Delta X_{t-3} + d_1(Y_{t-1} - a X_{t-1}) + e_{1t} \quad (5)$$

$$\Delta X_{t-1} = b_{20} + b_{21}\Delta Y_{t-1} + b_{22}\Delta Y_{t-2} + b_{23}\Delta Y_{t-3} + c_{21}\Delta X_{t-1} + c_{22}\Delta X_{t-2} + c_{23}\Delta X_{t-3} + d_2(Y_{t-1} - a X_{t-1}) + e_{2t} \quad (6)$$

	Equation 5		Equation 6		
	Coefficient	t-stat	Coefficient	t-stat	
b ₀₁	-.009	-2.42	b ₂₀	.001	2.91
b ₁₁	.251	4.00	b ₂₁	.002	.63
b ₁₂	-.043	-.66	b ₂₂	-.003	-.88
b ₁₃	.081	1.27	B ₂₃	.004	1.07
c ₁₁	.130	.11	c ₂₁	.939	14.60
c ₁₂	-.737	-.46	c ₂₂	-.005	-.06
c ₁₃	-.78	-0.65	c ₂₃	-.006	.87
d ₁	-.07	-3.64	d ₂	.000	.30

The change in the S&P 500 index is denoted as ΔY_t and the change in dividends is denoted as ΔX_t .

cointegrating equation is estimated (even in a two-variable problem), we may use any variable as the dependent variable. In our last example, this would entail placing dividends on the left-hand side of equation (2) and the S&P 500 index on the right-hand side. As the sample size approaches infinity, Engle and Granger (1987) showed the cointegration tests produce the same results irrespective of what variable you use as the dependent variable. The question then is: How large a sample is large enough?

A second problem is that the errors we use to test for cointegration are only estimates and not the true errors. Thus any mistakes made in estimating the error term, z_t , in equation (2) are carried forward into the equation (3) regression. Finally, the Engle-Granger procedure is unable to detect multiple cointegrating relationships.

The procedures developed by Johansen and Juselius (1990) avoid these problems. Consider the following multivariate model:

$$y_t = Ay_{t-1} + u_t \quad (7)$$

where

y_t is an $n \times 1$ vector ($y_{1t}, y_{2t}, \dots, y_{nt}$)'

u_t is an n -dimensional error term at t

A is an $n \times n$ matrix of coefficients

If the variables display a time trend, we may wish to add the matrix A_0 to equation (7). This would reflect a deterministic time trend. The same applies to equation (8) presented below. It does not change the nature of our analysis.

The model (without the deterministic time trend) can then be represented as:

$$\Delta y_t = (I - A)y_{t-1} + u_t \quad (8)$$

Let $B = I - A$. I is the identity matrix of dimension n . The cointegration of the system is determined by the rank of B matrix. The highest rank of B one can obtain is n , the number of variables under consideration. If the rank of the matrix equals zero, then the B matrix is null. This means $\Delta y_t = 0 + u_t$, where 0 is the null vector. In this case y_{it} will follow a random walk

($y_t = y_{t-1} + u_t$) and no linear combination of y_t will be stationary, so there are no cointegrating vectors.

If the rank of B is n , then each y_{it} is an autoregressive process. This means each y_{it} is stationary and thus they cannot be cointegrated. For any rank between 1 and $n - 1$, the system is cointegrated and the rank of the matrix is the number of cointegrating vectors.

The higher-order autoregressive representation is similar. Although it is more involved, the Johansen and Juselius estimation procedure can still handle it easily. Since the rank of a matrix equals the number of distinct nonzero characteristic roots of a matrix, the Johansen-Juselius procedure attempts to determine the number of nonzero characteristic roots of the relevant matrices. The procedure estimates the matrices and hence the characteristic roots with a maximum likelihood method.

The Johansen-Juselius procedure employs two statistics to test for nonzero characteristic roots. First they order the characteristic roots from high to low, $\lambda_1^* > \lambda_2^* > \dots > \lambda_n^*$. λ_i^* to estimate nonzero characteristic roots.

The first statistic, the trace test statistic, verifies the null hypothesis that at most i characteristic roots are different from zero. The alternative hypothesis is that more than i characteristic roots are nonzero. The statistic employed is:

$$\lambda_{\text{trace}}(i) = -T[\ln(1 - \lambda_i^*) + \ln(1 - \lambda_{i+1}^*) + \dots + \ln(1 - \lambda_n^*)]. \quad (9)$$

where T is the number of included time periods. If all the characteristic roots are zero since $\ln(1) = 0$, the statistic will equal zero. Thus low values of the test statistic will lead us to fail to reject the null hypothesis. The larger any characteristic root is, the more negative $1 - \lambda_i^*$ and the larger the test statistic and the more likely we will reject the null hypothesis.

The alternative test is called the maximum eigenvalue test since it is based on the largest eigenvalue. This statistic tests the null hypothesis that there are i cointegrating vectors against

the alternative hypothesis of $i + 1$. This statistic is:

$$\lambda_{\max}(i, i + 1) = -T \ln(1 - \lambda_{i+1}^*) \quad (10)$$

Again, if $\lambda_{i+1}^* = 0$, then the test statistic will equal zero. So low (high) values of λ_{i+1}^* will lead to a failure to reject (rejection of) the null hypothesis.

Johansen and Juselius derive critical values for both test statistics. The critical values are different if there is a deterministic time trend and an A_0 matrix is included. Enders (1995) provides tables for both critical statistics with and without the trend terms. Software programs often provide critical values and the relevant p -values.

Testing of the Dynamic Relationships among Country Stock Markets

Many portfolio managers seek international diversification. If stock market returns in different countries were not highly correlated, then portfolio managers could obtain risk reduction without significant loss of return by investing in different countries. But with the advent of globalization and the simultaneous integration of capital markets, the risk-diversifying benefits of international investing have been subject to challenge. In this section, we illustrate how cointegration can shed light on this issue and apply the Johansen-Juselius technique.

The idea of a common currency for the European countries is to reduce transactions costs and more closely link the economies. We shall use cointegration to examine whether the stock markets of France, Germany, and the Netherlands are linked following the introduction of the Euro in 1999. We use monthly data for the period 1999–2006.

The first step to test for cointegration is to establish that the three stock indexes are nonstationary in the levels and stationary in the first differences. In testing the present value model, we presented the autocorrelation function (the

ADF statistic), and the Phillips-Perron statistic. For reasons of space, we will not repeat this. Next we should establish the appropriate lag length for equation (8). This is typically done by estimating a traditional *vector autoregressive* (VAR) model and applying a multivariate version of the Akaike information criterion or Schwarz criterion. For our model, we use one lag, and thus the model takes the form:

$$y_t = A_0 + A_1 y_{t-1} + u_t \quad (11)$$

where y_t is the $n \times 3$ vector $(y_{1t}, y_{2t}, y_{3t})'$ of the logs of the stock market index for France, Germany, and the Netherlands (i.e., element y_{1t} is the log of the French index at time t ; y_{2t} is the log of the German index at time t ; and y_{3t} is the log of the Netherlands index at time t). We use logs of the stock market indexes to smooth the series. A_0 and A_1 are $n \times n$ matrices of parameters and u_t is the $n \times n$ error matrix.

The next step is to estimate the model. This means fitting equation (8). We incorporated a linear time trend, hence the inclusion of the matrix A_0 . Since there are restrictions across the equations, the procedure uses a maximum likelihood estimation procedure and not OLS. The focus of this estimation is not on the parameters of the A matrices. Few software programs present these estimates; rather, the emphasis is on the characteristic roots of the matrix B , which are estimated to determine the rank of the matrix.

The estimates of the characteristic roots are presented in Table 6. We want to establish whether i indexes are cointegrated. Thus, we test the null hypothesis that the indexes lack cointegration. To accomplish this, the λ_{trace} (0) statistic is calculated. Table 6 also provides this statistic. To ensure comprehension of this important statistic, we detail its calculation.

We have 96 usable observations.

$$\begin{aligned} \lambda_{\text{trace}}(0) &= -T[\ln(1 - \lambda_1^*) + \ln(1 - \lambda_2^*) \\ &\quad + \ln(1 - \lambda_3^*)] = -96[\ln(1 - 0.227) \\ &\quad + \ln(1 - 0.057) + \ln(1 - 0.028)] \\ &= 33.05 \end{aligned}$$

Table 6 Cointegration Test

Hypothesized No. of Cointegrating Vectors	Characteristic Roots	Trace Statistics λ_{trace}	5% Critical Value	p -Value	Max Statistic λ_{max}	5% Critical Value	p -Value
None	.227	33.05	29.80	.02	24.72	21.13	.01
At most 1	.057	8.32	15.49	.43	5.61	14.26	.66
At most 2	.028	2.72	3.84	.10	2.72	3.84	.10

As reported in Table 6, this exceeds the critical value for 5% significance of 29.80 and has a p -value of 0.02. Thus, we may reject the null hypothesis at a 5% level of significance and conclude that the evidence is consistent with at least one cointegrating vector. Next we can examine $\lambda_{\text{trace}}(1)$ to test the null hypothesis of at most 1 cointegrating vector against the alternative of 2 cointegrating vectors. Table 6 shows that λ_1 at 8.33 is less than the critical value of 15.49 necessary to establish statistical significance at the 5% level. We do not reject the null hypothesis. We therefore conclude that there is at least one cointegrating vector. There is no need to evaluate $\lambda_{\text{trace}}(2)$.

The λ_{max} statistic reinforces our conclusion. We can use $\lambda_{\text{max}}(0, 1)$ to test the null hypothesis that the variables lack cointegration against the alternative that they are cointegrated with one cointegrating vector. Table 6 presents the value of $\lambda_{\text{max}}(0, 1)$. Again, for pedagogic reasons we outline the calculation of $\lambda_{\text{max}}(0, 1)$.

$$\begin{aligned}\lambda_{\text{max}}(0, 1) &= (-T \ln(1 - \lambda_i^*)) = -96 \ln(1 - 0.227) \\ &= 24.72\end{aligned}$$

The computed value of 24.72 exceeds the critical value of 21.13 at the 5% significance level and has a p -value of 0.01. Once again, this leads us to reject the null hypothesis that the indexes lack cointegration and conclude that there exists at least one cointegrating vector.

The next step requires a presentation of the cointegrating equation and an analysis of the error-correction model. Table 7 presents both. The cointegrating equation is a multivariate representation of z_{t-1} in the Engle-Granger

method. This is presented in panel A of Table 7. The error-correction model takes the following representation.

$$\begin{aligned}\Delta y_t &= b_{10} + \sum_{i=1}^n b_{1i} \Delta y_{t-i} + \sum_{j=1}^n c_{1j} \Delta x_{t-j} \\ &\quad + d_1(y_{t-1} - ax_{t-1}) + e_{1t}\end{aligned}\quad (12)$$

The notation of equation (12) differs somewhat from the notation of equations (5) and (6). The notation used in equation (12) reflects the matrix notation adopted for the Johansen-Juselius method in equation (8). Nevertheless, for expositional convenience, we did not use the matrix notation for the error-correction term. Again, the Δ means the first difference of the variable; thus Δy_{1t-1} means the change in the log of the French stock index in period $t - 1$, $(y_{1t-1} - y_{1t-2})$. Equation (12) claims that changes in the log of the French stock index are due to changes in the French stock index during the last two (2) periods; changes in the German stock index during the last two periods; changes in the Netherlands stock index during the last two periods; and finally deviations of the French stock index from its stochastic trend with Germany and the Netherlands. An analogous equation could be written for both Germany and the Netherlands.

Panel B of Table 7 presents the error-correction model estimates for each of the three countries. The software used a two-period lag for the past values of the changes in the stock indexes as indicated by the Schwarz criterion.

The error-correction term in each equation reflects the deviation from the long-run stochastic trend of that stock index in the last period.

Table 7 Cointegrating Equation and Error Correction Equations 1999–2007

Panel A: Cointegrating Equation

$$\text{France} = 4.82 + 2.13 \text{ Germany} - 1.71 \text{ Netherlands}$$

$$[-8.41] \quad [5.25]$$

Panel B: Error Correction Equations

Country	$\Delta(\text{France})$	$\Delta(\text{Germany})$	$\Delta(\text{Netherlands})$
Z_{t-1}	-0.151477 [-2.21470]	-0.057454 [-0.66835]	-0.179129 [-2.52373]
$\Delta(\text{France}(-1))$	0.087360 [0.27222]	0.245750 [0.60927]	0.225357 [0.67667]
$\Delta(\text{France}(-2))$	-0.200773 [-0.68179]	-0.218331 [-0.58990]	-0.324250 [-1.06105]
$\Delta(\text{Germany}(-1))$	-0.189419 [-0.82197]	-0.024306 [-0.08392]	-0.094891 [-0.39680]
$\Delta(\text{Germany}(-2))$	-0.155386 [-0.67237]	-0.109070 [-0.37551]	-0.127301 [-0.53081]
$\Delta(\text{Netherlands}(-1))$	0.079881 [0.34284]	-0.189775 [-0.64805]	-0.188295 [-0.77875]
$\Delta(\text{Netherlands}(-2))$	0.439569 [1.89288]	0.446368 [1.52936]	0.483929 [2.00810]
C	0.005967 [1.02860]	0.002575 [0.35321]	0.002688 [0.44641]

France (-1) represents the log return of the French stock index one month ago. Germany (-1) and Netherlands (-1) have a similar interpretation the [] represent the *t*-statistic.

It should be noted that in contrast to the two-step procedure of the Engle-Granger approach, the Johansen-Juselius approach estimates the speed of adjustment coefficient in one step. It provides insight into the short-run dynamics. This coefficient is insignificant (at the 5% level) for Germany. This means that stock prices in Germany do not change in response to deviations from their stochastic trend with France and the Netherlands. Because the variables are cointegrated, we are guaranteed that at least one speed of adjustment coefficient will be significant. In fact, the speed of adjustment coefficients of both France and the Netherlands attain statistical significance (at the 5% level) and are about the same size. This shows that when the economies of France and the Netherlands deviate from the common stochastic trend, they adjust. In France about 15% and in the Netherlands about 17% of the last-period deviation is corrected during this period.

For France, neither past changes in its own stock index nor the past changes in Germany's stock index appear to affect French stock prices.

The changes in the lagged values of both indexes lack statistical significance. Only the second lag of the Netherlands stock index attained significance. For Germany, the past changes in its own stock prices and the past changes in the stock indexes of the other countries failed to obtain significance at the 5% level. For the Netherlands, its own second-period lag obtained statistical significance. Nevertheless, the failure of individual lags to obtain significance does not mean that jointly the lags are insignificant.

To see this, we turn to an examination of *Granger causality* in the error-correction models. Granger causality helps us to classify the variables into dependent and independent. A variable Granger causes another variable when past values of that variable improve our ability to forecast the original variable. To test for Granger causality, an *F*-test is employed to verify whether the lagged changes in, say, the stock index of France jointly zero in the German equation. Table 8 reports the results of pairwise Granger causality tests. We find that France and Germany do not Granger cause each other

Table 8 Cointegration Test Results 1975–2000

Cointegration between	Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob. **
Home Prices vs. Household Debt Ratio	None At most 1	0.09 0.05	14.72 5.28	25.87 12.52	0.60 0.56
Home Prices vs. Housing Affordability Index	None At most 1	0.13 0.07	20.34 7.06	25.87 12.52	0.21 0.34
Home Prices vs. Mortgage Rate	None At most 1	0.10 0.08	18.16 7.73	25.87 12.52	0.33 0.27
Home Prices vs. Homebuilders Stock Index	None At most 1	0.15 0.09	25.69 9.49	25.87 12.52	0.05 0.15
Home Prices vs. Unemployment Rate	None * At most 1	0.15 0.09	25.90 9.66	25.87 12.52	0.05 0.14
Home Prices vs. Mean of Middle Fifth of Income	None * At most 1	0.20 0.10	32.58 9.90	25.87 12.52	0.01 0.13
Home Prices vs. Mean of Top Fifth of Income	None * At most 1	0.21 0.08	32.29 8.70	25.87 12.52	0.01 0.20

Source: This table is reprinted from Arshanapalli and Nelson (2008) with permission of *The International Journal of Business and Finance Research*.

* Denotes rejection of the null hypothesis of no cointegration at the 0.05 level

** Denotes the p -value

at any conventional levels of significance. The smaller Netherlands economy finds its stock prices Granger caused by both France and Germany but the Netherlands does not Granger cause either French or German stock prices at conventional levels of significance.

Empirical Illustration of a Test for the Presence of a Housing Bubble

The third application demonstrates the use of cointegration to test the possibility of a bubble in the housing market. As we illustrated in our previous examples, the beginning of any analysis is a picture of the time series under examination. Figure 2 shows the trend in the U.S. housing index from 1975 to the third quarter of 2007. Clearly, since 2001 the United States has experienced several years of strong home price increases. Also, the figure illustrates that the rise began to slow in 2005. At this time we know that housing prices collapsed in 2008. This sort of evidence has led the financial and the general press to conclude that the U.S. housing market has experienced a bubble. However, the detection of a bubble after the fact is of little practical use. The question is, can cointegration

provide evidence of a bubble before the bubble bursts?

The widely accepted efficient market theory claims that financial asset prices reflect all the publicly available information at all times. This denies the possibility of a bubble. While some may believe prices are too high relative to fundamental factors, according to the theory they are wrong, because investors recognize immediately if the price of anything is too high (or too low) and respond by selling (or purchasing) the asset until the over-(under) pricing is eliminated. A mountainous body of academic research (see Fama, 1970, for a sampling) supports this view.

Nevertheless, the efficient market theory has been subject to much serious criticism (Shiller, 2003). Furthermore, much of the research focused on financial assets. The efficient market theory assumes that investors can sell an asset short to eliminate overpricing. Real estate is a real and illiquid asset. During the period of the housing price run-up there was no mechanism known to us for shorting a residential home. A futures market for housing is a relatively recent innovation. These markets do not function well enough to fulfill the assumptions of the

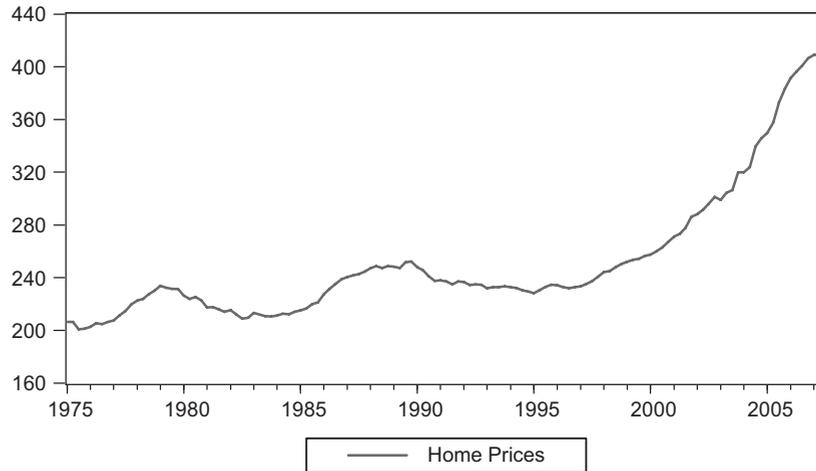


Figure 2 Home Prices in the United States

*The bursting of a real estate bubble has important implications for the U.S. economy. Residential real estate is an important component of householder wealth. In 1996, it represented 39% of household wealth. This figure is reprinted from Arshanapalli and Nelson (2008) with the permission of the Institute for Business and Finance Research.

efficient market theory. Thus, we should not dismiss the possibility of a housing bubble out of hand.

Arshanapalli and Nelson (2008) tested for the existence of a housing bubble, examining the stability of the underlying relationship of home prices and the economic forces that determine them. A relationship suddenly becomes unstable when rising home prices are not justified by the underlying economic fundamentals. Cointegration is well suited to test for this. Cointegration implies that two variables share a common stochastic trend. A common stochastic trend does not simply mean that they move upward or downward together, but rather that the variables may share both prolonged upward and prolonged downward movements.

Suppose housing prices are cointegrated with an economic variable and a bubble develops in the housing market, then housing prices rise without a corresponding rise in the variable. This implies the severing of a long-term relationship between housing prices and the variable. In other words, the cointegration should cease. In summary, if there were a housing bubble beginning in about 2000, then the variables,

which were cointegrated with housing prices before 2000, will no longer remain cointegrated after 2000.

Data

Quarterly data are used and the study covers the period 1975Q1–2007Q2. We employ the U.S. Office of Federal Housing Enterprise Oversight (OFHEO) Home Price quarterly index to measure housing prices. The index is not seasonally adjusted.

Next, we consider a series of seven variables that reflect the fundamental economic forces determining housing prices. The most important of these is income. Case and Shiller (2003) conclude that in nonbubble markets income explains most of the rise in housing prices. We employ two separate measures of income. The first is the mean of the middle quintile of the income distribution, denoted as the Middle Fifth. Second, we use the mean of the highest quintile of the income distribution, denoted as the Top Fifth. This attempts to account for the possibility that the wealthiest segment of the population influences housing prices

disproportionately because of their greater mobility. The U.S. Census Bureau, *Historical Income Tables-Families* (all races), and the National Association of Realtors provided these data.

The mortgage rate represents a strong influence on consumer demand for housing. We obtained the 30-year conventional mortgage rate (fixed rate, first mortgages) from the Board of Governors of the U.S. Federal Reserve System. The civilian unemployment rate measures the state of the economy. The U.S. Bureau of Labor Statistics provided the seasonally adjusted percentage of civilian unemployment. We converted the monthly data for both variables to quarterly data by a simple mean. The Homebuilders Stock Index provides an indication of the state of the housing market. A capitalization-weighted, price-level index of homebuilding stocks based on stocks included in the S&P 500 stock index was obtained from Merrill Lynch.

The final variables measure the ability of consumers to handle mortgage debt. The household debt ratio is the ratio of household credit market debt outstanding to annualized personal disposable income. The data also came from the Board of Governors of the U.S. Federal Reserve System. The Housing Affordability Index for all homebuyers (HAI) measures whether or not a typical family could qualify for a mortgage loan on a typical home, assuming a 20% down payment. We define a typical home as the national median-priced, existing single-family home as calculated by NAR. In its final form used here, the HAI is essentially "median family income divided by qualifying income." The index is interpreted as follows: A value of 100 means that a family with the median family income (from the U.S. Bureau of the Census and NAR) has exactly enough income to qualify for a mortgage on a median-priced home. National Association of Realtors (NAR) provided the data. In this research, the monthly HAI values result from quarterly samples.²

Again, the first step in establishing cointegration is to test the variables for stationarity. To establish nonstationarity we employed the ADF (augmented Dickey Fuller) test and the Phillips-Peron test. Although we do not display the results here, we conclude all the variables are nonstationary.

Next, we examine whether home prices and the seven fundamental variables are cointegrated. This is accomplished by examining a cointegrating regression for each of the seven variables with home prices. Table 8 presents the results of these cointegration tests for the 1975Q1–2000Q4 period. The Trace Statistic Test shows that for three of the seven variables, top fifth, middle fifth, and the unemployment rate, we may reject the null hypothesis of no cointegration at a 5% level of significance. Furthermore, we may reject the null hypothesis of no cointegration at the 10% level of statistical significance for one additional variable, the Homebuilders stock index. Thus for the period preceding the runup in home prices there appears to have been a strong link between home prices and both the income variables and the unemployment rate and a marginal link with the Homebuilders Stock Index.

Table 9 presents the results of these cointegration tests for the period 1975–2007Q3. The trace tests indicate the eigenvalues are not statistically distinguishable from zero in any equation at the 5% level. However, the P-value for the middle fifth of income was .0502. Recognizing the belief that the bubble burst in late 2005, we did the cointegration test for the period 1975–2005Q2. The P-value (hypothesized no. of CE(s) = none) for home prices vs. middle fifth of income was 11%. (Although we did not display the results, we cannot reject the hypothesis of no cointegration for any of the other fundamental variables during this period. This suggests that in the post-2005 period the normal relationship between home prices and income was reasserting itself. This result suggests that the linkage between home prices and fundamental variables has been substantially reduced

Table 9 Cointegration Test Results for the Whole Period: 1975–2007Q3

Cointegration between	Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob. *
Home Prices vs. Household Debt Ratio	None	0.07	7.68	15.49	0.50
	At most 1	0.00	0.29	3.84	0.59
Home Prices vs. Housing Affordability Index	None	0.12	13.64	15.49	0.09
	At most 1	0.01	0.89	3.84	0.35
Home Prices vs. Mortgage Rate	None	0.10	10.63	15.49	0.24
	At most 1	0.00	0.42	3.84	0.52
Home Prices vs. Homebuilders Stock Index	None	0.10	12.28	15.49	0.14
	At most 1	0.02	1.78	3.84	0.18
Home Prices vs. Unemployment Rate	None	0.10	12.13	15.49	0.15
	At most 1	0.02	1.81	3.84	1.18
Home Prices vs. Mean of Middle Fifth of Income	None	0.13	15.48	15.49	0.05
	At most 1	0.02	2.02	3.84	0.16
Home Prices vs. Mean of Top Fifth of Income	None	0.09	8.85	15.49	0.38
	At most 1	0.00	0.03	3.84	0.87

Source: This table is reprinted from Arshanapalli and Nelson (2008) with permission of *The International Journal of Business and Finance Research*.

* Denotes the *p*-value

after 2000. The evidence is consistent with a real estate bubble.

This is consistent with the presence of an asset price bubble.

KEY POINTS

- Many of the variables of interest to finance professionals are nonstationary.
- The relationships among them can be fruitfully analyzed if they share a common stochastic trend. A way of capturing this common stochastic trend is the application of cointegration.
- Cointegration analysis can reveal interesting long-run relationships between the variables.
- It is possible that cointegrating variables may deviate in the short run from their relationship, but the error correction model shows how these variables adjust to the long-run equilibrium.
- Cointegration analysis can reveal interesting short-run asset pricing adjustments.
- The error-correction models tend to have a better forecasting performance than simple vector autoregressive models.
- Cointegration analysis shows when fundamental long-run relationships are severed.

NOTES

1. For a summary of these criteria, see Chapter 12 in Focardi and Fabozzi (2004).
2. For more details on the exact calculation, go to www.realtor.org.

REFERENCES

- Arshanapalli, B., and Doukas, J. (1993). International stock market linkage: Evidence from the pre and post October 1987 period. *Journal of Banking and Finance* 17: 193–208.
- Arshanapalli, B., and Nelson, W. (2008). A cointegration test to verify the housing bubble. *The International Journal of Business and Finance Research* 2: 35–44.
- Case, K., and Shiller, R. (2003). Is there a bubble in the housing market. *Brooking Papers on Economic Activities* 2: 299–362.
- Dickey, D., and Fuller, W. A. (1979). Distribution of the estimates for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74: 427–431.
- Dickey, D., and Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* 49: 427–431.

- Enders, W. (1988). ARIMA and cointegration tests of purchasing power parity. *Review of Economics and Statistics* 70: 504–508.
- Enders, W. (1995). *Applied Econometric Time Series*. Hoboken, NJ: Wiley.
- Engle, R. F., and Granger, C. W. J. (1987). Cointegration and error-correction: Representation, estimation and testing. *Econometrica* 55: 251–276.
- Engle, R. F., and Yoo, S. (1987). Forecasting and testing in co-integrated systems. *Journal of Econometrics* 35: 143–159.
- Focardi, S. M., and Fabozzi, F. J. (2004). *The Mathematics of Financial Modeling and Investment Management*. Hoboken, NJ: John Wiley & Sons.
- Gurkaynak, R. (2005). Econometric tests of asset price bubbles: Taking stock. Finance and Economics Discussion Series, Federal Reserve Board. Washington, D.C.
- Hendry, D., and Juselius, K. (2000). Explaining cointegration analysis: Part II. Discussion Papers, University of Copenhagen, Department of Economics.
- Johansen, S., and Juselius, K. (1990). Maximum likelihood estimation and inference on cointegration with application to the demand for money. *Oxford Bulletin of Economics and Statistics* 52: 169–209.
- Liu, P., and Maddala, G. (1992). Using survey data to market efficiency in the foreign exchange market. *Empirical Economics*, 17: 303–314.
- Phillips, P., and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika* 75: 335–346.
- Shiller, R. (2003). From efficient markets theory to behavioral finance. *Journal of Economic Perspectives* 17: 83–104.
- Stock, J., and Watson, M. (2003). *Introduction to Econometrics*. New York: Addison Wesley.

Nonlinearity and Nonlinear Econometric Models in Finance

RUEY S. TSAY, PhD

H.G.B. Alexander Professor of Econometrics and Statistics,
University of Chicago Booth School of Business

Abstract: Many financial and economic data exhibit nonlinear characteristics. Prices of commodities such as crude oil often rise quickly but decline slowly. The monthly U.S. unemployment rate exhibits sharp increases followed by slow decreases. To model these characteristics in a satisfactory manner, one must employ nonlinear econometric models or use nonparametric statistical methods. For most applications, it suffices to employ simple nonlinear models. For example, the quarterly growth rate of the U.S. gross domestic product can be adequately described by the Markov switching or threshold autoregressive models. These models typically classify the state of the U.S. economy into two categories corresponding roughly to expansion and contraction.

In this entry, we study nonlinearity in financial data, discuss various nonlinear models available in the literature, and demonstrate application of nonlinear models in finance with real examples. The models discussed include bilinear models, *threshold autoregressive models*, smooth threshold autoregressive models, *Markov switching models*, and nonlinear additive autoregressive models. We also consider *nonparametric methods* and *neural networks*, and apply nonparametric methods to estimate interest models. To detect nonlinearity in financial data, we introduce various *nonlinearity tests* available in the literature and apply the tests to some financial series. Finally, we analyze the monthly U.S. unemployment rate and compare out-of-sample prediction of nonlinear models with linear ones via several criteria.

STUDY OF NONLINEARITY IN ECONOMETRICS AND STATISTICS

Assume, for simplicity, a univariate time series x_t is observed at equally spaced time points. We denote the observations by $\{x_t | t = 1, \dots, T\}$, where T is the sample size. A purely stochastic time series x_t is said to be linear if it can be written as

$$x_t = \mu + \sum_{i=0}^{\infty} \psi_i a_{t-i} \quad (1)$$

where μ is a constant, ψ_i are real numbers with $\psi_0 = 1$, and $\{a_i\}$ is a sequence of independent and identically distributed (IID) random variables with a well-defined distribution function. We assume that the distribution of a_t is

continuous and $E(a_t) = 0$. In many cases, we further assume that $\text{Var}(a_t) = \sigma_a^2$ or, even stronger, that a_t is Gaussian. If $\sigma_a^2 \sum_{i=1}^{\infty} \psi_i^2 < \infty$, then X_t is weakly stationary (i.e., the first two moments of x_t are time-invariant). The well-known autoregressive moving-average (ARMA) process of Box et al. (2008) is linear because it has an moving-average (MA) representation in equation (1). Any stochastic process that does not satisfy the condition of equation (1) is said to be nonlinear. The prior definition of nonlinearity is for purely stochastic time series. One may extend the definition by allowing the mean of x_t to be a linear function of some exogenous variables, including the time index and some periodic functions. But such a mean function can be handled easily by using a regression model with time series errors discussed in Tsay (2010, Chapter 2), and we shall not consider the extension here. Mathematically, a purely stochastic time series model for x_t is a function of an IID sequence consisting of the current and past shocks—that is,

$$x_t = f(a_t, a_{t-i}, \dots) \quad (2)$$

The linear model in equation (1) says that $f(\cdot)$ is a linear function of its arguments. Any nonlinearity in $f(\cdot)$ results in a nonlinear model. The general nonlinear model in equation (2) is too vague to be useful in practice. Further assumptions are needed to make the model applicable.

To put nonlinear models available in the literature in a proper perspective, we write the model of x_t in terms of its conditional moments. Let F_{t-1} be the σ -field generated by available information at time $t - 1$ (inclusive). Typically, F_{t-1} denotes the collection of linear combinations of elements in $\{x_{t-1}, x_{t-2}, \dots\}$ and $\{a_{t-1}, a_{t-2}, \dots\}$. The conditional mean and variance of x_t given F_{t-1} are

$$\begin{aligned} \mu_t &= E(x_t | F_{t-1}) \equiv g(F_{t-1}) \\ \sigma_t^2 &= \text{Var}(x_t | F_{t-1}) \equiv h(F_{t-1}) \end{aligned} \quad (3)$$

where $g(\cdot)$ and $h(\cdot)$ are well-defined functions with $h(\cdot) > 0$. Thus, we restrict the model to

$$x_t = g(F_{t-1}) + \sqrt{h(F_{t-1})}e_t$$

where $e_t = a_t/\sigma_t$ is a standardized shock (or innovation). For the linear series x_t in equation (1), $g(\cdot)$ is a linear function of elements of F_{t-1} and $h(\cdot) = \sigma_a^2$. The development of nonlinear models involves making extensions of the two equations in equation (3). If $g(\cdot)$ is nonlinear, x_t is said to be nonlinear in mean. If $h(\cdot)$ is time-variant, then x_t is nonlinear in variance. The conditional heteroscedastic models, for example, the GARCH model of Bollerslev (1986), are nonlinear in variance because their conditional variances σ_t^2 evolve over time. Based on the well-known Wold decomposition, a weakly stationary and purely stochastic time series can be expressed as a linear function of uncorrelated shocks. For stationary volatility series, these shocks are uncorrelated, but dependent. The models discussed in this entry represent another extension to nonlinearity derived from modifying the conditional mean equation in equation (3).

Many nonlinear time series models have been proposed in the statistical literature, such as the bilinear models of Granger and Andersen (1978), the threshold autoregressive (TAR) model of Tong (1978), the state-dependent model of Priestley (1980), and the Markov switching model of Hamilton (1989). The basic idea underlying these nonlinear models is to let the conditional mean μ_t evolve over time according to some simple parametric nonlinear function. Recently, a number of nonlinear models have been proposed by making use of advances in computing facilities and computational methods. Examples of such extensions include the nonlinear state-space modeling of Carlin, Polson, and Stoffer (1992), the functional-coefficient autoregressive model of Chen and Tsay (1993a), the nonlinear additive autoregressive model of Chen and Tsay (1993b), the multivariate adaptive regression spline of Lewis and Stevens (1991), and the generalized autoregressive score (GAS) model of Creal et al. (2010). The basic idea of these extensions is either using simulation methods to describe the evolution of the conditional

distribution of x_t or using data-driven methods to explore the nonlinear characteristics of a series. Finally, nonparametric and semiparametric methods such as kernel regression and artificial neural networks have also been applied to explore the nonlinearity in a time series. We discuss some nonlinear models in this entry that are applicable to financial time series. The discussion includes some nonparametric and semiparametric methods.

Apart from the development of various nonlinear models, there is substantial interest in studying test statistics that can discriminate linear series from nonlinear ones. Both parametric and nonparametric tests are available. Most parametric tests employ either the Lagrange multiplier or likelihood ratio statistics. Nonparametric tests depend on either higher order spectra of x_t or the concept of dimension correlation developed for chaotic time series. We review some nonlinearity tests, discuss modeling and forecasting of nonlinear models, and provide an application of nonlinear models.

NONLINEAR MODELS

Most nonlinear models developed in the statistical literature focus on the conditional mean equation in equation (3); see Priestley (1988) and Tong (1990) for summaries of nonlinear models. Our goal here is to introduce some nonlinear models that are useful in finance.

Bilinear Model

The linear model in equation (1) is simply the first-order Taylor series expansion of the $f(\cdot)$ function in equation (2). As such, a natural extension to nonlinearity is to employ the second-order terms in the expansion to improve the approximation. This is the basic idea of bilinear models, which can be defined as

$$x_t = c + \sum_{i=1}^p \phi_i x_{t-i} - \sum_{j=1}^q \theta_j a_{t-j} + \sum_{i=1}^m \sum_{j=1}^s \beta_{ij} x_{t-i} a_{t-j} + a_t \tag{4}$$

where $p, q, m,$ and s are nonnegative integers. This model was introduced by Granger and Andersen (1978) and has been widely investigated. Subba Rao and Gabr (1984) discuss some properties and applications of the model, and Liu and Brockwell (1988) study general bilinear models. Properties of bilinear models such as stationarity conditions are often derived by (a) putting the model in a state-space form and (b) using the state transition equation to express the state as a product of past innovations and random coefficient vectors. A special generalization of the bilinear model in equation (4) has conditional heteroscedasticity. For example, consider the model

$$x_t = \mu + \sum_{i=1}^s \beta_i a_{t-i} a_t + a_t \tag{5}$$

where $\{a_t\}$ is a white noise series. The first two conditional moments of x_t are

$$E(x_t | F_{t-1}) = \mu$$

$$\text{Var}(x_t | F_{t-1}) = \left(1 + \sum_{i=1}^s \beta_i a_{t-i} \right)^2 \sigma_a^2$$

which confirm that the model has time-varying volatility.

Example 1. Consider the monthly simple returns of the CRSP equal-weighted index from January 1926 to December 2008 for 996 observations. Denote the series by R_t . The sample partial autocorrelation function (PACF) of R_t shows significant serial correlations at lags 1 and 3 so that an AR(3) model is used for the mean equation. The squared series of the AR(3) residuals suggests that the conditional heteroscedasticity might depend on lags 1, 3 and 8 of the residuals. Therefore, we employ the special bilinear model

$$R_t = \mu + \phi_1 R_{t-1} + \phi_3 R_{t-3} + (1 + \beta_1 a_{t-1} + \beta_3 a_{t-3}) a_t$$

for the series, where $a_t = \beta_0 \epsilon_t$ with ϵ_t being an IID series with mean zero and variance 1. Note that lag 8 is omitted for simplicity. Assuming that the conditional distribution of a_t is normal, we use the conditional maximum

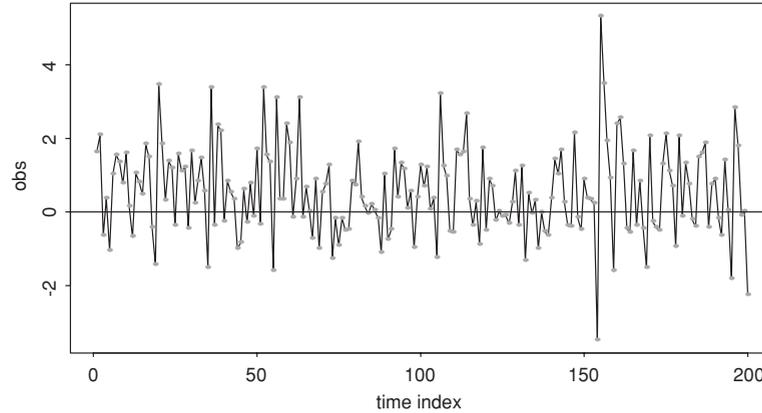


Figure 1 Time Plot of a Simulated 2-Regime TAR(1) Series

likelihood method and obtain the fitted model

$$R_t = 0.0114 + 0.167R_{t-1} - 0.095R_{t-3} + 0.071(1 + 0.377a_{t-1} - 0.646a_{t-3})\epsilon_t \quad (6)$$

where the standard errors of the parameters are, in the order of appearance, 0.0023, 0.032, 0.027, 0.002, 0.147, and 0.136, respectively. All estimates are significantly different from zero at the 5% level. Define

$$\hat{\epsilon}_t = \frac{R_t - 0.0114 - 0.167R_{t-1} + 0.095R_{t-3}}{0.071(1 + 0.377\hat{a}_{t-1} - 0.646\hat{a}_{t-3})}$$

where $\hat{\epsilon}_t = 0$ for $t \leq 3$ as the standardized residual series of the model. The sample autocorrelation function (ACF) of $\hat{\epsilon}_t$ shows no significant serial correlations, but the series is not independent because the squared series $\hat{\epsilon}_t^2$ has significant serial correlations. The validity of model (6) deserves further investigation. For comparison, we also consider an AR(3)-ARCH(3) model for the series and obtain

$$R_t = 0.013 + 0.223R_{t-1} + 0.006R_{t-2} - 0.013R_{t-3} + a_t \quad (7)$$

$$\sigma_t^2 = 0.002 + 0.185a_{t-1}^2 + 0.301a_{t-2}^2 + 0.197a_{t-3}^2$$

where all estimates but the coefficients of R_{t-2} and R_{t-3} are highly significant. The standardized residual series of the model shows no serial correlations, but the squared residuals show $Q(10) = 19.78$ with a p -value of 0.031. Models (6) and (7) appear to be similar, but the latter seems

to fit the data better. Further study shows that an AR(1)-GARCH(1,1) model fits the data well.

Threshold Autoregressive (TAR) Model

This model is motivated by several nonlinear characteristics commonly observed in practice such as asymmetry in declining and rising patterns of a process. It uses piecewise linear models to obtain a better approximation of the conditional mean equation. However, in contrast to the traditional piecewise linear model that allows for model changes to occur in the “time” space, the TAR model uses threshold space to improve linear approximation. Let us start with a simple 2-regime AR(1) model

$$x_t = \begin{cases} -1.5x_{t-1} + a_t & \text{if } x_{t-1} < 0 \\ 0.5x_{t-1} + a_t & \text{if } x_{t-1} \geq 0 \end{cases} \quad (8)$$

where the a_t are IID $N(0,1)$. Here the threshold variable is x_{t-1} and the threshold is 0.

Figure 1 shows the time plot of a simulated series of x_t with 200 observations. A horizontal line of zero is added to the plot, which illustrates several characteristics of TAR models. First, despite the coefficient -1.5 in the first regime, the process x_t is geometrically ergodic and stationary. In fact, the necessary and sufficient condition for model (8) to be geometrically ergodic is $\phi_1^{(1)} < 1$, $\phi_1^{(2)} < 1$ and $\phi_1^{(1)} \phi_1^{(2)} < 1$, where

$\phi_1^{(i)}$ is the AR coefficient of regime i ; see Petrucci and Woolford (1984) and Chen and Tsay (1991).

Ergodicity is an important concept in time series analysis. For example, the statistical theory showing that the sample mean $\bar{x} = (\sum_{t=1}^T x_t)/T$ of x_t converges to the mean of x_t is referred to as the ergodic theorem, which can be regarded as the counterpart of the central limit theory for the IID case. Second, the series exhibits an asymmetric increasing and decreasing pattern. If x_{t-1} is negative, then x_t tends to switch to a positive value due to the negative and explosive coefficient -1.5 . Yet when x_{t-1} is positive, it tends to take multiple time periods for x_t to reduce to a negative value. Consequently, the time plot of x_t shows that regime 2 has more observations than regime 1, and the series contains large upward jumps when it becomes negative. The series is therefore not time-reversible. Third, the model contains no constant terms, but $E(x_t)$ is not zero. The sample mean of the particular realization is 0.61 with a standard deviation of 0.07. In general, $E(x_t)$ is a weighted average of the conditional means of the two regimes, which are nonzero. The weight for each regime is simply the probability that x_t is in that regime under its stationary distribution. It is also clear from the discussion that, for a TAR model to have zero mean, nonzero constant terms in some of the regimes are needed. This is very different from a stationary linear model for which a nonzero constant implies that the mean of x_t is not zero.

A time series x_t is said to follow a k -regime self-exciting TAR (SETAR) model with threshold variable x_{t-d} if it satisfies

$$x_t = \phi_0^{(j)} + \phi_1^{(j)} x_{t-1} - \dots - \phi_p^{(j)} x_{t-p} + a_t^{(j)} \quad (9)$$

if $\gamma_{j-1} \leq x_{t-d} < \gamma_j$

where k and d are positive integers, $j = 1, \dots, k$, γ_i are real numbers such that $-\infty = \gamma_0 < \gamma_1 < \dots < \gamma_{k-1} < \gamma_k = \infty$, the superscript (j) is used to signify the regime, and $\{a_t^{(j)}\}$ are IID sequences with mean 0 and variance σ_j^2 and are mutually independent for different j . The parameter d is referred to as the delay pa-

rameter and γ_j are the thresholds. Here it is understood that the AR models are different for different regimes; otherwise, the number of regimes can be reduced. Equation (9) says that a SETAR model is a piecewise linear AR model in the threshold space. It is similar in spirit to the usual piecewise linear models in regression analysis, where model changes occur in the order in which observations are taken. The SETAR model is nonlinear provided that $k > 1$.

Properties of general SETAR models are hard to obtain, but some of them can be found in Tong (1990), Chan (1993), Chan and Tsay (1998), and the references therein. In recent years, there is increasing interest in TAR models and their applications; see, for instance, Hansen (1997), Tsay (1998), and Montgomery et al. (1998). Tsay (1989) proposed a testing and modeling procedure for univariate SETAR models. The model in equation (9) can be generalized by using a threshold variable z_t that is measurable with respect to F_{t-1} (i.e., a function of elements of F_{t-1}). The main requirements are that z_t is stationary with a continuous distribution function over a compact subset of the real line and that z_{t-d} is known at time t . Such a generalized model is referred to as an open-loop TAR model.

Example 2. To demonstrate the application of TAR models, consider the U.S. monthly civilian unemployment rate, seasonally adjusted and measured in percentage, from January 1948 to March 2009 for 735 observations. The data are obtained from the Bureau of Labor Statistics, Department of Labor, and are shown in Figure 2. The plot shows two main characteristics of the data. First, there appears to be a slow but upward trend in the overall unemployment rate. Second, the unemployment rate tends to increase rapidly and decrease slowly. Thus, the series is not time-reversible and may not be unit-root stationary, either.

Because the sample autocorrelation function decays slowly, we employ the first differenced series $y_t = (1-B)u_t$ in the analysis, where u_t is the monthly unemployment rate. Using

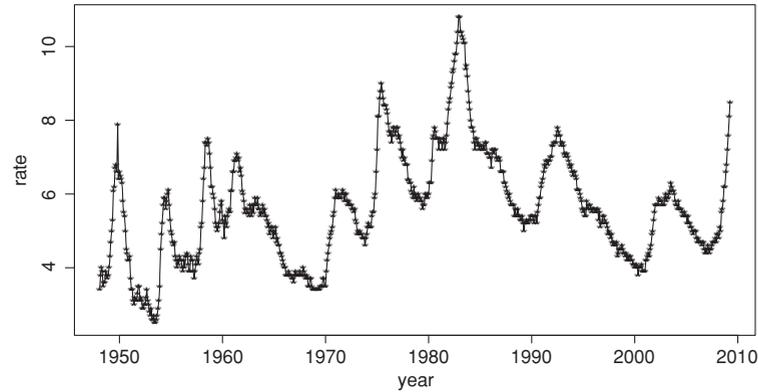


Figure 2 Time Plot of Monthly U.S. Civilian Unemployment Rate, Seasonally Adjusted, from January 1948 to March 2009

univariate ARIMA models, we obtain the model

$$\begin{aligned} (1 - 1.13B + 0.27B^2)(1 - 0.51B^{12})y_t \\ = (1 - 1.12B + 0.44B^2)(1 - 0.82B^{12})a_t \end{aligned} \quad (10)$$

where $\hat{\sigma}_a = 0.187$ and all estimates but the AR(2) coefficient are statistically significant at the 5% level. The t-ratio of the estimate of AR(2) coefficient is -1.66 . The residuals of model (10) give $Q(12) = 12.3$ and $Q(24) = 25.5$, respectively. The corresponding p -values are 0.056 and 0.11, respectively, based on χ^2 distributions with 6 and 18 degrees of freedom. Thus, the fitted model adequately describes the serial dependence of the data. Note that the seasonal AR and MA coefficients are highly significant with standard error 0.049 and 0.035, respectively, even though the data were seasonally adjusted. The adequacy of seasonal adjustment deserves further study. Using model (10), we obtain the 1-step ahead forecast of 8.8 for the April 2009 unemployment rate, which is close to the actual data of 8.9.

To model nonlinearity in the data, we employ TAR models and obtain the model

$$y_t = \begin{cases} 0.083y_{t-2} + 0.158y_{t-3} + 0.118y_{t-4} \\ \quad - 0.180y_{t-12} + a_{1t} & \text{if } y_{t-i} \leq 0.1 \\ 0.421y_{t-2} + 0.239y_{t-3} - 0.127y_{t-12} \\ \quad + a_{2t} & \text{if } y_{t-i} > 0.1 \end{cases} \quad (11)$$

where the standard errors of a_{it} are 0.180 and 0.217, respectively, the standard errors of the AR parameters in regime 1 are 0.046, 0.043, 0.042, and 0.037 whereas those of the AR parameters in regime 2 are 0.054, 0.057, and 0.075, respectively. The number of data points in regimes 1 and 2 are 460 and 262, respectively. The standardized residuals of model (11) only shows some minor serial correlation at lag 12. Based on the fitted TAR model, the dynamic dependence in the data appears to be stronger when the change in monthly unemployment rate is greater than 0.1%. This is understandable because a substantial increase in the unemployment rate is indicative of weakening in the U.S. economy, and policy makers might be more inclined to take action to help the economy, which in turn may affect the dynamics of the unemployment rate series. Consequently, model (11) is capable of describing the time-varying dynamics of the U.S. unemployment rate.

The MA representation of model (10) is

$$\begin{aligned} \psi(B) \approx 1 + 0.01B + 0.18B^2 + 0.20B^3 \\ + 0.18B^4 + 0.15B^5 + \dots \end{aligned}$$

It is then not surprising to see that no y_{t-1} term appears in model (11).

Threshold models can be used in finance to handle the leverage effect, that is, volatility responds differently to prior positive and negative returns. The models can also be used to

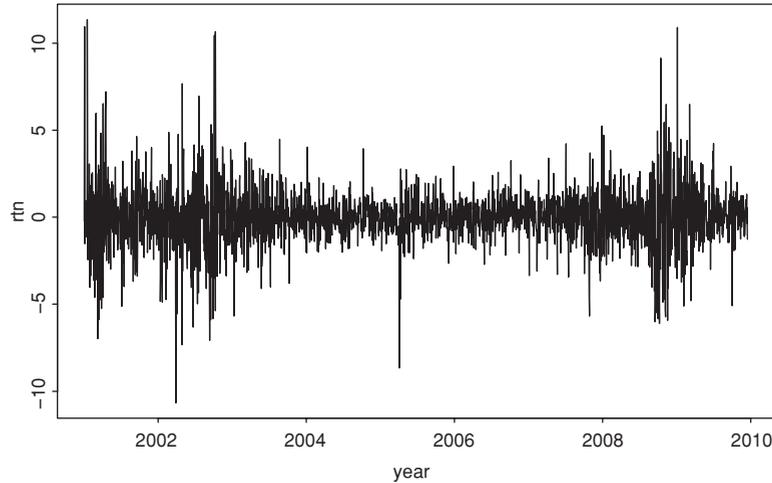


Figure 3 Time Plot of the Daily Log Returns, in Percentages, for IBM Stock from January 2, 2001 to December 31, 2009

study arbitrage trading in index futures and cash prices. See Tsay (2010, chap. 8) for discussions and demonstration. Here we focus on volatility modeling and introduce an alternative approach to parameterization of threshold GARCH (TGARCH) models. In some applications, this new general TGARCH model fares better than the model of Glosten et al. (1993).

Example 3. Consider the daily log returns, in percentages and including dividends, of IBM stock from January 2, 2001 to December 31, 2009 for 2,263 observations. Figure 3 shows the time plot of the series. The volatility seems to be larger at the beginning and end of the data span. If GARCH models are entertained, we obtain the following GARCH(1,1) model for the series:

$$\begin{aligned} r_t &= 0.058 + a_t, & a_t &= \sigma_t \epsilon_t \\ \sigma_t^2 &= 0.041 + 0.093a_{t-1}^2 + 0.894\sigma_{t-1}^2 \end{aligned} \quad (12)$$

where r_t is the log return, $\{\epsilon_t\}$ is a Gaussian white noise sequence with mean zero and variance 1.0, the standard error of the constant term in the mean equation is 0.026, and those of the volatility equation are 0.012, 0.020, and 0.021, respectively. All estimates are statistically significant at the 5% level. The Ljung-Box statistics of the standardized

residuals, $\hat{\epsilon}_t = \hat{a}_t / \hat{\sigma}_t$, give $Q(10) = 10.08(0.43)$ and $Q(20) = 23.24(0.28)$, where the number in parentheses denotes p -value obtained using the asymptotic X_m^2 distribution. For the squared standardized residuals, we obtain $Q(10) = 7.38(0.69)$ and $Q(20) = 15.43(0.75)$. The model is adequate in modeling the serial dependence and conditional heteroscedasticity of the data. But the unconditional mean for r_t of model (12) is 0.058, which is substantially larger than the sample mean 0.024, indicating that the model might be misspecified.

Next, we employ the TGARCH model of Glosten et al. (1993) and obtain

$$\begin{aligned} r_t &= 0.015 + a_t, & a_t &= \sigma_t \epsilon_t \\ \sigma_t^2 &= 0.032 + 0.033a_{t-1}^2 + 0.091N_{t-1}a_{t-1}^2 \\ &\quad + 0.911\sigma_{t-1}^2 \end{aligned} \quad (13)$$

where N_{t-1} is the indicator for negative a_{t-1} such that $N_{t-1} = 1$ if $a_{t-1} < 0$ and $= 0$ otherwise, the standard error of the parameter in the mean equation is 0.026, and those of the volatility equation are 0.005, 0.005, 0.006, and 0.008, respectively. All estimates except the constant term of the mean equation are highly significant. Let \tilde{a}_t be the standardized residuals of model (13). We have $Q(10) = 9.81(0.46)$ and $Q(20) = 22.17(0.33)$ for the $\{\tilde{a}_t\}$ series and

$Q(10) = 22.12(0.01)$ and $Q(20) = 31.15(0.05)$ for $\{\hat{a}_t^2\}$. The model fails to describe the conditional heteroscedasticity of the data at the 5% level.

The idea of TAR models can be used to refine the prior TGARCH model by allowing for increased flexibility in modeling the asymmetric response in volatility. More specifically, we consider a TAR-GARCH(1,1) model for the series and use the constrained optimization method L-BFGS-B to perform estimation. The resulting model is

$$\begin{aligned} r_t &= 0.023 + a_t, \quad a_t = \sigma_t \epsilon_t \\ \sigma_t^2 &= 0.086 + 0.044a_{t-1}^2 + 0.815\sigma_{t-1}^2 \\ &\quad + (-0.114 + 0.052a_{t-1}^2 + 0.214\sigma_{t-1}^2)N_{t-1} \end{aligned} \quad (14)$$

where all estimates are significant at the 5% level and N_{t-1} is defined in equation (13). The estimate -0.114 is only marginally significant because its standard error is 0.055. The coefficient of σ_{t-1}^2 is greater than 1 when $a_{t-1} < 0$, but it is not significantly different from 1 in view of its standard error.

Let \hat{a}_t be the standardized residuals of model (14). We obtain $Q(10) = 9.10(0.52)$ and $Q(20) = 21.82(0.35)$ for $\{\hat{a}_t\}$ and $Q(10) = 19.80(0.03)$ and $Q(20) = 27.41(0.12)$ for $\{\hat{a}_t^2\}$. Thus, model (14) is adequate in modeling the serial correlation and conditional heteroscedasticity of the daily log returns of IBM stock considered. The unconditional mean return of model (14) is 0.023, which is much closer to the sample mean 0.024 than those implied by models (12) and (13). Comparing the fitted TAR-GARCH and TGARCH models, we see that the asymmetric behavior in daily IBM stock volatility is much stronger than what is allowed in a TGARCH model. Specifically, the coefficient of σ_{t-1}^2 also depends on the sign of a_{t-1} .

Smooth Transition AR (STAR) Model

A criticism of the SETAR model is that its conditional mean equation is not continuous. The

thresholds $\{\gamma_j\}$ are the discontinuity points of the conditional mean function μ_t . In response to this criticism, smooth TAR models have been proposed; see Chan and Tong (1986) and Teräsvirta (1994) and the references therein. A time series x_t follows a 2-regime STAR(p) model if it satisfies

$$\begin{aligned} x_t &= c_0 + \sum_{i=1}^p \phi_{0,i} x_{t-i} + F\left(\frac{x_{t-d} - \Delta}{s}\right) \\ &\quad \times \left(c_1 + \sum_{i=1}^p \phi_{1,i} x_{t-i}\right) + a_t \end{aligned} \quad (15)$$

where d is the delay parameter, Δ and s are parameters representing the location and scale of model transition, and $F(\cdot)$ is a smooth transition function. In practice, $F(\cdot)$ often assumes one of three forms—namely, logistic, exponential, or a cumulative distribution function. From equation (15) and with $0 \leq F(\cdot) \leq 1$, the conditional mean of a STAR model is a weighted linear combination between the following two equations:

$$\begin{aligned} \mu_{1t} &= c_0 + \sum_{i=1}^p \phi_{0,i} x_{t-i} \\ \mu_{2t} &= (c_0 + c_1) + \sum_{i=1}^p (\phi_{0,i} + \phi_{1,i}) x_{t-i} \end{aligned}$$

The weights are determined in a continuous manner by $F((x_{t-d} - \Delta)/s)$. The prior two equations also determine properties of a STAR model. For instance, a prerequisite for the stationarity of a STAR model is that all zeros of both AR polynomials are outside the unit circle. An advantage of the STAR model over the TAR model is that the conditional mean function is differentiable. However, experience shows that the transition parameters Δ and s of a STAR model are hard to estimate. In particular, most empirical studies show that standard errors of the estimates of Δ and s are often quite large, resulting in t -ratios about 1.0; see Teräsvirta (1994). This uncertainty leads to various complications in interpreting an estimated STAR model.

Example 4. To illustrate the application of STAR models in financial time series analysis,

we consider the monthly simple stock returns for Minnesota Mining and Manufacturing (3M) Company from February 1946 to December 2008. If ARCH models are entertained, we obtain the following ARCH(2) model

$$\begin{aligned} R_t &= 0.013 + a_t, \quad a_t = \sigma_t \epsilon_t \\ \sigma_t^2 &= 0.003 + 0.088a_{t-1}^2 + 0.109a_{t-2}^2 \end{aligned} \quad (16)$$

where standard errors of the estimates are 0.002, 0.0003, 0.047, and 0.050, respectively. As discussed before, such an ARCH model fails to show the asymmetric responses of stock volatility to positive and negative prior shocks. The STAR model provides a simple alternative that may overcome this difficulty. Applying STAR models to the monthly returns of 3M stock, we obtain the model

$$\begin{aligned} R_t &= 0.015 + at, \quad a_t = \sigma_t \epsilon_t \\ \sigma_t^2 &= (0.003 + 0.205a_{t-1}^2 + 0.092a_{t-2}^2) \\ &\quad + \frac{0.001 - 0.239a_{t-1}^2}{1 + \exp(-1000a_{t-1})} \end{aligned} \quad (17)$$

where the standard error of the constant term in the mean equation is 0.002 and the standard errors of the estimates in the volatility equation are 0.0002, 0.074, 0.043, 0.0004, and 0.080, respectively. The scale parameter 1000 of the logistic transition function is fixed a priori to simplify the estimation. This STAR model provides some support for asymmetric responses to positive and negative prior shocks. For a large negative a_{t-1} , the volatility model approaches the ARCH(2) model

$$\sigma_t^2 = 0.003 + 0.205a_{t-1}^2 + 0.092a_{t-2}^2$$

Yet for a large positive a_{t-1} , the volatility process behaves like the ARCH(2) model

$$\sigma_t^2 = 0.004 - 0.034a_{t-1}^2 + 0.092a_{t-2}^2$$

The negative coefficient of a_{t-1}^2 in the prior model is counterintuitive, but the magnitude is small. As a matter of fact, for a large positive shock a_{t-1} , the ARCH effects appear to be weak even though the parameter estimates remain statistically significant.

Markov Switching Model

The idea of using probability switching in nonlinear time series analysis is discussed in Tong (1983). Using a similar idea, but emphasizing aperiodic transition between various states of an economy, Hamilton (1989) considers the Markov switching autoregressive (MSA) model. Here the transition is driven by a hidden two-state Markov chain. A time series x_t follows an MSA model if it satisfies

$$\begin{cases} c_1 + \sum_{i=1}^p \phi_{1,i} x_{t-i} + a_{1t} & \text{if } s_t = 1 \\ c_2 + \sum_{i=1}^p \phi_{2,i} x_{t-i} + a_{2t} & \text{if } s_t = 2 \end{cases} \quad (18)$$

where s_t assumes values in $\{1,2\}$ and is a first-order Markov chain with transition probabilities

$$P(s_t = 2 | s_{t-1} = 1) = w_1, \quad P(s_t = 1 | s_{t-1} = 2) = w_2$$

The innovational series $\{a_{1t}\}$ and $\{a_{2t}\}$ are sequences of IID random variables with mean zero and finite variance and are independent of one another. A small w_i means that the model tends to stay longer in state i . In fact, $1/w_i$ is the expected duration of the process to stay in state i . From the definition, an MSA model uses a hidden Markov chain to govern the transition from one conditional mean function to another. This is different from that of a SETAR model for which the transition is determined by a particular lagged variable. Consequently, a SETAR model uses a deterministic scheme to govern the model transition whereas an MSA model uses a stochastic scheme.

In practice, the stochastic nature of the states implies that one is never certain about which state x_t belongs to in an MSA model. When the sample size is large, one can use some filtering techniques to draw inference on the state of x_t . Yet as long as x_{t-d} is observed, the regime of x_t is known in a SETAR model. This difference has important practical implications in forecasting. For instance, forecasts of an MSA model are always a linear combination of forecasts produced by submodels of individual states. But those of a SETAR model only come from a single regime provided that x_{t-d} is observed.

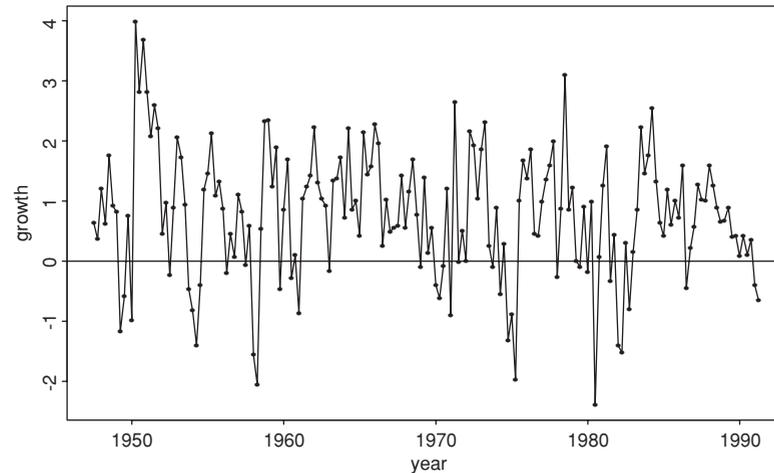


Figure 4 Time Plot of the Growth Rate of the U.S. Quarterly Real GNP from 1947.II to 1991.I
Note: The data are seasonally adjusted and in percentages.

Forecasts of a SETAR model also become a linear combination of those produced by models of individual regimes when the forecast horizon exceeds the delay d . It is much harder to estimate an MSA model than other models because the states are not directly observable. Hamilton (1990) uses the EM algorithm, which is a statistical method iterating between taking expectation and maximization. McCulloch and Tsay (1994) consider a Markov chain Monte Carlo (MCMC) method to estimate general MSA models. For applications of MCMC methods in finance, see Tsay (2010, Chapter 12).

McCulloch and Tsay (1993) generalize the MSA model in equation (18) by letting the transition probabilities w_1 and w_2 be logistic, or probit, functions of some explanatory variables available at time $t - 1$. Chen, McCulloch, and Tsay (1997) use the idea of Markov switching as a tool to perform model comparison and selection between nonnested nonlinear time series models (e.g., comparing bilinear and SETAR models). Each competing model is represented by a state. This approach to select a model is a generalization of the odds ratio commonly used in Bayesian analysis. Finally, the MSA model can easily be generalized to the case of more than two states. The computational intensity

involved increases rapidly, however. For more discussions of Markov switching models in econometrics, see Hamilton (1994, Chapter 22).

Example 5. Consider the growth rate, in percentages, of the U.S. quarterly real gross national product (GNP) from the second quarter of 1947 to the first quarter of 1991. The data are seasonally adjusted and shown in Figure 4, where a horizontal line of zero growth is also given. It is reassuring to see that a majority of the growth rates are positive. This series has been widely used in nonlinear analysis of economic time series. Tiao and Tsay (1994) and Potter (1995) use TAR models, whereas Hamilton (1989) and McCulloch and Tsay (1994) employ Markov switching models.

Employing the MSA model in equation (18) with $p = 4$ and using a Markov chain Monte Carlo method, McCulloch and Tsay (1994) obtain the estimates shown in Table 1. The results have several interesting findings. First, the mean growth rate of the marginal model for state 1 is $0.909 / (1 - 0.265 - 0.029 + 0.126 + 0.11) = 0.965$ and that of state 2 is $-0.42 / (1 - 0.216 - 0.628 + 0.073 + 0.097) = -1.288$. Thus, state 1 corresponds to quarters with positive growth, or expansion periods, whereas state 2

Table 1 Estimation Results of a Markov Switching Model with $p = 4$ for the Growth Rate of U.S. Quarterly Real GNP, Seasonally Adjusted

Parameter	State 1						
	c_i	ϕ_1	ϕ_2	ϕ_3	ϕ_4	σ_i	w_i
Estimate	0.909	0.265	0.029	-0.126	-0.110	0.816	0.118
Std. Error	0.202	0.113	0.126	0.103	0.109	0.125	0.053
Parameter	State 2						
	c_i	ϕ_1	ϕ_2	ϕ_3	ϕ_4	σ_i	w_i
Estimate	-0.420	0.216	0.628	-0.073	-0.097	1.017	0.286
Std. Error	0.324	0.347	0.377	0.364	0.404	0.293	0.064

Note: The estimates and their standard errors are posterior means and standard errors of a Gibbs sampling with 5000 iterations.

consists of quarters with negative growth, or a contraction period. Second, the relatively large posterior standard deviations of the parameters in state 2 reflect that there are few observations in that state. This is expected as Figure 4 shows few quarters with negative growth. Third, the transition probabilities appear to be different for different states. The estimates indicate that it is more likely for the U.S. GNP to get out of a contraction period than to jump into one -0.286 versus 0.118 . Fourth, treating $1/w_i$ as the expected duration for the process to stay in state i , we see that the expected durations for a contraction period and an expansion period are approximately 3.69 and 11.31 quarters. Thus, on average, a contraction in the U.S. economy lasts about a year, whereas an expansion can last for 3 years. Finally, the estimated AR coefficients of x_{t-2} differ substantially between the two states, indicating that the dynamics of the U.S. economy are different between expansion and contraction periods.

Nonparametric Methods

In some financial applications, we may not have sufficient knowledge to prespecify the nonlinear structure between two variables Y and X . In other applications, we may wish to take advantage of the advances in computing facilities and computational methods to explore the functional relationship between Y and X . These considerations lead to the use of nonparametric

methods and techniques. Nonparametric methods, however, are not without cost. They are highly data dependent and can easily result in overfitting. Our goal here is to introduce some nonparametric methods for financial applications and some nonlinear models that make use of nonparametric methods and techniques. The nonparametric methods discussed include kernel regression, local least squares estimation, and neural network.

The essence of nonparametric methods is smoothing. Consider two financial variables Y and X , which are related by

$$Y_t = m(X_t) + a_t \quad (19)$$

where $m(\cdot)$ is an arbitrary, smooth, but unknown function and $\{a_t\}$ is a white noise sequence. We wish to estimate the nonlinear function $m(\cdot)$ from the data. For simplicity, consider the problem of estimating $m(\cdot)$ at a particular date for which $X = x$. That is, we are interested in estimating $m(x)$. Suppose that at $X = x$ we have repeated independent observations y_1, \dots, y_T . Then the data become

$$y_t = m(x) + a_t, \quad t = 1, \dots, T$$

Taking the average of the data, we have

$$\frac{\sum_{t=1}^T y_t}{T} = m(x) + \frac{\sum_{t=1}^T a_t}{T}$$

By the law of large numbers, the average of the shocks converges to zero as T increases. Therefore, the average $\hat{y} = (\sum_{t=1}^T y_t)/T$ is a consistent

estimate of $m(x)$. That the average \bar{y} provides a consistent estimate of $m(x)$ or, alternatively, that the average of shocks converges to zero shows the power of smoothing.

In financial time series, we do not have repeated observations available at $X = x$. What we observed are $\{(y_t, x_t)\}$ for $t = 1, \dots, T$. But if the function $m(\cdot)$ is sufficiently smooth, then the value of Y_t for which $X_t \approx x$ continues to provide accurate approximation of $m(x)$. The value of Y_t for which X_t is far away from x provides less accurate approximation for $m(x)$. As a compromise, one can use a weighted average of y_t instead of the simple average to estimate $m(x)$. The weight should be larger for those Y_t with X_t close to x and smaller for those Y_t with X_t far away from x . Mathematically, the estimate of $m(x)$ for a given x can be written as

$$\hat{m}(x) = \frac{1}{T} \sum_{t=1}^T w_t(x) y_t \quad (20)$$

where the weights $w_t(x)$ are larger for those y_t with x_t close to x and smaller for those y_t with x_t far away from x . In equation (20), we assume that the weights sum to T . One can treat $1/T$ as part of the weights and make the weights sum to one.

From equation (20), the estimate $\hat{m}(x)$ is simply a local weighted average with weights determined by two factors. The first factor is the distance measure (i.e., the distance between x_t and x). The second factor is the assignment of weight for a given distance. Different ways to determine the distance between x_t and x and to assign the weight using the distance give rise to different nonparametric methods. In what follows, we discuss the commonly used kernel regression and local linear regression methods.

Kernel Regression

Kernel regression is perhaps the most commonly used nonparametric method in smoothing. The weights here are determined by a kernel, which is typically a probability density

function, is denoted by $K(x)$, and satisfies

$$K(x) \geq 0, \quad \int K(z) dz = 1$$

However, to increase the flexibility in distance measure, one often rescales the kernel using a variable $h > 0$, which is referred to as the bandwidth. The rescaled kernel becomes

$$K_h(x) = \frac{1}{h} K(x/h), \quad \int K_h(z) dz = 1 \quad (21)$$

The weight function can now be defined as

$$w_t(x) = \frac{K_h(x - x_t)}{\sum_{t=1}^T K_h(x - x_t)} \quad (22)$$

where the denominator is a normalization constant that makes the smoother adaptive to the local intensity of the X variable and ensures the weights sum to one. Plugging equation (22) into the smoothing formula (20), we have the well-known Nadaraya-Watson kernel estimator

$$\hat{m}(x) = \sum_{t=1}^T w_t(x) y_t = \frac{\sum_{t=1}^T K_h(x - x_t) y_t}{\sum_{t=1}^T K_h(x - x_t)} \quad (23)$$

see Nadaraya (1964) and Watson (1964). In practice, many choices are available for the kernel $K(x)$. However, theoretical and practical considerations lead to a few choices, including the Gaussian kernel

$$K_h(x) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{x^2}{2h^2}\right)$$

and the Epanechnikov kernel (Epanechnikov, 1969)

$$K_h(x) = \frac{0.75}{h} \left(1 - \frac{x^2}{h^2}\right) I\left(\left|\frac{x}{h}\right| \leq 1\right)$$

where $I(A)$ is an indicator such that $I(A) = 1$ if A holds and $I(A) = 0$ otherwise. Figure 5 shows the Gaussian and Epanechnikov kernels for $h = 1$.

To gain insight into the bandwidth h , we evaluate the Nadaraya-Watson estimator with the Epanechnikov kernel at the observed values $\{x_t\}$ and consider two extremes. First, if $h \rightarrow 0$, then

$$\hat{m}(x_t) \rightarrow \frac{K_h(0) y_t}{K_h(0)} = y_t$$

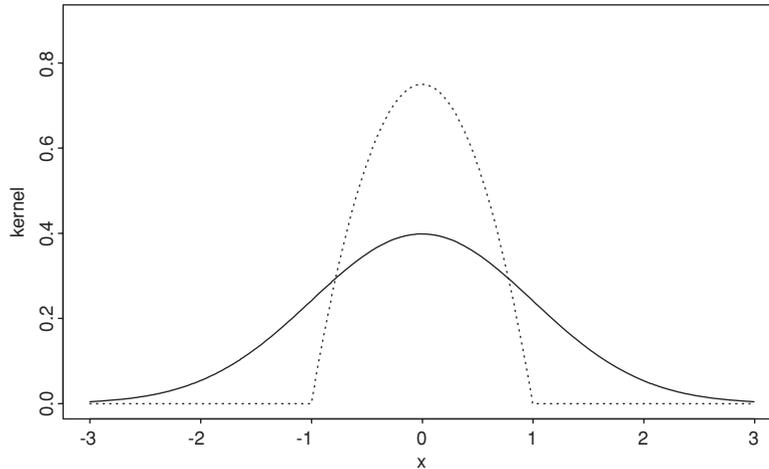


Figure 5 Standard Normal Kernel (Solid Line) and Epanechnikov Kernel (Dashed Line) with Bandwidth $h = 1$

indicating that small bandwidths reproduce the data. Second, if $h \rightarrow \infty$, then

$$\hat{m}(x_t) \rightarrow \frac{\sum_{t=1}^T K_h(0)y_t}{\sum_{t=1}^T K_h(0)} = \frac{1}{T} \sum_{t=1}^T y_t = \bar{y}$$

suggesting that large bandwidths lead to an oversmoothed curve—the sample mean. In general, the bandwidth function h acts as follows. If h is very small, then the weights focus on a few observations that are in the neighborhood around each x_t . If h is very large, then the weights will spread over a larger neighborhood of x_t . Consequently, the choice of h plays an important role in kernel regression. This is the well-known problem of bandwidth selection in kernel regression.

Bandwidth Selection

There are several approaches for bandwidth selection; see Härdle (1990) and Fan and Yao (2003). The first approach is the plug-in method, which is based on the asymptotic expansion of the mean integrated squared error (MISE) for *kernel smoothers*

$$\text{MISE} = E \int_{-\infty}^{\infty} [\hat{m}(x) - m(x)]^2 dx$$

where $m(\cdot)$ is the true function. The quantity $E[\hat{m}(x) - m(x)]^2$ of the MISE is a pointwise measure of the mean squared error (MSE) of $\hat{m}(x)$ evaluated at x .

Under some regularity conditions, one can derive the optimal bandwidth that minimizes the MISE. The optimal bandwidth typically depends on several unknown quantities that must be estimated from the data with some preliminary smoothing. Several iterations are often needed to obtain a reasonable estimate of the optimal bandwidth. In practice, the choice of preliminary smoothing can become a problem. Fan and Yao (2003) give a normal reference bandwidth selector as

$$\hat{h}_{\text{opt}} \begin{cases} 1.06sT^{-1/5} & \text{for the Gaussian kernel} \\ 2.34sT^{-1/5} & \text{for the Epanechnikov kernel} \end{cases}$$

where s is the sample standard error of the independent variable, which is assumed to be stationary.

The second approach to bandwidth selection is the leave-one-out cross-validation. First, one observation (x_j, y_j) is left out. The remaining $T - 1$ data points are used to obtain the following smoother at x_j :

$$\hat{m}_{h,j}(x_j) = \frac{1}{T-1} \sum_{t \neq j} w_t(x_j)y_t$$

which is an estimate of y_j , where the weights $w_t(x_j)$ sum to $T-1$. Second, perform step-1 for $j = 1, \dots, T$ and define the function

$$CV(h) = \frac{1}{T} \sum_{j=1}^T [y_j - \hat{m}_{h,j}(x_j)]^2 W(x_j)$$

where $w(\cdot)$ is a nonnegative weight function satisfying $\sum_{j=1}^n W(x_j) = T$, that can be used to down-weight the boundary points if necessary. Decreasing the weights assigned to data points close to the boundary is needed because those points often have fewer neighboring observations. The function $CV(h)$ is called the cross-validation function because it validates the ability of the smoother to predict $\{y_t\}_{t=1}^T$. One chooses the bandwidth h that minimizes the $CV(\cdot)$ function.

Local Linear Regression Method

Assume that the second derivative of $m(\cdot)$ in model (19) exists and is continuous at x , where x is a given point in the support of $m(\cdot)$. Denote the data available by $\{(y_t, x_t)\}_{t=1}^T$. The *local linear regression* method to nonparametric regression is to find a and b that minimize

$$L(a, b) = \sum_{t=1}^T [y_t - a - b(x - x_t)]^2 K_h(x - x_t) \quad (24)$$

where $K_h(\cdot)$ is a kernel function defined in equation (21) and h is a bandwidth. Denote the resulting value of a by \hat{a} . The estimate of $m(x)$ is then defined as \hat{a} . In practice, x assumes an observed value of the independent variable. The estimate \hat{b} can be used as an estimate of the first derivative of $m(\cdot)$ evaluated at x .

Under the least squares theory, equation (24) is a weighted least squares problem and one can derive a closed-form solution for a . Specifically, taking the partial derivatives of $L(a, b)$ with respect to both a and b and equating the derivatives to zero, we have a system of two equations with two unknowns:

$$\begin{aligned} \sum_{t=1}^T K_h(x - x_t) y_t &= a \sum_{t=1}^T K_h(x - x_t) \\ &+ b \sum_{t=1}^T (x - x_t) K_h(x - x_t) \end{aligned}$$

$$\begin{aligned} \sum_{t=1}^T y_t (x - x_t) K_h(x - x_t) &= a \sum_{t=1}^T (x - x_t) K_h(x - x_t) \\ &+ b \sum_{t=1}^T (x - x_t)^2 K_h(x - x_t) \end{aligned}$$

Define

$$s_{T,\ell} = \sum_{t=1}^T K_h(x - x_t) (x - x_t)^\ell, \quad \ell = 0, 1, 2.$$

The prior system of equations becomes

$$\begin{aligned} \begin{bmatrix} s_{T,0} & s_{T,1} \\ s_{T,1} & s_{T,2} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \\ = \begin{bmatrix} \sum_{t=1}^T K_h(x - x_t) y_t \\ \sum_{t=1}^T (x - x_t) K_h(x - x_t) y_t \end{bmatrix} \end{aligned}$$

Consequently, we have

$$\hat{a} = \frac{s_{T,2} \sum_{t=1}^T K_h(x - x_t) y_t - s_{T,1} \sum_{t=1}^T (x - x_t) K_h(x - x_t) y_t}{s_{T,0} s_{T,2} - s_{T,1}^2}$$

The numerator and denominator of the prior fraction can be further simplified as

$$\begin{aligned} s_{T,2} &= \sum_{t=1}^T K_h(x - x_t) y_t \\ &\quad - s_{T,1} \sum_{t=1}^T (x - x_t) K_h(x - x_t) y_t \\ &= \sum_{t=1}^T [K_h(x - x_t) (s_{T,2} - (x - x_t) s_{T,1})] y_t \\ s_{T,0} s_{T,2} - s_{T,1}^2 &= \sum_{t=1}^T K_h(x - x_t) s_{T,2} \\ &\quad - \sum_{t=1}^T (x - x_t) K_h(x - x_t) s_{T,1} \\ &= \sum_{t=1}^T K_h(x - x_t) [(s_{T,2} - (x - x_t) s_{T,1})] \end{aligned}$$

In summary, we have

$$\hat{a} = \frac{\sum_{t=1}^T w_t y_t}{\sum_{t=1}^T w_t} \quad (25)$$

where w_t is defined as

$$w_t = K_h(x - x_t) [s_{T,2} - (x - x_t) s_{T,1}]$$

In practice, to avoid possible zero in the denominator, we use $\hat{m}(x)$ next to estimate $m(x)$:

$$\hat{m}(x) = \frac{\sum_{t=1}^T w_t y_t}{\sum_{t=1}^T w_t + 1/T^2} \quad (26)$$

Notice that a nice feature of equation (26) is that the weight w_t satisfies

$$\sum_{t=1}^T (x - x_t)w_t = 0$$

Also, if one assumes that $m(\cdot)$ of equation (19) has the first derivative and finds the minimizer of

$$\sum_{t=1}^T (y_t - a)^2 K_h(x - x_t)$$

then the resulting estimator is the Nadaraya-Watson estimator mentioned earlier. In general, if one assumes that $m(x)$ has a bounded k th derivative, then one can replace the linear polynomial in equation (24) by a $(k - 1)$ -order polynomial. We refer to the estimator in equation (26) as the local linear regression smoother. Fan (1993) shows that, under some regularity conditions, the local linear regression estimator has some important sampling properties. The selection of bandwidth can be carried out via the same methods as before.

Financial Time Series Application

In time series analysis, the explanatory variables are often the lagged values of the series. Consider the simple case of a single explanatory variable. Here model (19) becomes

$$x_t = m(x_{t-1}) + a_t$$

and the kernel regression and local linear regression method discussed before are directly applicable. When multiple explanatory variables exist, some modifications are needed to implement the nonparametric methods. For the kernel regression, one can use a multivariate kernel such as a multivariate normal density function with a prespecified covariance matrix:

$$K_h(\mathbf{x}) = \frac{1}{(h\sqrt{2\pi})^p |\Sigma|^{1/2}} \exp\left(-\frac{1}{2h^2} \mathbf{x}' \Sigma^{-1} \mathbf{x}\right)$$

where p is the number of explanatory variables and Σ is a prespecified positive-definite matrix. Alternatively, one can use the product of

univariate kernel functions as a multivariate kernel—for example,

$$K_h(\mathbf{x}) = \prod_{i=1}^p \frac{0.75}{h_i} \left(1 - \frac{x_i^2}{h_i^2}\right) I\left(\left|\frac{x_i}{h_i}\right| < 1\right)$$

This latter approach is simple, but it overlooks the relationship between the explanatory variables.

Example 6. To illustrate the application of nonparametric methods in finance, consider the weekly 3-month Treasury bill secondary market rate from 1970 to 1997 for 1,461 observations. The data are obtained from the Federal Reserve Bank of St. Louis and are shown in Figure 6. This series has been used in the literature as an example of estimating stochastic diffusion equations using discretely observed data. Here we consider a simple model

$$y_t = \mu(x_{t-1})dt + \sigma(x_{t-1})dw_t$$

where x_t is the 3-month Treasury bill rate, $y_t = x_t - x_{t-1}$, w_t is a standard Brownian motion, and $\mu(\cdot)$ and $\sigma(\cdot)$ are smooth functions of x_{t-1} , and apply the local smoothing function *lowess* of R or S-Plus to obtain nonparametric estimates of $\mu(\cdot)$ and $\sigma(\cdot)$; see Cleveland (1979). For simplicity, we use $|y_t|$ as a proxy of the volatility of x_t .

For the simple model considered, $\mu(x_{t-1})$ is the conditional mean of y_t given x_{t-1} , that is, $\mu(x_{t-1}) = E(y_t|x_{t-1})$. Figure 7(a) shows the scatterplot of $y(t)$ versus x_{t-1} . The plot also contains the local smooth estimate of $\mu(x_{t-1})$ obtained by the method of *lowess* in the statistical package R. The estimate is essentially zero. However, to better understand the estimate, Figure 7(b) shows the estimate $\hat{\mu}(x_{t-1})$ on a finer scale. It is interesting to see that $\hat{\mu}(x_{t-1})$ is positive when x_{t-1} is small, but becomes negative when x_{t-1} is large. This is in agreement with the common sense that when the interest rate is high, it is expected to come down, and when the rate is low, it is expected to increase. Figure 7(c) shows the scatterplot of $|y(t)|$ versus x_{t-1} and the estimate of $\hat{\sigma}(x_{t-1})$ via *lowess*. The plot confirms that the

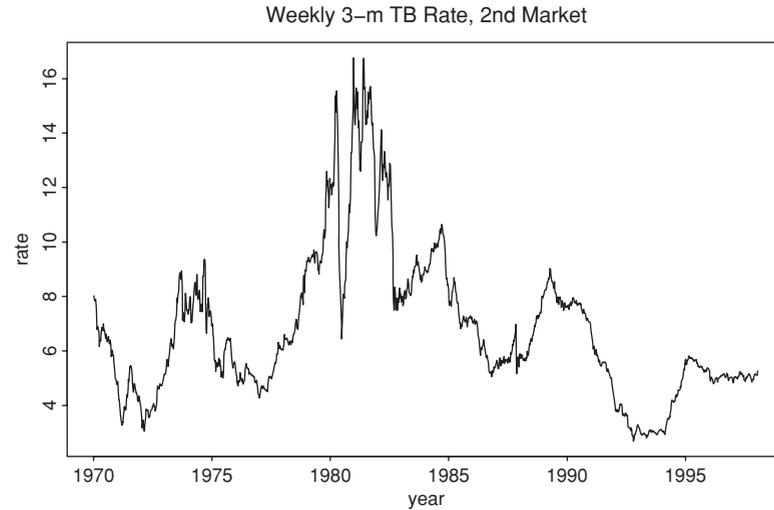


Figure 6 Time Plot of U.S. Weekly 3-Month Treasury Bill Rate in the Secondary Market from 1970 to 1997

higher the interest rate, the larger the volatility. Figure 7(d) shows the estimate $\hat{\sigma}(x_{t-1})$ on a finer scale. Clearly the volatility is an increasing function of x_{t-1} and the slope seems to accelerate when x_{t-1} is approaching 10%. This exam-

ple demonstrates that simple non-parametric methods can be helpful in understanding the dynamic structure of a financial time series.

The following nonlinear models are derived with the help of nonparametric methods.

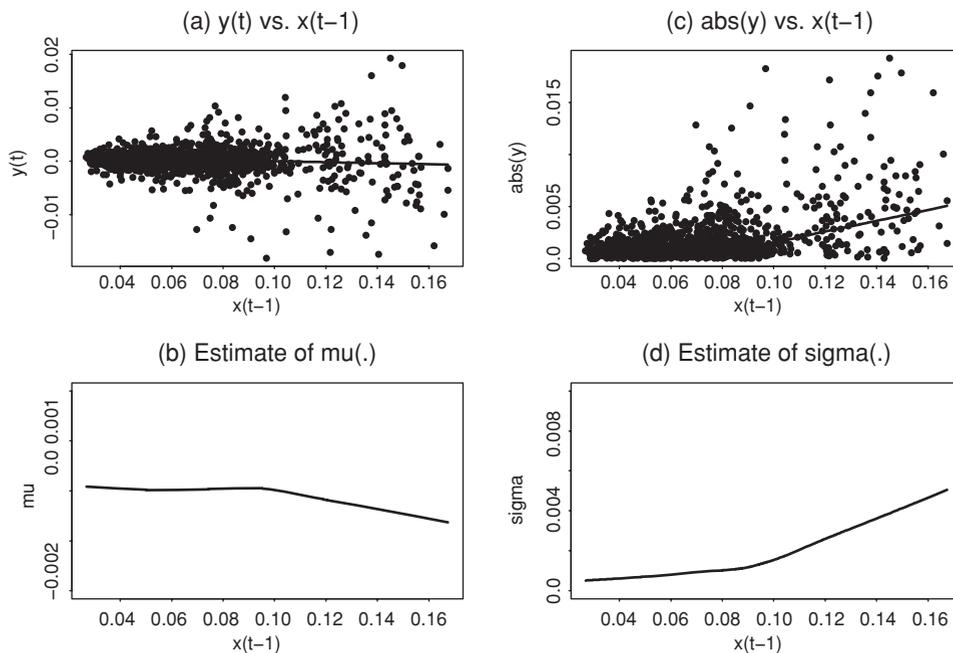


Figure 7 Estimation of Conditional Mean and Volatility of Weekly 3-Month Treasury Bill Rate via a Local Smoothing Method: (a) y_t versus x_{t-1} , where $y_t = x_t - x_{t-1}$ and x_t is the interest rate; (b) estimate of $\mu(x_{t-1})$; (c) $|y_t|$ versus x_{t-1} ; and (d) estimate of $\sigma(x_{t-1})$

Functional Coefficient AR Model

Recent advances in nonparametric techniques enable researchers to relax parametric constraints in proposing nonlinear models. In some cases, nonparametric methods are used in a preliminary study to help select a parametric nonlinear model. This is the approach taken by Chen and Tsay (1993a) in proposing the functional-coefficient autoregressive (FAR) model that can be written as

$$x_t = f_1(X_{t-1})x_{t-1} + \cdots + f_p(X_{t-1})x_{t-p} + a_t \quad (27)$$

where $X_{t-1} = (x_{t-1}, \dots, X_{t-k})'$ is a vector of lagged values of x_t . If necessary x_{t-1} may also include other explanatory variables available at time $t-1$. The functions $f_i(\cdot)$ of equation (27) are assumed to be continuous, even twice differentiable, almost surely with respect to their arguments. Most of the nonlinear models discussed before are special cases of the FAR model. In application, one can use nonparametric methods such as kernel regression or local linear regression to estimate the functional coefficients $f_i(\cdot)$, especially when the dimension of X_{t-1} is low (e.g., X_{t-1} is a scalar). Recently, Cai, Fan, and Yao (2000) applied the local linear regression method to estimate $f_i(\cdot)$ and showed that substantial improvements in 1-step ahead forecasts can be achieved by using FAR models.

Nonlinear Additive AR Model

A major difficulty in applying nonparametric methods to nonlinear time series analysis is the "curse of dimensionality." Consider a general nonlinear AR(p) process $x_t = f(x_{t-1}, \dots, x_{t-p}) + a_t$. A direct application of nonparametric methods to estimate $f(\cdot)$ would require p -dimensional smoothing, which is hard to do when p is large, especially if the number of data points is not large. A simple, yet effective way to overcome this difficulty is to entertain an additive model that only requires lower dimensional smoothing. A time series x_t follows a nonlinear

additive AR (NAAR) model if

$$x_t = f_0(t) + \sum_{i=1}^p f_i(x_{t-i}) + a_t \quad (28)$$

where the $f_i(\cdot)$ are continuous functions almost surely. Because each function $f_i(\cdot)$ has a single argument, it can be estimated nonparametrically using one-dimensional smoothing techniques and hence avoids the curse of dimensionality. In application, an iterative estimation method that estimates $f_i(\cdot)$ nonparametrically conditioned on estimates of $f_j(\cdot)$ for all $j \neq i$ is used to estimate a NAAR model; see Chen and Tsay (1993b) for further details and examples of NAAR models.

The additivity assumption is rather restrictive and needs to be examined carefully in application. Chen, Liu, and Tsay (1995) consider test statistics for checking the additivity assumption.

Nonlinear State-Space Model

Making use of recent advances in MCMC methods (Gelfand and Smith, 1990), Carlin, Polson, and Stoffer (1992) propose a Monte Carlo approach for nonlinear state-space modeling. The model considered is

$$S_t = f_t(S_{t-1}) + u_t, \quad x_t = g_t(S_t) + v_t \quad (29)$$

where S_t is the state vector, $f_t(\cdot)$ and $g_t(\cdot)$ are known functions depending on some unknown parameters, $\{u_t\}$ is a sequence of IID multivariate random vectors with zero mean and non-negative definite covariance matrix Σu , $\{v_t\}$ is a sequence of IID random variables with mean zero and variance σ_v^2 , and $\{u_t\}$ is independent of $\{v_t\}$.

Monte Carlo techniques are employed to handle the nonlinear evolution of the state transition equation because the whole conditional distribution function of S_t given S_{t-1} is needed for a nonlinear system. Other numerical smoothing methods for nonlinear time series analysis have been considered by Kitagawa

(1998) and the references therein. MCMC methods (or computing-intensive numerical methods) are powerful tools for nonlinear time series analysis. Their potential has not been fully explored. However, the assumption of knowing $f_i(\cdot)$ and $g_i(\cdot)$ in model (29) may hinder practical use of the proposed method. A possible solution to overcome this limitation is to use nonparametric methods such as the analyses considered in FAR and NAAR models to specify $f_i(\cdot)$ and $g_i(\cdot)$ before using nonlinear state-space models.

Neural Networks

A popular topic in modern data analysis is neural network, which can be classified as a semiparametric method. The literature on neural network is enormous, and its application spreads over many scientific areas with varying degrees of success; see Ripley (1993, Sections 2 and 10). Cheng and Titterton (1994) provide information on neural networks from a statistical viewpoint. In this subsection, we focus solely on the feed-forward neural networks in which inputs are connected to one or more neurons, or nodes, in the input layer, and these nodes are connected forward to further layers until they reach the output layer. Figure 8 shows an example of a simple feed-forward network for univariate time series analysis with one hidden layer. The input layer has two nodes, and the hidden layer has three. The input nodes are connected forward to each and every node in the hidden layer, and these hidden nodes

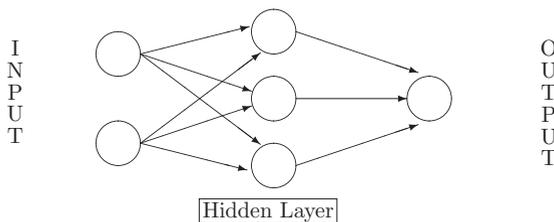


Figure 8 A Feed-Forward Neural Network with One Hidden Layer for Univariate Time Series Analysis

are connected to the single node in the output layer. We call the network a 2-3-1 feed-forward network. More complicated neural networks, including those with feedback connections, have been proposed in the literature, but the feed-forward networks are most relevant to our study.

Feed-Forward Neural Networks

A neural network processes information from one layer to the next by an “activation function.” Consider a feed-forward network with one hidden layer. The j th node in the hidden layer is defined as

$$h_j = f_j(\alpha_{0j} + \sum_{i \rightarrow j} w_{ij}x_i) \quad (30)$$

where x_i is the value of the i th input node, $f_j(\cdot)$ is an activation function typically taken to be the logistic function

$$f_j(z) = \frac{\exp(z)}{1 + \exp(z)}$$

α_{0j} is called the bias, the summation $i \rightarrow j$ means summing over all input nodes feeding to j , and w_{ij} are the weights. For illustration, the j th node of the hidden layer of the 2-3-1 feed-forward network in Figure 8 is

$$h_j = \frac{\exp(\alpha_{0j} + w_{1j}x_1 + w_{2j}x_2)}{1 + \exp(\alpha_{0j} + w_{1j}x_1 + w_{2j}x_2)}, \quad j = 1, 2, 3. \quad (31)$$

For the output layer, the node is defined as

$$o = f_o(\alpha_{0o} + \sum_{j \rightarrow o} w_{jo}h_j) \quad (32)$$

where the activation function $f_o(\cdot)$ is either linear or a Heaviside function. If $f_o(\cdot)$ is linear, then

$$o = \alpha_{0o} + \sum_{j=1}^k w_{jo}h_j$$

where k is the number of nodes in the hidden layer. By a Heaviside function, we mean $f_o(z) = 1$ if $z > 0$ and $f_o(z) = 0$ otherwise. A neuron with a Heaviside function is called a threshold neuron, with “1” denoting that the neuron fires its message. For example, the output of the

2-3-1 network in Figure 8 is

$$o = \alpha_{0o} + w_{1o}h_1 + w_{2o}h_2 + w_{3o}h_3$$

if the activation function is linear; it is

$$o = \begin{cases} 1 & \text{if } \alpha_{0o} + w_{1o}h_1 + w_{2o}h_2 + w_{3o}h_3 > 0 \\ 0 & \text{if } \alpha_{0o} + w_{1o}h_1 + w_{2o}h_2 + w_{3o}h_3 \leq 0 \end{cases}$$

if $f_o(\cdot)$ is a Heaviside function.

Combining the layers, the output of a feed-forward neural network can be written as

$$o = f_o \left[\alpha_{0o} + \sum_{j \rightarrow o} w_{jo} f_j \left(\alpha_{0j} + \sum_{i \rightarrow j} w_{ij} x_i \right) \right] \quad (33)$$

If one also allows for direct connections from the input layer to the output layer, then the network becomes

$$o = f_o \left[\alpha_{0o} + \sum_{i \rightarrow o} \alpha_{io} x_i + \sum_{j \rightarrow o} w_{jo} f_j \left(\alpha_{0j} + \sum_{i \rightarrow j} w_{ij} x_i \right) \right] \quad (34)$$

where the first summation is summing over the input nodes. When the activation function of the output layer is linear, the direct connections from the input nodes to the output node represent a linear function between the inputs and output. Consequently, in this particular case model (34) is a generalization of linear models. For the 2-3-1 network in Figure 8, if the output activation function is linear, then equation (33) becomes

$$o = \alpha_{0o} + \sum_{j=1}^3 w_{jo} h_j$$

where h_j is given in equation (31). The network thus has 13 parameters. If equation (34) is used, then the network becomes

$$o = \alpha_{0o} + \sum_{i=1}^2 \alpha_{io} x_i + \sum_{j=1}^3 w_{jo} h_j$$

where again h_j is given in equation (31). The number of parameters of the network increases to 15.

We refer to the function in equation (33) or (34) as a semiparametric function because its func-

tional form is known, but the number of nodes and their biases and weights are unknown. The direct connections from the input layer to the output layer in equation (34) mean that the network can skip the hidden layer. We refer to such a network as a skip-layer feed-forward network.

Feed-forward networks are known as multilayer perceptrons in the neural network literature. They can approximate any continuous function uniformly on compact sets by increasing the number of nodes in the hidden layer; see Hornik, Stinchcombe, and White (1989), Hornik (1993), and Chen and Chen (1995). This property of neural networks is the universal approximation property of the multilayer perceptrons. In short, feed-forward neural networks with a hidden layer can be seen as a way to parameterize a general continuous nonlinear function.

Training and Forecasting

Application of neural networks involves two steps. The first step is to train the network (i.e., to build a network, including determining the number of nodes and estimating their biases and weights). The second step is inference, especially forecasting. The data are often divided into two nonoverlapping subsamples in the training stage. The first subsample is used to estimate the parameters of a given feed-forward neural network. The network so built is then used in the second subsample to perform forecasting and compute its forecasting accuracy. By comparing the forecasting performance, one selects the network that outperforms the others as the "best" network for making inference. This is the idea of cross-validation widely used in statistical model selection. Other model selection methods are also available.

In a time series application, let $\{(r_t, \mathbf{x}_t) | t = 1, \dots, T\}$ be the available data for network training, where \mathbf{x}_t denotes the vector of inputs and r_t is the series of interest (e.g., log returns of an asset). For a given network, let o_t be the output of the network with input \mathbf{x}_t ; see equation (34). Training a neural network amounts to choosing

its biases and weights to minimize some fitting criterion—for example, the least squares

$$S^2 = \sum_{t=1}^T (r_t - o_t)^2$$

This is a nonlinear estimation problem that can be solved by several iterative methods. To ensure the smoothness of the fitted function, some additional constraints can be added to the prior minimization problem. In the neural network literature, the back propagation (BP) learning algorithm is a popular method for network training. The BP method, introduced by Bryson and Ho (1969), works backward starting with the output layer and uses a gradient rule to modify the biases and weights iteratively. (Appendix 2A of Ripley, 1993, provides a derivation of back propagation.) Once a feed-forward neural network is built, it can be used to compute forecasts in the forecasting subsample.

Example 7. To illustrate applications of the neural network in finance, we consider the monthly log returns, in percentages and including dividends, for IBM stock from January 1926 to December 1999. We divide the data into two subsamples. The first subsample consisting of returns from January 1926 to December 1997 for 864 observations is used for modeling. Using model (34) with three inputs and two nodes in the hidden layer, we obtain a 3-2-1 network for the series. The three inputs are r_{t-1} , r_{t-2} , and r_{t-3} and the biases and weights are given next:

$$\hat{r}_t = 3.22 - 1.81f_1(r_{t-1}) - 2.28f_2(r_{t-1}) - 0.09r_{t-1} - 0.05r_{t-2} - 0.12r_{t-3} \quad (35)$$

where $r_{t-1} = (r_{t-1}, r_{t-2}, r_{t-3})$ and the two logistic functions are

$$f_1(r_{t-1}) = \frac{\exp(-8.34 - 18.97r_{t-1} + 2.17r_{t-2} - 19.17r_{t-3})}{1 + \exp(-8.34 - 18.97r_{t-1} + 2.17r_{t-2} - 19.17r_{t-3})}$$

$$f_2(r_{t-1}) = \frac{\exp(39.25 - 22.17r_{t-1} - 17.34r_{t-2} - 5.98r_{t-3})}{1 + \exp(39.25 - 22.17r_{t-1} - 17.34r_{t-2} - 5.98r_{t-3})}$$

The standard error of the residuals for the prior model is 6.56. For comparison, we also built an AR model for the data and obtained

$$r_t = 1.101 + 0.077r_{t-1} + a_t, \quad \sigma_a = 6.61 \quad (36)$$

The residual standard error is slightly greater than that of the feed-forward model in equation (35).

Forecast Comparison

The monthly returns of IBM stock in 1998 and 1999 form the second subsample and are used to evaluate the out-of-sample forecasting performance of neural networks. As a benchmark for comparison, we use the sample mean of r_t in the first subsample as the 1-step ahead forecast for all the monthly returns in the second subsample. This corresponds to assuming that the log monthly price of IBM stock follows a random walk with drift. The mean squared forecast error (MSFE) of this benchmark model is 91.85. For the AR(1) model in equation (36), the MSFE of 1-step ahead forecasts is 91.70. Thus, the AR(1) model slightly outperforms the benchmark. For the 3-2-1 feed-forward network in equation (35), the MSFE is 91.74, which is essentially the same as that of the AR(1) model.

Example 8. Nice features of the feed-forward network include its flexibility and wide applicability. For illustration, we use the network with a Heaviside activation function for the output layer to forecast the direction of price movement for IBM stock considered in Example 7. Define a direction variable as

$$d_t = \begin{cases} 1 & \text{if } r_t \geq 0 \\ 0 & \text{if } r_t < 0 \end{cases}$$

We use eight input nodes consisting of the first four lagged values of both r_t and d_t and four nodes in the hidden layer to build an 8-4-1 feed-forward network for d_t in the first subsample. The resulting network is then used to compute the 1-step ahead probability of an “upward movement” (i.e., a positive return) for the following month in the second subsample.

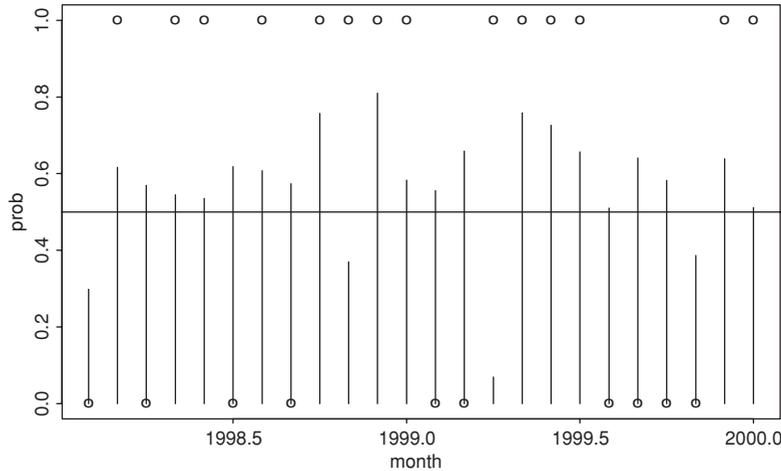


Figure 9 One-Step Ahead Probability Forecasts for a Positive Monthly Return for IBM Stock Using an 8-4-1 Feed-Forward Neural Network

Note: The forecasting period is from January 1998 to December 1999.

Figure 9 shows a typical output of probability forecasts and the actual directions in the second subsample with the latter denoted by circles. A horizontal line of 0.5 is added to the plot. If we take a rigid approach by letting $\hat{d}_t = 1$ if the probability forecast is greater than or equal to 0.5 and $\hat{d}_t = 0$ otherwise, then the neural network has a successful rate of 0.58. The success rate of the network varies substantially from one estimation to another, and the network uses 49 parameters.

To gain more insight, we did a simulation study of running the 8-4-1 feed-forward network 500 times and computed the number of errors in predicting the upward and downward movement using the same method as before. The mean and median of errors over the 500 runs are 11.28 and 11, respectively, whereas the maximum and minimum number of errors are 18 and 4. For comparison, we also did a simulation with 500 runs using a random walk with drift—that is,

$$\hat{d}_t = \begin{cases} 1 & \text{if } \hat{r}_t = 1.19 + \epsilon_t \geq 0, \\ 0 & \text{otherwise} \end{cases}$$

where 1.19 is the average monthly log return for IBM stock from January 1926 to December

1997 and $\{\epsilon_t\}$ is a sequence of IID $N(0,1)$ random variables. The mean and median of the number of forecast errors become 10.53 and 11, whereas the maximum and minimum number of errors are 17 and 5, respectively. Figure 10 shows the histograms of the number of forecast errors for the two simulations. The results show that the 8-4-1 feed-forward neural network does not outperform the simple model that assumes a random walk with drift for the monthly log price of IBM stock.

NONLINEARITY TESTS

In this section, we discuss some nonlinearity tests available in the literature that have decent power against the nonlinear models considered earlier in this entry. The tests discussed include both parametric and nonparametric statistics. The Ljung-Box statistics of squared residuals, the bispectral test, and the Brock, Dechert, and Scheinkman (BDS) test are nonparametric methods. The *RESET test* (Ramsey, 1969), the *F tests* of Tsay (1986, 1989), and other Lagrange multiplier and likelihood ratio tests depend on specific parametric functions. Because nonlinearity may occur in many ways, there exists no

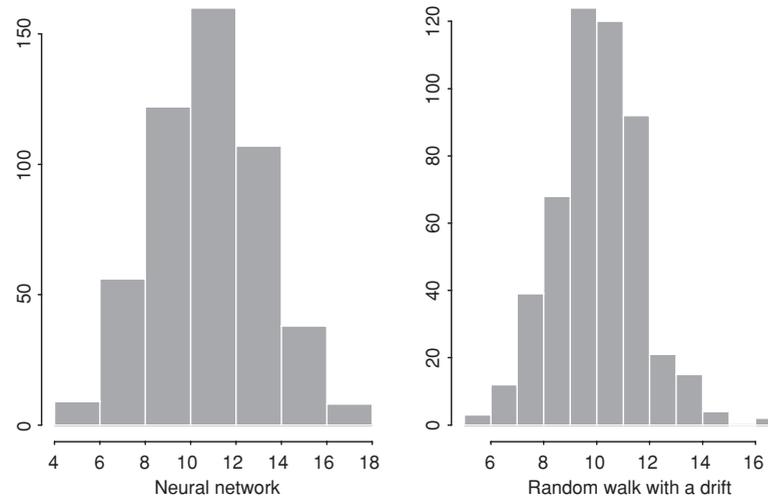


Figure 10 Histograms of the Number of Forecasting Errors for the Directional Movements of Monthly Log Returns of IBM Stock

Note: The forecasting period is from January 1998 to December 1999.

single test that dominates the others in detecting nonlinearity.

Nonparametric Tests

Under the null hypothesis of linearity, residuals of a properly specified linear model should be independent. Any violation of independence in the residuals indicates inadequacy of the entertained model, including the linearity assumption. This is the basic idea behind various nonlinearity tests. In particular, some of the nonlinearity tests are designed to check for possible violation in quadratic forms of the underlying time series.

Q-Statistic of Squared Residuals

McLeod and Li (1983) apply the Ljung-Box statistics to the squared residuals of an ARMA(p, q) model to check for model inadequacy. The test statistic is

$$Q(m) = T(T + 2) \sum_{i=1}^m \frac{\hat{\rho}_i^2(a_t^2)}{T - i}$$

where T is the sample size, m is a properly chosen number of autocorrelations used in the test,

a_t denotes the residual series, and $\hat{\rho}_i(a_t^2)$ is the lag- i ACF of a_t^2 . If the entertained linear model is adequate, $Q(m)$ is asymptotically a chi-squared random variable with $m - p - q$ degrees of freedom. The prior Q -statistic is useful in detecting conditional heteroscedasticity of a_t and is asymptotically equivalent to the Lagrange multiplier test statistic of Engle (1982) for ARCH models. The null hypothesis of the statistics is $H_0: \beta_1 = \dots = \beta_m = 0$, where β_i is the coefficient of a_{t-i}^2 in the linear regression

$$a_t^2 = \beta_0 + \beta_1 a_{t-1}^2 + \dots + \beta_m a_{t-m}^2 + e_t$$

for $t = m + 1, \dots, T$. Because the statistic is computed from residuals (not directly from the observed returns), the number of degrees of freedom is $m - p - q$.

Bispectral Test

This test can be used to test for linearity and Gaussianity. It depends on the result that a properly normalized bispectrum of a linear time series is constant over all frequencies and that the constant is zero under normality. The bispectrum of a time series is the Fourier transform of its third-order moments. For a

stationary time series x_t in equation (1), the third-order moment is defined as

$$c(u, v) = g \sum_{k=-\infty}^{\infty} \psi_k \psi_{k+u} \psi_{k+v} \quad (37)$$

where u and v are integers, $g = E(a_t^3)$, $\psi_0 = 1$, and $\psi_k = 0$ for $k < 0$. Taking Fourier transforms of equation (37), we have

$$b_3(w_1, w_2) = \frac{g}{4\pi^2} \Gamma[-(w_1 + w_2)] \Gamma(w_1) \Gamma(w_2) \quad (38)$$

where $\Gamma(w) = \sum_{u=0}^{\infty} \psi_u \exp(-i w u)$ with $i = \sqrt{-1}$, and w_i are frequencies. Yet the spectral density function of x_t is given by

$$p(w) = \frac{\sigma_a^2}{2\pi} |\Gamma(w)|^2$$

where w denotes the frequency. Consequently, the function

$$b(w_1, w_2) = \frac{|b_3(w_1, w_2)|^2}{p(w_1)p(w_2)p(w_1 + w_2)} \quad (39)$$

= constant for all (w_1, w_2)

The bispectrum test makes use of the property in equation (39). Basically, it estimates the function $b(w_1, w_2)$ in equation (39) over a suitably chosen grid of points and applies a test statistic similar to Hotelling's T^2 statistic to check the constancy of $b(w_1, w_2)$. For a linear Gaussian series, $E(a_t^3) = g = 0$ so that the bispectrum is zero for all frequencies (w_1, w_2) . For further details of the bispectral test, see Priestley (1988), Subba Rao and Gabr (1984), and Hinich (1982). Limited experience shows that the test has decent power when the sample size is large.

BDS Statistic

Brock, Dechert, and Scheinkman (1987) propose a test statistic, commonly referred to as the *BDS test*, to detect the IID assumption of a time series. The statistic is, therefore, different from other test statistics discussed because the latter mainly focus on either the second- or third-order properties of x_t . The basic idea of the BDS test is to make use of a "correlation integral"

popular in chaotic time series analysis. Given a k -dimensional time series X_t and observations $\{X_t\}_{t=1}^{T_k}$, define the correlation integral as

$$C_k(\delta) = \lim_{T_k \rightarrow \infty} \frac{2}{T_k(T_k - 1)} \sum_{i < j} I_\delta(X_i, X_j) \quad (40)$$

where $I_\delta(u, v)$ is an indicator variable that equals one if $\|u - v\| < \delta$, and zero otherwise, where $\|\cdot\|$ is the supnorm. The correlation integral measures the fraction of data pairs of $\{X_t\}$ that are within a distance of δ from each other.

Consider next a time series x_t . Construct k -dimensional vectors $X_t^k = (x_t, x_{t+1}, \dots, x_{t+k-1})'$, which are called k -histories. The idea of the BDS test is as follows. Treat a k -history as a point in the k -dimensional space. If $\{x_t\}_{t=1}^T$ are indeed IID random variables, then the k -histories $\{X_t^k\}_{t=1}^{T_k}$ should show no pattern in the k -dimensional space. Consequently, the correlation integrals should satisfy the relation $C_k(\delta) = [C_1(\delta)]^k$. Any departure from the prior relation suggests that x_t are not IID. As a simple but informative example, consider a sequence of IID random variables from the uniform distribution over $[0, 1]$. Let $[a, b]$ be a subinterval of $[0, 1]$ and consider the "2-history" (x_t, x_{t+1}) , which represents a point in the two-dimensional space. Under the IID assumption, the expected number of 2-histories in the subspace $[a, b] \times [a, b]$ should equal the square of the expected number of x_t in $[a, b]$.

This idea can be formally examined by using sample counterparts of correlation integrals. Define

$$C_\ell(\delta, T) = \frac{2}{T_\ell(T_\ell - 1)} \sum_{i < j} I_\delta(X_i^*, X_j^*), \quad \ell = 1, k$$

where $T_\ell = T - \ell + 1$ and $X_i^* = x_i$ if $\ell = 1$ and $X_i^* = X_i^k$ if $\ell = k$. Under the null hypothesis that $\{x_t\}$ are IID with a nondegenerated distribution function $F(\cdot)$, Brock, Dechert, and Scheinkman (1987) show that

$$C_k(\delta, T) \rightarrow [C_1(\delta)]^k \quad \text{with probability 1,} \\ \text{as } T \rightarrow \infty$$

for any fixed k and δ . Furthermore, the statistic $\sqrt{T}\{C_k(\delta, T) - [C_1(\delta, T)]^k\}$ is asymptotically distributed as normal with mean zero and variance

$$\sigma_k^2(\delta) = 4 \left(N^k + 2 \sum_{j=1}^{k-1} N^{k-j} C^{2j} + (k-1)^2 C^{2k} - k^2 N C^{2k-2} \right)$$

where $C = \int [F(z + \delta) - F(z - \delta)] dF(z)$ and $N = \int [F(z + \delta) - F(z - \delta)]^2 dF(z)$. Note that $C_1\{\delta, T\}$ is a consistent estimate of C , and N can be consistently estimated by

$$N(\delta, T) = \frac{6}{T_k(T_k - 1)(T_k - 2)} \sum_{t < s < u} I_\delta(x_t, x_s) I_\delta(x_s, x_u)$$

The BDS test statistic is then defined as

$$D_k(\delta, T) = \sqrt{T} \{C_k(\delta, T) - [C_1(\delta, T)]^k\} / \sigma_k(\delta, T) \quad (41)$$

where $\sigma_k(\delta, T)$ is obtained from $\sigma_k(\delta)$ when C and N are replaced by $C_1(\delta, T)$ and $N(\delta, T)$, respectively. This test statistic has a standard normal limiting distribution. For further discussion and examples of applying the BDS test, see Hsieh (1989) and Brock, Hsieh, and LeBaron (1991). In application, one should remove linear dependence, if any, from the data before applying the BDS test. The test may be sensitive to the choices of δ and k , especially when k is large.

Parametric Tests

Turning to parametric tests, we consider the RESET test of Ramsey (1969) and its generalizations. We also discuss some test statistics for detecting threshold nonlinearity.

The RESET Test

Ramsey (1969) proposes a specification test for linear least squares regression analysis. The test is referred to as a RESET test and is readily applicable to linear AR models. Consider the linear AR(p) model

$$x_t = \mathbf{X}'_{t-1} \boldsymbol{\phi} + a_t \quad (42)$$

where $\mathbf{X}_{t-1} = (1, x_{t-1}, \dots, x_{t-p})'$ and $\boldsymbol{\phi} = (\phi_0, \phi_1, \dots, \phi_p)'$. The first step of the RESET test is to obtain the least squares estimate $\hat{\boldsymbol{\phi}}$ of equation (42) and compute the fit $\hat{x}_t = \mathbf{X}'_{t-1} \hat{\boldsymbol{\phi}}$, the residual $\hat{a}_t = x_t - \hat{x}_t$, and the sum of squared residuals $SSR_0 = \sum_{t=p+1}^T \hat{a}_t^2$, where T is the sample size. In the second step, consider the linear regression

$$\hat{a}_t = \mathbf{X}'_{t-1} \boldsymbol{\alpha}_1 + \mathbf{M}'_{t-1} \boldsymbol{\alpha}_2 + v_t \quad (43)$$

where $\mathbf{M}_{t-1} = (\hat{x}_t^2, \dots, \hat{x}_t^{s+1})'$ for some $s \geq 1$, and compute the least squares residuals

$$\hat{v}_t = \hat{a}_t - \mathbf{X}'_{t-1} \hat{\boldsymbol{\alpha}}_1 - \mathbf{M}'_{t-1} \hat{\boldsymbol{\alpha}}_2$$

and the sum of squared residuals $SSR_1 = \sum_{t=p+1}^T \hat{v}_t^2$ of the regression. The basic idea of the RESET test is that if the linear AR(p) model in equation (42) is adequate, then $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ of equation (43) should be zero. This can be tested by the usual F statistic of equation (43) given by

$$F = \frac{(SSR_0 - SSR_1)/g}{SSR_1/(T - p - g)} \quad \text{with} \quad g = s + p + 1 \quad (44)$$

which, under the linearity and normality assumption, has an F distribution with degrees of freedom g and $T - p - g$.

Because \hat{x}_t^k for $k = 2, \dots, s + 1$ tend to be highly correlated with \mathbf{X}_{t-1} and among themselves, principal components of \mathbf{M}_{t-1} that are not colinear with \mathbf{X}_{t-1} are often used in fitting equation (43).

Keenan (1985) proposes a nonlinearity test for time series that uses \hat{x}_t^2 only and modifies the second step of the RESET test to avoid multicollinearity between \hat{x}_t^2 and \mathbf{X}_{t-1} . Specifically the linear regression (43) is divided into two steps. In step 2(a), one removes linear dependence of \hat{x}_t^2 on \mathbf{X}_{t-1} by fitting the regression

$$\hat{x}_t^2 = \mathbf{X}'_{t-1} \boldsymbol{\beta} + u_t$$

and obtaining the residual $\hat{u}_t = \hat{x}_t^2 - \mathbf{X}_{t-1} \hat{\boldsymbol{\beta}}$. In step 2(b), consider the linear regression

$$\hat{a}_t = \hat{u}_t \boldsymbol{\alpha} + v_t$$

and obtain the sum of squared residuals $SSR_1 = \sum_{t=p+1}^T (\hat{a}_t - \hat{u}_t \hat{\alpha})^2 = \sum_{t=p+1}^T \hat{v}_t^2$ to test the null hypothesis $\alpha = 0$.

The F Test

To improve the power of Keenan’s test and the RESET test, Tsay (1986) uses a different choice of the regressor M_{t-1} . Specifically, he suggests using $M_{t-1} = vech(X_{t-1} X'_{t-1})$, where $vech(A)$ denotes the half-stacking vector of the matrix A using elements on and below the diagonal only. For example, if $p = 2$, then $M_{t-1} = (x_{t-1}^2, x_{t-1}x_{t-2}, x_{t-2}^2)'$. The dimension of M_{t-1} is $p(p + 1)/2$ for an $AR(p)$ model. In practice, the test is simply the usual partial F statistic for testing $\alpha = 0$ in the linear least squares regression

$$x_t = X'_{t-1} \phi + M'_{t-1} \alpha + e_t$$

where e_t denotes the error term. Under the assumption that x_t is a linear $AR(p)$ process, the partial F statistic follows an F distribution with degrees of freedom g and $T - p - g - 1$, where $g = p(p + 1)/2$. We refer to this F test as the Ori- F test. Luukkonen, Saikkonen, and Teräsvirta (1988) further extend the test by augmenting M_{t-1} with cubic terms x_{t-i}^3 for $i = 1, \dots, p$.

Threshold Test

When the alternative model under study is a SETAR model, one can derive specific test statistics to increase the power of the test. One of the specific tests is the likelihood ratio statistic. This test, however, encounters the difficulty of undefined parameters under the null hypothesis of linearity because the threshold is undefined for a linear AR process. Another specific test seeks to transform testing threshold nonlinearity into detecting model changes. It is then interesting to discuss the differences between these two specific tests for threshold nonlinearity.

To simplify the discussion, let us consider the simple case that the alternative model is a 2-regime SETAR model with threshold variable x_{t-d} . The null hypothesis H_0 : x_t follows the lin-

ear $AR(p)$ model

$$x_t = \phi_0 + \sum_{i=1}^p \phi_i x_{t-i} + a_t \tag{45}$$

whereas the alternative hypothesis H_a : x_t follows the SETAR model

$$x_t = \begin{cases} \phi_0^{(1)} + \sum_{i=1}^p \phi_i^{(1)} x_{t-i} + a_{1t} & \text{if } x_{t-d} < r_1 \\ \phi_0^{(2)} + \sum_{i=1}^p \phi_i^{(2)} x_{t-i} + a_{2t} & \text{if } x_{t-d} \geq r_1 \end{cases} \tag{46}$$

where r_1 is the threshold. For a given realization $\{x_t\}_{t=1}^T$ and assuming normality let $l_0(\hat{\phi}, \hat{\sigma}_a^2)$ be the log likelihood function evaluated at the maximum likelihood estimates of $\phi = (\phi_0, \dots, \phi_p)'$ and σ_a^2 . This is easy to compute. The likelihood function under the alternative is also easy to compute if the threshold r_1 is given. Let $l_1(r_1; \hat{\phi}_1, \hat{\sigma}_1^2; \hat{\phi}_2, \hat{\sigma}_2^2)$ be the log likelihood function evaluated at the maximum likelihood estimates of $\phi_i = (\phi_0^{(i)}, \dots, \phi_p^{(i)})'$ and σ_i^2 conditioned on knowing the threshold r_1 . The log likelihood ratio $l(r_1)$ defined as

$$l(r_1) = l_1(r_1; \hat{\phi}_1, \hat{\sigma}_1^2; \hat{\phi}_2, \hat{\sigma}_2^2) - l_0(\hat{\phi}, \hat{\sigma}_a^2)$$

is then a function of the threshold r_1 , which is unknown. Yet under the null hypothesis, there is no threshold and r_1 is not defined. The parameter r_1 is referred to as a nuisance parameter under the null hypothesis. Consequently, the asymptotic distribution of the likelihood ratio is very different from that of the conventional likelihood ratio statistics. (See Chan, 1991, for further details and critical values of the test.) A common approach is to use $l_{\max} = \sup_{v < r_1 < u} l(r_1)$ as the test statistic, where v and u are pre-specified lower and upper bounds of the threshold. Davis (1987) and Andrews and Ploberger (1994) provide further discussion on hypothesis testing involving nuisance parameters under the null hypothesis. Simulation is often used to obtain empirical critical values of the test statistic l_{\max} , which depends on the choices of v and u . The average of $l(r_1)$ over $r_1 \in [v, u]$ is also considered by Andrews and Ploberger as a test statistic.

Tsay (1989) makes use of arranged autoregression and recursive estimation to derive an alternative test for threshold nonlinearity. The arranged autoregression seeks to transfer the SETAR model under the alternative hypothesis H_a into a model change problem with the threshold r_1 serving as the change point. To see this, the SETAR model in equation (46) says that x_t follows essentially two linear models depending on whether $x_{t-d} < r_1$ or $x_{t-d} \geq r_1$. For a realization $\{x_t\}_{t=1}^T$, x_{t-d} can assume values $\{x_1, \dots, X_{T-d}\}$. Let $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(T-d)}$ be the ordered statistics of $\{x_t\}_{t=1}^{T-d}$ (i.e., arranging the observations in increasing order). The SETAR model can then be written as

$$x_{(j)+d} = \beta_0 + \sum_{i=1}^p \beta_i x_{(j)+d-i} + a_{(j)+d}, \quad (47)$$

$$j = 1, \dots, T-d$$

where $\beta_i = \phi_i^{(1)}$ if $x_{(j)} < r_1$ and $\beta_i = \phi_i^{(2)}$ if $x_{(j)} \geq r_1$. Consequently, the threshold r_1 is a change point for the linear regression in equation (47), and we refer to equation (47) as an arranged autoregression (in increasing order of the threshold x_{t-d}). Note that the arranged autoregression in (47) does not alter the dynamic dependence of x_t on x_{t-i} for $i = 1, \dots, p$ because $x_{(j)+d}$ still depends on $x_{(j)+d-i}$ for $i = 1, \dots, p$. What is done is simply to present the SETAR model in the threshold space instead of in the time space. That is, the equation with a smaller x_{t-d} appears before that with a larger x_{t-d} . The *threshold test* of Tsay (1989) is obtained as follows.

- *Step 1.* Fit equation (47) using $j = 1, \dots, m$, where m is a prespecified positive integer (e.g., 30). Denote the least squares estimates of β_i by $\hat{\beta}_{i,m}$, where m denotes the number of data points used in estimation.
- *Step 2.* Compute the predictive residual

$$\hat{a}_{(m+1)+d} = x_{(m+1)+d} - \hat{\beta}_{0,m} - \sum_{i=1}^p \hat{\beta}_{i,m} x_{(m+1)+d-i}$$

and its standard error. Let $\hat{e}_{(m+1)+d}$ be the standardized predictive residual.

- *Step 3.* Use the recursive least squares method to update the least squares estimates to $\hat{\beta}_{i,m+1}$ by incorporating the new data point $x_{(m+1)+d}$.
- *Step 4.* Repeat steps 2 and 3 until all data points are processed.
- *Step 5.* Consider the linear regression of the standardized predictive residual

$$\hat{e}_{(m+j)+d} = \alpha_0 + \sum_{i=1}^p \alpha_i x_{(m+j)+d-i} + v_t, \quad (48)$$

$$j = 1, \dots, T-d-m$$

and compute the usual F statistic for testing $\alpha_i = 0$ in equation (48) for $i = 0, \dots, p$. Under the null hypothesis that x_t follows a linear AR(p) model, the F ratio has a limiting F distribution with degrees of freedom $p+1$ and $T-d-m-p$.

We refer to the earlier F test as a TAR- F test. The idea behind the test is that under the null hypothesis there is no model change in the arranged autoregression in equation (47) so that the standardized predictive residuals should be close to IID with mean zero and variance 1. In this case, they should have no correlations with the regressors $x_{(m+j)+d-i}$. For further details including formulas for a recursive least squares method and some simulation study on performance of the TAR- F test, see Tsay (1989). The TAR- F test avoids the problem of nuisance parameters encountered by the likelihood ratio test. It does not require knowing the threshold r_1 . It simply tests that the predictive residuals have no correlations with regressors if the null hypothesis holds. Therefore, the test does not depend on knowing the number of regimes in the alternative model. Yet the TAR- F test is not as powerful as the likelihood ratio test if the true model is indeed a 2-regime SETAR model with a known innovational distribution.

Applications

In this subsection, we apply some of the nonlinearity tests discussed previously to five time series. For a real financial time series, an AR model is used to remove any serial correlation

Table 2 Nonlinearity Tests for Simulated Series and Some Log Stock Returns

Data	Q		BDS ($\delta = 1.5\hat{\sigma}_a$)			
	(5)	(10)	2	3	4	5
N(0,1)	3.2	6.5	-0.32	-0.14	-0.15	-0.33
t_6	0.9	1.7	-0.87	-1.18	-1.56	-1.71
ln(ew)	2.9	4.9	9.94	11.72	12.83	13.65
ln(vw)	1.0	9.8	8.61	9.88	10.70	11.29
ln(ibm)	0.6	7.1	4.96	6.09	6.68	6.82
	$d = 1$		BDS($\delta = \hat{\sigma}_a$)			
Data	Ori-F	TAR-F	2	3	4	5
N(0,1)	1.13	0.87	-0.77	-0.71	-1.04	-1.27
t_6	0.69	0.81	-0.35	-0.76	-1.25	-1.49
ln(ew)	5.05	6.77	10.01	11.85	13.14	14.45
ln(vw)	4.95	6.85	7.01	7.83	8.64	9.53
ln(ibm)	1.32	1.51	3.82	4.70	5.45	5.72

Note: The sample size of simulated series is 500 and that of stock returns is 864. The BDS test uses $k = 2, \dots, 5$.

in the data, and the tests apply to the residual series of the model. The five series employed are as follows:

1. r_{1t} : A simulated series of IID $N(0,1)$ with 500 observations.
2. r_{2t} : A simulated series of IID Student- t distribution with 6 degrees of freedom. The sample size is 500.
3. a_{3t} : The residual series of monthly log returns of CRSP equal-weighted index from 1926 to 1997 with 864 observations. The linear AR model used is

$$(1 - 0.180B + 0.099B^3 - 0.105B^9)r_{3t} = 0.0086 + a_{3t}$$

4. a_{4t} : The residual series of monthly log returns of CRSP value-weighted index from 1926 to 1997 with 864 observations. The linear AR model used is

$$(1 - 0.098B + 0.111B^3 - 0.088B^5)r_{4t} = 0.0078 + a_{4t}$$

5. a_{5t} : The residual series of monthly log returns of IBM stock from 1926 to 1997 with 864 observations. The linear AR model used is

$$(1 - 0.077B)r_{5t} = 0.011 + a_{5t}$$

Table 2 shows the results of the nonlinearity test. For the simulated series and IBM returns, the F tests are based on an AR(6) model. For the index returns, the AR order is the same as the model given earlier. For the BDS test, we chose $\delta = \hat{\sigma}_a$ and $\delta = 1.5\hat{\sigma}_a$ with $k = 2, \dots, 5$. Also given in the table are the Ljung-Box statistics that confirm no serial correlation in the residual series before applying nonlinearity tests. Compared with their asymptotic critical values, the BDS test and F tests are insignificant at the 5% level for the simulated series. However, the BDS tests are highly significant for the real financial time series. The F tests also show significant results for the index returns, but they fail to suggest nonlinearity in the IBM log returns. In summary, the tests confirm that the simulated series are linear and suggest that the stock returns are nonlinear.

1 MODELING

Nonlinear time series modeling necessarily involves subjective judgment. However, there are some general guidelines to follow. It starts with building an adequate linear model on which nonlinearity tests are based. For financial time series, the Ljung-Box statistics and Engle's test

are commonly used to detect conditional heteroscedasticity. For general series, other tests discussed in the previous section apply. If nonlinearity is statistically significant, then one chooses a class of nonlinear models to entertain. The selection here may depend on the experience of the analyst and the substantive matter of the problem under study.

For volatility models, the order of an ARCH process can often be determined by checking the partial autocorrelation function of the squared series. For GARCH and exponential GARCH models, only lower orders such as (1,1), (1,2), and (2,1) are considered in most applications. Higher order models are hard to estimate and understand. For TAR models, one may use the procedures given in Tong (1990) and Tsay (1989, 1998) to build an adequate model. When the sample size is sufficiently large, one may apply nonparametric techniques to explore the nonlinear feature of the data and choose a proper nonlinear model accordingly; see Chen and Tsay (1993a) and Cai, Fan, and Yao (2000). The MARS procedure of Lewis and Stevens (1991) can also be used to explore the dynamic structure of the data.

Finally, information criteria such as the Akaike information criterion (Akaike, 1974) and the generalized odd ratios in Chen, McCulloch, and Tsay (1997) can be used to discriminate between competing nonlinear models. The chosen model should be carefully checked before it is used for prediction.

FORECASTING

Unlike the linear model, there exist no closed-form formulas to compute forecasts of most nonlinear models when the forecast horizon is greater than 1. We use parametric bootstraps to compute nonlinear forecasts. It is understood that the model used in forecasting has been rigorously checked and is judged to be adequate for the series under study. By a model, we mean the dynamic structure and innovational distri-

butions. In some cases, we may treat the estimated parameters as given.

Parametric Bootstrap

Let T be the forecast origin and ℓ be the forecast horizon ($\ell > 0$). That is, we are at time index T and interested in forecasting $x_{T+\ell}$. The parametric bootstrap considered computes realizations $x_{T+1}, \dots, x_{T+\ell}$ sequentially by (a) drawing a new innovation from the specified innovational distribution of the model, and (b) computing x_{T+1} using the model, data, and previous forecasts $x_{T+1}, \dots, x_{T+i-1}$. This results in a realization for $x_{T+\ell}$. The procedure is repeated M times to obtain M realizations of $x_{T+\ell}$ denoted by $\{x_{T+\ell}^{(j)}\}_{j=1}^M$. The point forecast of $X_{T+\ell}$ is then the sample average of $x_{T+\ell}^{(j)}$. Let the forecast be $x_T(\ell)$. We used $M = 3000$ in some applications and the results seem fine. The realizations $\{x_{T+\ell}^{(j)}\}_{j=1}^M$ can also be used to obtain an empirical distribution of $x_{T+\ell}$. We make use of this empirical distribution later to evaluate forecasting performance.

Forecasting Evaluation

There are many ways to evaluate the forecasting performance of a model, ranging from directional measures to magnitude measures to distributional measures. A directional measure considers the future direction (up or down) implied by the model. Predicting that tomorrow's S&P 500 index will go up or down is an example of directional forecasts that are of practical interest. Predicting the year-end value of the daily S&P 500 index belongs to the case of magnitude measure. Finally, assessing the likelihood that the daily S&P 500 index will go up 10% or more between now and the year end requires knowing the future conditional probability distribution of the index. Evaluating the accuracy of such an assessment needs a distributional measure.

In practice, the available data set is divided into two subsamples. The first subsample of the data is used to build a nonlinear model, and the second subsample is used to evaluate the

forecasting performance of the model. We refer to the two subsamples of data as estimation and forecasting subsamples. In some studies, a rolling forecasting procedure is used in which a new data point is moved from the forecasting subsample into the estimation subsample as the forecast origin advances. In what follows, we briefly discuss some measures of forecasting performance that are commonly used in the literature. Keep in mind, however, that there exists no widely accepted single measure to compare models. A utility function based on the objective of the forecast might be needed to better understand the comparison.

Directional Measure

A typical measure here is to use a 2 × 2 contingency table that summarizes the number of “hits” and “misses” of the model in predicting ups and downs of $x_{T+\ell}$ in the forecasting subsample. Specifically, the contingency table is given as

Actual	Predicted		
	up	down	
up	m_{11}	m_{12}	m_{10}
down	m_{21}	m_{22}	m_{20}
	m_{01}	m_{02}	m

where m is the total number of ℓ -step ahead forecasts in the forecasting subsample, m_{11} is the number of “hits” in predicting upward movements, m_{21} is the number of “misses” in predicting downward movements of the market, and so on. Larger values in m_{11} and m_{22} indicate better forecasts. The test statistic

$$X^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(m_{ij} - m_{i0}m_{0j}/m)^2}{m_{i0}m_{0j}/m}$$

can then be used to evaluate the performance of the model. A large χ^2 signifies that the model outperforms the chance of random choice. Under some mild conditions, χ^2 has an asymptotic chi-squared distribution with 1 degree of free-

dom. For further discussion of this measure, see Dahl and Hylleberg (1999).

For illustration of the directional measure, consider the 1-step ahead probability forecasts of the 8-4-1 feed-forward neural network shown in Figure 9. The 2 × 2 table of “hits” and “misses” of the network is

Actual	Predicted		
	up	down	
up	12	2	14
down	8	2	10
	20	4	24

The table shows that the network predicts the upward movement well, but fares poorly in forecasting the downward movement of the stock. The chi-squared statistic of the table is 0.137 with 77-value 0.71. Consequently, the network does not significantly outperform a random-walk model with equal probabilities for “upward” and “downward” movements.

Magnitude Measure

Three statistics are commonly used to measure performance of point forecasts. They are the mean squared error (MSE), mean absolute deviation (MAD), and mean absolute percentage error (MAPE). For ℓ -step ahead forecasts, these measures are defined as

$$MSE(\ell) = \frac{1}{m} \sum_{j=0}^{m-1} [x_{T+\ell+j} - x_{T+j}(\ell)]^2 \tag{49}$$

$$MAD(\ell) = \frac{1}{m} \sum_{j=0}^{m-1} |x_{T+\ell+j} - x_{T+j}(\ell)| \tag{50}$$

$$MAPE(\ell) = \frac{1}{m} \sum_{j=0}^{m-1} \left| \frac{x_{T+j}(\ell)}{x_{T+j+\ell}} - 1 \right| \tag{51}$$

where m is the number of ℓ -step ahead forecasts available in the forecasting subsample.

In application, one often chooses one of the above three measures, and the model with the smallest magnitude on that measure is regarded as the best ℓ -step ahead forecasting

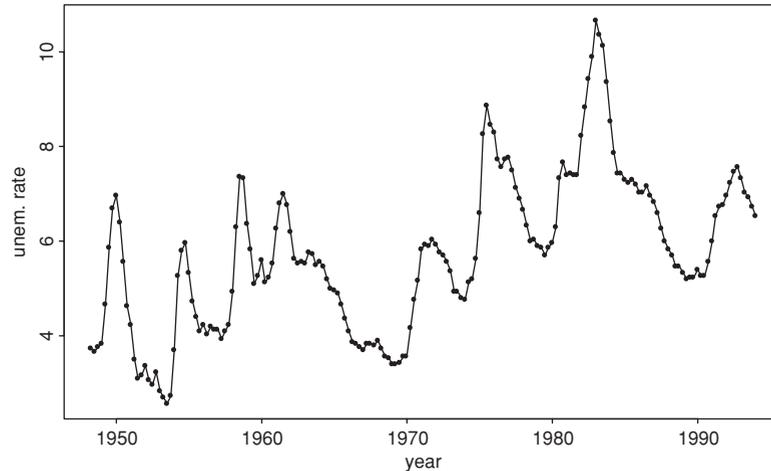


Figure 11 Time Plot of the U.S. Quarterly Unemployment Rate, Seasonally Adjusted, from 1948 to 1993

model. It is possible that different ℓ may result in selecting different models. The measures also have other limitations in model comparison; see, for instance, Clements and Hendry (1993).

Distributional Measure

Practitioners recently began to assess forecasting performance of a model using its predictive distributions. Strictly speaking, a predictive distribution incorporates parameter uncertainty in forecasts. We call it conditional predictive distribution if the parameters are treated as fixed. The empirical distribution of $x_{T+\ell}$ obtained by the parametric bootstrap is a conditional predictive distribution. This empirical distribution is often used to compute a distributional measure. Let $u_T(\ell)$ be the percentile of the observed $x_{T+\ell}$ in the prior empirical distribution. We then have a set of m percentiles $\{u_{T+j}(\ell)\}_{j=0}^{m-1}$, where again m is the number of ℓ -step ahead forecasts in the forecasting subsample. If the model entertained is adequate, $\{u_{T+j}(\ell)\}$ should be a random sample from the uniform distribution on $[0, 1]$. For a sufficiently large m , one can compute the Kolmogorov-Smirnov statistic of $\{u_{T+j}(\ell)\}$ with respect to uniform $[0, 1]$. The statistic can be used for both model checking and forecasting comparison.

2 APPLICATION

In this section, we illustrate nonlinear time series models by analyzing the quarterly U.S. civilian unemployment rate, seasonally adjusted, from 1948 to 1993. This series was analyzed in detail by Montgomery, Zarnowitz, Tsay, and Tiao (1998). We repeat some of the analyses here using nonlinear models. Figure 11 shows the time plot of the data. Well-known characteristics of the series include that (a) it tends to move countercyclically with U.S. business cycles, and (b) the rate rises quickly but decays slowly. The latter characteristic suggests that the dynamic structure of the series is nonlinear.

Denote the series by x_t and let $\Delta x_t = x_t - x_{t-1}$ be the change in unemployment rate. The linear model

$$(1 - 0.31B^4)(1 - 0.65B)\Delta x_t = (1 - 0.78B^4)a_t, \\ \hat{\sigma}_a^2 = 0.090 \quad (52)$$

was built by Montgomery et al. (1998), where the standard errors of the three coefficients are 0.11, 0.06, and 0.07, respectively. This is a seasonal model even though the data were seasonally adjusted. It indicates that the seasonal adjustment procedure used did not successfully remove the seasonality. This model is used as a benchmark model for forecasting comparison.

Table 3 Nonlinearity Test for Changes in the U.S. Quarterly Unemployment Rate: 1948.II–1993.IV

Type	Ori-F	LST	TAR(1)	TAR(2)	TAR(3)	TAR(4)
Test	2.80	2.83	2.41	2.16	2.84	2.98
<i>p</i> value	.0007	.0002	.0298	.0500	.0121	.0088

Note: An AR(5) model was used in the tests, where LST denotes the test of Luukkonen et al. (1988) and TAR(*d*) means threshold test with delay *d*.

To test for nonlinearity, we apply some of the nonlinearity tests discussed earlier in this entry with an AR(5) model for the differenced series Δx_t . The results are given in Table 3. All of the tests reject the linearity assumption. In fact, the linearity assumption is rejected for all AR(*p*) models we applied, where $p = 2, \dots, 10$.

Using a modeling procedure similar to that of Tsay (1989), Montgomery et al. (1998) build the following TAR model for the Δx_t series:

$$\Delta x_t = \begin{cases} 0.01 + 0.73\Delta x_{t-1} + 0.10\Delta x_{t-2} + a_{1t} & \text{if } \Delta x_{t-2} \leq 0.1, \\ 0.18 + 0.80\Delta x_{t-1} - 0.56\Delta x_{t-2} + a_{2t} & \text{otherwise} \end{cases} \quad (53)$$

The sample variances of a_{1t} and a_{2t} are 0.76 and 0.165, respectively, the standard errors of the three coefficients of regime 1 are 0.03, 0.10, and 0.12, respectively, and those of regime 2 are 0.09, 0.1, and 0.16. This model says that the change in the U.S. quarterly unemployment rate, Δx_t , behaves like a piecewise linear model in the reference space of $x_{t-2} - x_{t-3}$ with threshold 0.1. Intuitively, the model implies that the dynamics of unemployment act differently depending on the recent change in the unemployment rate. In the first regime, the unemployment rate has had either a decrease or a minor increase. Here the economy should be stable, and essentially the change in the rate follows a simple AR(1) model because the lag-2 coefficient is insignificant. In the second regime, there is a substantial jump in the unemployment rate (0.1 or larger). This typically corresponds to the contraction phase in the business cycle. It is also the period during which government interventions and industrial restructuring are likely to occur. Here Δx_t fol-

lows an AR(2) model with a positive constant, indicating an upward trend in x_t . The AR(2) polynomial contains two complex characteristic roots, which indicate possible cyclical behavior in Δx_t . Consequently, the chance of having a turning point in x_t increases, suggesting that the period of large increases in x_t should be short. This implies that the contraction phases in the U.S. economy tend to be shorter than the expansion phases.

Applying a Markov chain Monte Carlo method, Montgomery et al. (1998) obtain the following Markov switching model for Δx_t :

$$\Delta x_t = \begin{cases} -0.07 + 0.38\Delta x_{t-1} - 0.05\Delta x_{t-2} + \epsilon_{1t} & \text{if } s_t = 1 \\ 0.16 + 0.86\Delta x_{t-1} - 0.38\Delta x_{t-2} + \epsilon_{2t} & \text{if } s_t = 2 \end{cases} \quad (54)$$

The conditional means of Δx_t are -0.10 for $s_t = 1$ and 0.31 for $s_t = 2$. Thus, the first state represents the expansionary periods in the economy, and the second state represents the contractions. The sample variances of ϵ_{1t} and ϵ_{2t} are 0.031 and 0.192, respectively. The standard errors of the three parameters in state $s_t = 1$ are 0.03, 0.14, and 0.11, and those of state $s_t = 2$ are 0.04, 0.13, and 0.14, respectively. The state transition probabilities are $P(s_t = 2 | s_{t-1} = 1) = 0.084(0.060)$ and $P(s_t = 1 | s_{t-1} = 2) = 0.126(0.053)$, where the number in parentheses is the corresponding standard error. This model implies that in the second state the unemployment rate x_t has an upward trend with an AR(2) polynomial possessing complex characteristic roots. This feature of the model is similar to the second regime of the TAR model in equation (53). In the first state, the unemployment

rate x_t has a slightly decreasing trend with a much weaker autoregressive structure.

Forecasting Performance

A rolling procedure was used by Montgomery et al. (1998) to forecast the unemployment rate x_t . The procedure works as follows:

1. Begin with forecast origin $T = 83$, corresponding to 1968.II, which was used in the literature to monitor the performance of various econometric models in forecasting unemployment rate. Estimate the linear, TAR, and MSA models using the data from 1948.I to the forecast origin (inclusive).
2. Perform 1-quarter to 5-quarter ahead forecasts and compute the forecast errors of each model. Forecasts of nonlinear models used are computed by using the parametric bootstrap method explained earlier in this entry.
3. Advance the forecast origin by 1 and repeat the estimation and forecasting processes until all data are employed.
4. Use MSE and mean forecast error to compare performance of the models.

Table 4 shows the relative MSE of forecasts and mean forecast errors for the linear model in equation (52), the TAR model in equation (53), and the MSA model in equation (54), using the linear model as a benchmark. The comparisons are based on overall performance as well as the status of the U.S. economy at the forecast origin. From the table, we make the following observations:

1. For the overall comparison, the TAR model and the linear model are very close in MSE, but the TAR model has smaller biases. Yet the MSA model has the highest MSE and smallest biases.
2. For forecast origins in economic contractions, the TAR model shows improvements over the linear model both in MSE and bias. The MSA model also shows some improvement over the linear model, but the improvement is not as large as that of the TAR model.

Table 4 Out-of-Sample Forecast Comparison Among Linear, TAR, and MSA Models for the U.S. Quarterly Unemployment Rate

(A) Relative MSE of Forecast					
Model	1-step	2-step	3-step	4-step	5-step
(a) Overall Comparison					
Linear	1.00	1.00	1.00	1.00	1.00
TAR	1.00	1.04	0.99	0.98	1.03
MSA	1.19	1.39	1.40	1.45	1.61
MSE	0.08	0.31	0.67	1.13	1.54
(b) Forecast Origins in Economic Contractions					
Linear	1.00	1.00	1.00	1.00	1.00
TAR	0.85	0.91	0.83	0.72	0.72
MSA	0.97	1.03	0.96	0.86	1.02
MSE	0.22	0.97	2.14	3.38	3.46
(c) Forecast Origins in Economic Expansions					
Linear	1.00	1.00	1.00	1.00	1.00
TAR	1.06	1.13	1.10	1.15	1.17
MSA	1.31	1.64	1.73	1.84	1.87
MSE	0.06	0.21	0.45	0.78	1.24
(B) Mean of Forecast Errors					
Model	1-step	2-step	3-step	4-step	5-step
(a) Overall Comparison					
Linear	0.03	0.09	0.17	0.25	0.33
TAR	-0.10	-0.02	-0.03	-0.03	-0.01
MSA	0.00	-0.02	-0.04	-0.07	-0.12
(b) Forecast Origins in Economic Contractions					
Linear	0.31	0.68	1.08	1.41	1.38
TAR	0.24	0.56	0.87	1.01	0.86
MSA	0.20	0.41	0.57	0.52	0.14
(c) Forecast Origins in Economic Expansions					
Linear	-0.01	0.00	0.03	0.08	0.17
TAR	-0.05	-0.11	-0.17	-0.19	-0.14
MSA	-0.03	-0.08	-0.13	-0.17	-0.16

Note: The starting forecast origin is 1968.II, where the row marked by "MSE" shows the MSE of the benchmark linear model.

3. For forecast origins in economic expansions, the linear model outperforms both nonlinear models.

The results suggest that the contributions of nonlinear models over linear ones in forecasting the U.S. quarterly unemployment rate are

mainly in the periods when the U.S. economy is in contraction. This is not surprising because, as mentioned before, it is during the economic contractions that government interventions and industrial restructuring are most likely to occur. These external events could introduce nonlinearity in the U.S. unemployment rate. Intuitively, such improvements are important because it is during the contractions that people pay more attention to economic forecasts.

KEY POINTS

- Nonlinearity exists in many financial data, including log returns of widely used market indexes such as CRSP equal- and value-weight indexes.
- Nonlinearity also appears in asset volatility. Indeed, simple threshold models such as the threshold GARCH model can be used to better describe the behavior of asset volatility. The model has been used to model the leverage effect between return and volatility.
- Simple nonparametric methods such as the local linear regression method can be used to provide a deeper understanding of interest rate dynamics.
- The unemployment rate example shows that, even though nonlinear models may not outperform linear ones in all forecast origins, they can provide more accurate forecasts when the U.S. economy is under contraction. This is useful because people in general pay more attention to forecasts during economic recession.
- Among the nonlinear models, the Markov switching model has the smallest bias in out-of-sample prediction. The model, however, has a larger mean square of forecast errors than the threshold autoregressive model. This behavior is consistent with the structure of the model because the true states of the economy are never certain under the switching model.

REFERENCES

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC-19: 716–723.
- Andrews, D. W. K., and Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica* 62: 1383–1414.
- Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (2008). *Time Series Analysis: Forecasting and Control*, 4th ed. Hoboken, NJ: John Wiley.
- Brock, W., Dechert, W. D., and Scheinkman, J. (1987). A test for independence based on the correlation dimension. Working paper, Department of Economics, University of Wisconsin, Madison.
- Brock, W., Hsieh, D. A., and LeBaron, B. (1991). *Nonlinear Dynamics, Chaos and Instability: Statistical Theory and Economic Evidence*. Cambridge, MA: MIT Press.
- Bryson, A. E., and Ho, Y. C. (1969). *Applied Optimal Control*. New York: Blaisdell.
- Cai, Z., Fan, J., and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association* 95: 941–956.
- Carlin, B. P., Polson, N. G., and Stoffer, D. S. (1992). A Monte Carlo approach to nonnormal and nonlinear state space modeling. *Journal of the American Statistical Association* 87: 493–500.
- Chan, K. S. (1991). Percentage points of likelihood ratio tests for threshold autoregression. *Journal of the Royal Statistical Society Series B* 53: 691–696.
- Chan, K. S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *The Annals of Statistics* 21: 520–533.
- Chan, K. S., and Tong H. (1986). On estimating thresholds in autoregressive models. *Journal of Time Series Analysis* 7: 179–190.
- Chan, K. S., and Tsay, R. S. (1998). Limiting properties of the conditional least squares estimator of a continuous TAR model. *Biometrika* 85: 413–426.
- Chen, C., McCulloch, R. E., and Tsay, R. S. (1997). A unified approach to estimating and modeling univariate linear and nonlinear time series. *Statistica Sinica* 7: 451–472.
- Chen, R., and Tsay, R. S. (1991). On the ergodicity of TAR(1) processes. *Annals of Applied Probability* 1: 613–634.

- Chen, R., and Tsay, R. S. (1993a). Functional-coefficient autoregressive models. *Journal of the American Statistical Association* 88: 298–308.
- Chen, R., and Tsay, R. S. (1993b). Nonlinear additive ARX models. *Journal of the American Statistical Association* 88: 955–967.
- Chen, R., Liu, J., and Tsay, R. S. (1995). Additivity tests for nonlinear autoregressive models. *Biometrika* 82: 369–383.
- Chen, T., and Chen, H. (1995). Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks* 6: 911–917.
- Cheng, B., and Titterton, D. M. (1994). Neural networks: A review from a statistical perspective. *Statistical Science* 9: 2–54.
- Clements, M. P., and Hendry, D. F. (1993). On the limitations of comparing mean square forecast errors. *Journal of Forecasting* 12: 617–637.
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* 74: 829–836.
- Creal, D., Koopman, S. J., and Lucas, A. (2010). Generalized autoregressive score models with applications. Working paper, Booth School of Business, University of Chicago.
- Dahl, C. M., and Hylleberg, S. (1999). Specifying nonlinear econometric models by flexible regression models and relative forecast performance. Working paper, Department of Economics, University of Aarhus, Denmark.
- Davis, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 74: 33–43.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflations. *Econometrica* 50: 987–1007.
- Epanechnikov, V. (1969). Nonparametric estimates of a multivariate probability density. *Theory of Probability and Its Applications* 14: 153–158.
- Fan, J. (1993). Local linear regression smoother and their minimax efficiencies. *The Annals of Statistics* 21: 196–216.
- Fan, J., and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer-Verlag.
- Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85: 398–409.
- Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of nominal excess return on stocks. *Journal of Finance* 48: 1779–1801.
- Granger, C. W. J., and Andersen, A. P. (1978). *An Introduction to Bilinear Time Series Models*. Gottingen: Vandenhoeck and Ruprecht.
- Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57: 357–384.
- Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics* 45: 39–70.
- Hamilton, J. D. (1994). *Time Series Analysis*. Princeton, NJ: Princeton University Press.
- Härdle, W. (1990). *Applied Nonparametric Regression*. New York: Cambridge University Press.
- Hansen, B. E. (1997). Inference in TAR models. *Studies in Nonlinear Dynamics and Econometrics* 1: 119–131.
- Hinich, M. (1982). Testing for Gaussianity and linearity of a stationary time series. *Journal of Time Series Analysis* 3: 169–176.
- Hornik, K. (1993). Some new results on neural network approximation. *Neural Networks* 6: 1069–1072.
- Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks* 2: 359–366.
- Hsieh, D. A. (1989). Testing for nonlinear dependence in daily foreign exchange rates. *Journal of Business* 62: 339–368.
- Keenan, D. M. (1985). A Tukey non-additivity-type test for time series nonlinearity. *Biometrika* 72: 39–44.
- Kitagawa, G. (1998). A self-organizing state space model. *Journal of the American Statistical Association* 93: 1203–1215.
- Lewis, P. A. W., and Stevens, J. G. (1991). Nonlinear modeling of time series using multivariate adaptive regression spline (MARS). *Journal of the American Statistical Association* 86: 864–877.
- Liu, J., and Brockwell, P. J. (1988). On the general bilinear time-series model. *Journal of Applied Probability* 25: 553–564.
- Luukkonen, R., Saikkonen, P., and Teräsvirta, T. (1988). Testing linearity against smooth transition autoregressive models. *Biometrika* 75: 491–499.
- McCulloch, R. E., and Tsay, R. S. (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *Journal of the American Statistical Association* 88: 968–978.

- McCulloch, R. E., and Tsay, R. S. (1994). Statistical inference of macroeconomic time series via Markov switching models. *Journal of Time Series Analysis* 15: 523–539.
- McLeod, A. I., and Li, W. K. (1983). Diagnostic checking ARMA time series models using squared-residual autocorrelations. *Journal of Time Series Analysis* 4: 269–273.
- Montgomery, A. L., Zarnowitz, V., Tsay, R. S., and Tiao, G. C. (1998). Forecasting the U. S. unemployment rate. *Journal of the American Statistical Association* 93: 478–493.
- Nadaraya, E. A. (1964). On estimating regression. *Theory and Probability Application* 10: 186–190.
- Petrucelli, J., and Woolford, S. W. (1984). A threshold AR(1) model. *Journal of Applied Probability* 21: 270–286.
- Potter, S. M. (1995). A nonlinear approach to U.S. GNP. *Journal of Applied Econometrics* 10: 109–125.
- Priestley, M. B. (1980). State-dependent models: A general approach to nonlinear time series analysis. *Journal of Time Series Analysis* 1: 47–71.
- Priestley, M. B. (1988). *Non-linear and Non-stationary Time Series Analysis*. London: Academic Press.
- Ramsey, J. B. (1969). Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society Series B* 31: 350–371.
- Ripley, B. D. (1993). Statistical aspects of neural networks. In O. E. Barndorff-Nielsen, J. L. Jensen, and W. S. Kendall (eds.), *Networks and Chaos—Statistical and Probabilistic Aspects*, pp. 40–123. London: Chapman and Hall.
- Subba Rao, T., and Gabr, M. M. (1984). An introduction to bispectral analysis and bilinear time series models. *Lecture Notes in Statistics*, 24. New York: Springer-Verlag.
- Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association* 89: 208–218.
- Tiao, G. C., and Tsay, R. S. (1994). Some advances in nonlinear and adaptive modeling in time series. *Journal of Forecasting* 13: 109–131.
- Tong, H. (1978). On a threshold model. In C. H. Chen (ed.), *Pattern Recognition and Signal Processing*. Amsterdam: Sijhoff & Noordhoff.
- Tong, H. (1983). Threshold models in nonlinear time series analysis. *Lecture Notes in Statistics*. New York: Springer-Verlag.
- Tong, H. (1990). *Non-Linear Time Series: A Dynamical System Approach*. Oxford: Oxford University Press.
- Tsay, R. S. (1986). Nonlinearity tests for time series. *Biometrika* 73: 461–466.
- Tsay, R. S. (1989). Testing and modeling threshold autoregressive processes. *Journal of the American Statistical Association* 84: 231–240.
- Tsay, R. S. (1998). Testing and modeling multivariate threshold models. *Journal of the American Statistical Association* 93: 1188–1202.
- Tsay, R. S. (2010). *Analysis of Financial Time Series*, 3rd ed. Hoboken, NJ: Wiley.
- Venables, W. N., and Ripley, B. D. (1999). *Modern Applied Statistics with S-Plus*, 3rd ed. New York: Springer-Verlag.
- Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A* 26: 359–372.

Robust Estimates of Betas and Correlations

THOMAS K. PHILIPS, PhD

Regional Head of Investment Risk and Performance, BNP Paribas Investment Partners

Abstract: The Theil-Sen estimator is an exceptionally simple and robust linear regression estimator, affording estimates of slope and intercept that are virtually identical to their ordinary least squares counterparts in the absence of outliers, but which do not change appreciably in the presence of outliers. In fact, with univariate data, it improves on ordinary least squares in almost every way imaginable, and it is therefore a striking fact that this remarkable estimator is not universally known and used. It can be used to derive robust estimates of beta and the correlation coefficient that are virtually identical to their classical counterparts when asset returns are normally distributed, and which are significantly more robust when asset returns are highly skewed or contaminated with outliers.

Point estimates of betas and correlations are most often obtained using *ordinary least squares (OLS)* and the standard maximum likelihood estimator, respectively. While these estimators are clearly optimal when asset returns are normally distributed, and when we hold no view on their prior distribution, they can be far from optimal when these conditions are not satisfied. In this entry, a novel explanation of OLS is provided and is then used to motivate a *robust* univariate *regression* algorithm due to Theil (1950) and Sen (1968). This estimator is then used to obtain remarkably robust (i.e., outlier resistant) estimates of asset *betas*, asset *correlations*, and *non-negative definite* correlation and *covariance* matrices.

OLS REVISITED

Generations of students have learned OLS in the way depicted pictorially in Figure 1. We are given a set of points, each with an abscissa (or x value) and an ordinate (or y value), and which are displayed on a scatter plot in the $X - Y$ plane. All errors are assumed to be concentrated in the ordinates. The abscissae are assumed to be known with certainty. The i^{th} point has coordinates (x_i, y_i) , and the collection of points visually evidences a noisy, but linear, relationship between the x 's and the y 's. The object of OLS is to find a straight line, the line of best fit, with slope β_{OLS} and intercept α_{OLS} , and which minimizes the sum of squared vertical distances (or errors) from the points to this line.

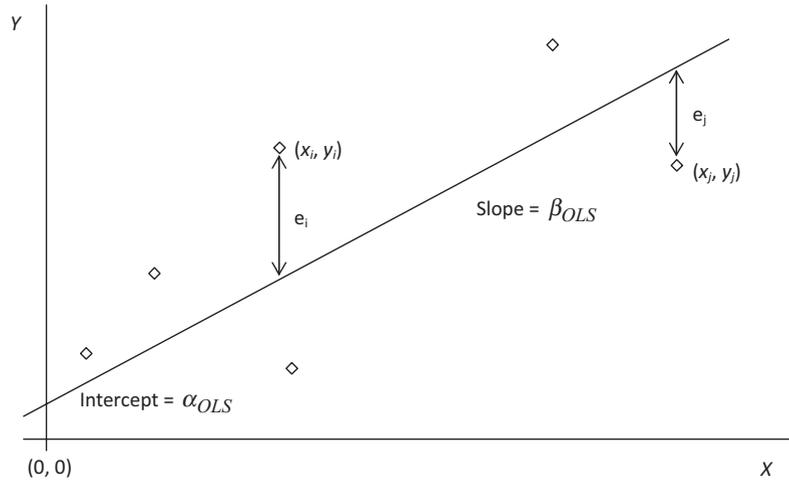


Figure 1 Ordinary Least Squares—Classical Depiction

This pictorial representation has become so firmly embedded in our consciousness that we take its geometry and its formulation for granted. But consider that the method dates back to 1800, and the fact that it was independently discovered by Carl Friedrich Gauss and Joseph-Louis Lagrange, who surely rank among the greatest mathematicians of all time, and it should come as no surprise that this textbook depiction of OLS hides more than one secret. In this section, we expose two of its secrets.

We start our exploration of OLS with Figure 2, which plots the same set of points as Figure 1, but now, instead of drawing a single line of best fit through the entire data set, we choose two specific points, (x_i, y_i) and (x_j, y_j) , draw the unique straight line joining them and project it back till it intersects the Y axis. This line has slope β_{ij} and intercept α_{ij} , where β_{ij} and α_{ij} are given by:

$$\beta_{ij} = \frac{y_i - y_j}{x_i - x_j}, \tag{1}$$

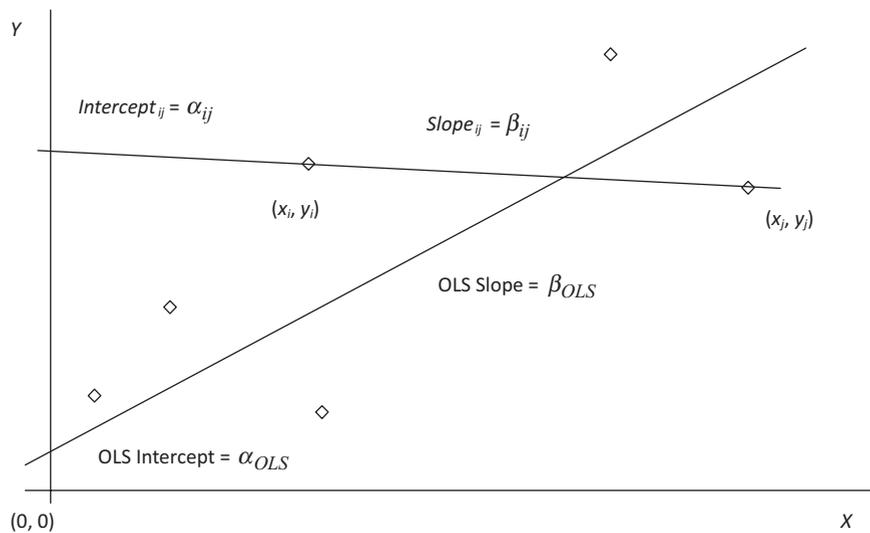


Figure 2 Ordinary Least Squares—Alternative Depiction

and

$$\alpha_{ij} = \frac{x_i \times y_j - x_j \times y_i}{x_i - x_j} \quad (2)$$

On comparing Figures 1 and 2, it is clear that β_{OLS} must necessarily lie between $\min_{i,j} \beta_{ij}$ and $\max_{i,j} \beta_{ij}$, both endpoints inclusive, and that α_{OLS} must likewise lie between $\min_{i,j} \alpha_{ij}$ and $\max_{i,j} \alpha_{ij}$. The OLS slope and intercept can therefore be written as weighted averages of all $\binom{N}{2} = \frac{N(N-1)}{2}$ pairwise slopes and intercepts for some nonnegative sets of weights, that is,

$$\beta_{OLS} = \sum_i \sum_j w_{ij} \beta_{ij}, \quad \sum_i \sum_j w_{ij} = 1, \quad w_{ij} \geq 0 \quad (3)$$

and

$$\alpha_{OLS} = \sum_i \sum_j v_{ij} \alpha_{ij}, \quad \sum_i \sum_j v_{ij} = 1, v_{ij} \geq 0 \quad (4)$$

In any particular situation, these weights are not unique, as equations (3) and (4) are enormously overdetermined, and we therefore seek a set of strictly positive weights that simultaneously solves both equations for an arbitrary collection of points. Such a set of weights can be identified using some clever guesswork motivated by the following observation: If (x_i, y_i) and (x_j, y_j) are close together, then any error in either ordinate will induce significant errors in β_{ij} and α_{ij} . Pairs of points that are far apart are much less susceptible to this problem. We ought, therefore, to overweight slopes and intercepts derived from pairs of points that are far apart relative to those that are derived from pairs of points that are close together.

Next, as all the error is concentrated in the abscissae, and as the ordinates are known with certainty, the weights must be a function only of $(x_i - x_j)$ —they cannot depend on $(y_i - y_j)$. Finally, the function must be even, because some weights would be negative if it were odd. Some tedious and not particularly enlightening alge-

bra shows our intuition to be correct, that is,

$$w_{ij} = v_{ij} = \frac{(x_i - x_j)^2}{\sum_k \sum_l (x_k - x_l)^2} \quad (5)$$

It follows that

$$\beta_{OLS} = \frac{\sum_i \sum_j \beta_{ij} (x_i - x_j)^2}{\sum_k \sum_l (x_k - x_l)^2} \quad (6)$$

and

$$\alpha_{OLS} = \frac{\sum_i \sum_j \alpha_{ij} (x_i - x_j)^2}{\sum_k \sum_l (x_k - x_l)^2} \quad (7)$$

Equations (6) and (7) yield OLS' first little secret—the line of best fit is just an appropriately weighted average of all possible lines that could be drawn using this data set! While this argument does not readily extend to the multivariate case, it does give us a fresh perspective on OLS, which now stands exposed as a clever and computationally efficient weighting scheme over the set of unique straight lines drawn through all possible pairs of points. A proof of this result, which is usually credited to Sen (1968), can be found in Heitman and Ord (1985).

Its second little secret lies in its focus on squared errors. Why should it be the second, and not the fourth or the sixty-fourth power of the error that is minimized? To answer this question, recall the way in which the OLS slope and intercept are defined:

$$\begin{aligned} \alpha_{OLS}, \beta_{OLS} &= \arg \min \sum_i error_i^2 \\ &= \arg \min \sum_i (y_i - \alpha_{OLS} - \beta_{OLS} \times x_i)^2 \end{aligned} \quad (8)$$

Solving this minimization problem requires us to compute the partial derivatives of the sum of squared errors with respect to α_{OLS} and β_{OLS} , and to equate the resulting expressions to 0. This results in a set of linear equations that can be solved in closed form (the solution was known to both Gauss and Lagrange). If, however, the error is raised to a power other than

two, we would have to solve a set of nonlinear equations, which do not, in general, admit a closed form solution—they must be solved numerically on a computer, a tool that neither Gauss nor Lagrange had access to. That said, raising the error to any even power (or even making it the argument of any even function) and then performing the indicated minimization numerically will result in a line that is optimal under that measure, though its slope and intercept will not, in general, equal the OLS slope and intercept.

All of this leads to our second insight—the mathematical formulation of OLS is driven by thoroughly practical considerations. In 1800, anything else simply couldn't be (and for that matter, still can't be) solved analytically! Having exposed these two little secrets of OLS, we now describe a better way in which to compute univariate regressions and explore its application to the estimation of beta and the correlation coefficient, as well as to the estimation of positive definite correlation and covariance matrices.

THEIL-SEN REGRESSION

The *Theil-Sen regression* algorithm (Thiel, 1950; Sen, 1968) is a robust alternative to univariate OLS that performs particularly well in the presence of outliers (loosely, in the presence of large, sporadic errors that are anything but Gaussian). It has long been known that OLS is acutely sensitive to errors in its inputs, and it is immediately apparent from equations (6) and (7) that even a single outlier can induce arbitrarily large errors in β_{OLS} and α_{OLS} .

Thiel (1950) and Sen (1968) propose a novel solution to this problem—they propose using the median of all $\binom{N}{2} = \frac{N(N-1)}{2}$ slopes to estimate the slope of the regression line, and choose the intercept to force the median error to 0. The primary difference between their methods is that Thiel uses all available observations, while Sen restricts attention to the subset of observations with distinct abscissae, that is, the set of

points for which $x_i \neq x_j$, and replaces each set of points that share the same abscissa with a single point whose ordinate is the average of their ordinates.

Formally, the Theil-Sen estimates of slope and intercept are given by:

$$\beta_{TS} = \text{median}_{i,j} \{ \beta_{ij} \} \quad (9)$$

and

$$\alpha_{TS} = \text{median}_{i,j} \{ y_i - \beta_{TS} \times x_i \} \quad (10)$$

This regression has been widely studied. Peng, Wang, and Wang (2008), for example, show that it is strongly consistent and super-efficient, and derive its asymptotic distribution. Interestingly, the median has long been used as a robust estimator of the mean for symmetric distributions, but this appears to be the first known application of the median to the estimation of regression coefficients.

We illustrate the superiority of Theil-Sen regression over OLS via simulations, the results of which are displayed in Tables 1 and 2. When the distribution of errors is normal, the distributions of β_{TS} and α_{TS} are almost identical to those of β_{OLS} and α_{OLS} . When the errors are drawn from a highly skewed distribution, or when the data are contaminated with significant amounts of noise, the distributions of β_{TS} and α_{TS} are far less variable than those of β_{OLS} and α_{OLS} .

These results are generated as follows. Using a high-quality random number generator (Mersenne twister), we create two independent random vectors, X and Y , both of length 100, and drawn from one of two distributions—unit normal and Pareto(2). We then regress Y against X using both OLS and the Theil-Sen regression. As the vectors are independent, the distribution of the slope and the intercept of the regression lines should be centered at 0 and $E[X]$, respectively.

We run 10,000 simulations to ensure that the 99% confidence intervals on our estimates are extremely tight (the width of the confidence interval is inversely proportional to the square root of the number of simulation runs), and

Table 1 Theil-Sen Regression vs. OLS: Normally Distributed Random Variables

Percentiles	1	5	10	25	50	75	90	95	99
Theil-Sen Slope	-0.26	-0.17	-0.14	-0.07	0	0.07	0.13	0.18	0.26
Least Squares Slope	-0.24	-0.17	-0.13	-0.07	0	0.07	0.13	0.17	0.24
Theil-Sen Intercept	-0.3	-0.21	-0.16	-0.09	0	0.08	0.16	0.21	0.3
Least Squares Intercept	-0.23	-0.17	-0.13	-0.07	0	0.07	0.13	0.17	0.23
Theil-Sen Mean Square	0.69	0.77	0.81	0.89	0.98	1.08	1.17	1.24	1.35
Least Squares Mean Square Err	0.69	0.76	0.81	0.88	0.98	1.07	1.16	1.23	1.33
Theil-Sen Median	0	0	0	0	0	0	0	0	0
Least Squares Median Error	-0.17	-0.12	-0.09	-0.05	0	0.05	0.09	0.12	0.18

Tables 1 and 2 display various percentiles of the distribution of the slope, the intercept, and the mean squared error (i.e., the sum of squared errors divided by 100) for both OLS and the Theil-Sen regression.

The first simulation, for normal random variables, illustrates how close the Theil-Sen algorithm is to OLS in the special case when OLS is clearly optimal. The second simulation illustrates its robustness with Pareto(2) random variables, whose distribution is highly skewed, and whose long tails serve as proxies for outliers.

When X and Y are normally distributed (Table 1), the median slope, the interquartile range for the slope (the difference between the 75th and the 25th percentiles), and the MSE for the two algorithms are essentially identical. The same holds true even when we look at a 90% range (the difference between the 5th and the 95th percentiles). However, when X and Y are drawn from a Pareto(2) distribution (Table 2), the performance of the two algorithms diverges: The interquartile range for the slope is 40% smaller for the Theil-Sen regression (0.06 vs. 0.10 for OLS) and an astonishing 60% smaller

for the 90% range (0.16 vs. 0.41), though the median MSE rises by about 12%.

The median slope remains 0 for the Theil-Sen regression, but exhibits a slight downward bias for OLS. The range of the intercept for the Theil-Sen regression is slightly larger than it is for OLS, but this is driven entirely by the fact that the Theil-Sen intercept is chosen to force the median error to 0, while the OLS intercept is chosen to minimize the sum of squared errors.

These simulations clearly show that the Theil-Sen regression gives up nothing to OLS when X and Y are normally distributed and is at a significant advantage when they are not. Similar results are obtained when either X or Y is contaminated with outliers. In all such experiments, the advantage of the Theil-Sen approach is readily apparent. In fact, it can be shown that as many as 29% of the data points can be corrupted with errors of arbitrary size before the Theil-Sen estimates of slope and intercept start to break down.

Given these results, and the accompanying fact that the vast majority of regressions run in practice are univariate, it is surprising that the Theil-Sen regression is not more widely used

Table 2 Theil-Sen Regression vs. OLS: Pareto(2) Distributed Random Variables

Percentiles	1	5	10	25	50	75	90	95	99
Theil-Sen Slope	-0.12	-0.07	-0.05	-0.03	0	0.03	0.06	0.09	0.14
Least Squares Slope	-0.35	-0.18	-0.13	-0.07	-0.02	0.03	0.13	0.23	0.65
Theil-Sen Intercept	1.17	1.24	1.28	1.34	1.41	1.49	1.57	1.62	1.74
Least Squares Intercept	0.98	1.49	1.62	1.78	1.95	2.16	2.41	2.62	3.44
Theil-Sen Mean Sqr Err	0.59	0.88	1.11	1.67	2.83	5.46	12.29	22.67	107.54
Least Squares Mean Sqr Err	0.52	0.78	0.98	1.47	2.53	4.97	11.54	21.47	104.8
Theil-Sen Median Error	0	0	0	0	0	0	0	0	0
Least Squares Median Error	-1.51	-0.97	-0.83	-0.64	-0.5	-0.4	-0.32	-0.28	-0.22

and appreciated. This may in part be driven by the fact that Theil-Sen regression, unlike OLS, does not generalize naturally to the case where there are many independent variables, as the median is inherently a one-dimensional measure.

A number of attempts have been made to create multivariate extensions of the Theil-Sen regression, the two most popular ones being the iterative Gauss-Seidel method described by Hastie and Tibshirani (1990) and the elemental subset method of Oja and Niinimaa (1984), which is described in Rousseeuw and Leroy (1987). Unfortunately, neither approach is entirely reliable in practice, and it is easy to find simple examples for which they converge to the wrong solution, particularly when the relationship being modeled is nonlinear.

ROBUST ESTIMATES OF BETA

The beta of an asset Y with respect to the market portfolio X plays a central role in modern finance as a result of the capital asset pricing model (Treynor, 1961; Sharpe, 1964; Lintner,

1965; and Mossin, 1966), and is defined to be

$$\beta_{Y|X} = \frac{\text{Cov}(X, Y)}{\sigma_X^2} \quad (11)$$

This quantity is, of course, just the slope coefficient in a univariate regression, and is precisely what OLS estimates. The application of the Theil-Sen regression algorithm to the estimation of beta is obvious—the Theil-Sen estimate of slope ought to provide us a more robust estimate of the historical of a security than the corresponding OLS estimate.

The advantages of the Theil-Sen estimator of beta are made clear by the following estimate of the beta of IBM around the crash of 1987. Starting on July 1, 1987, and ending on December 31, 1987, we estimate IBM's β by regressing its daily return for the most recent 132 days (or 6 months) against the corresponding daily return of the S&P 500. As can be seen from Figure 3, the Theil-Sen estimate is far more stable than the OLS estimate. In particular, it does not jump sharply after the 20% drop in the S&P 500 on October 19 as does the OLS beta, just as one would expect given its robustness.

While this is clearly an extreme example of a single outlier corrupting an estimate of beta,

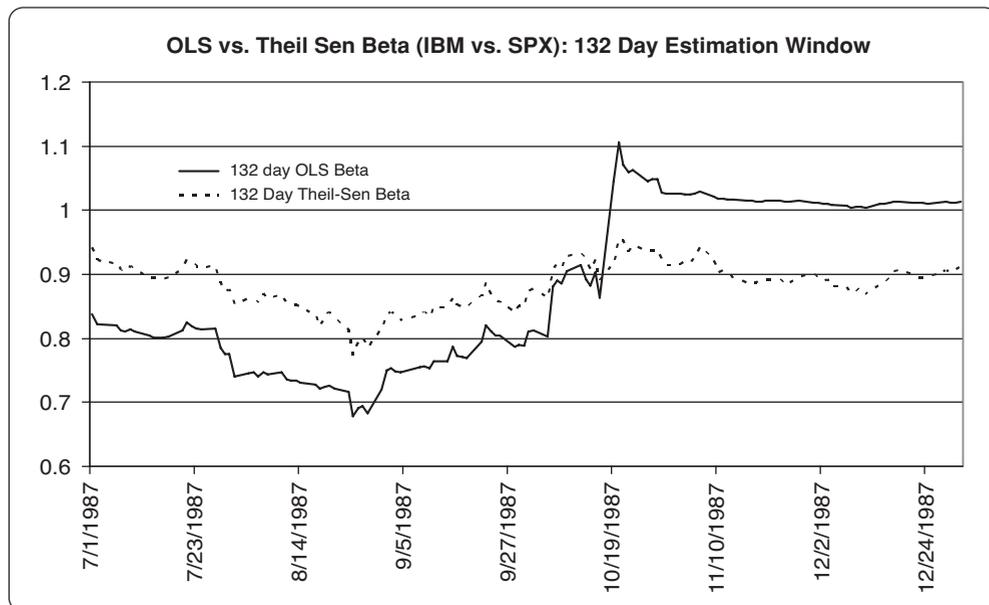


Figure 3 OLS vs. Theil-Sen Estimates of Beta: July 1, 1987, to December 31, 1987

outliers in financial data are far more common than is usually assumed, and they are not easily detected, as they influence many classical estimators in a way that masks their presence. One popular method of identifying and removing outliers is to remove points that lie more than three sample standard deviations away from the regression line.

Unfortunately, outliers can so distort the slope and intercept of the regression line, as well as the sample standard deviation of the errors, that all the points, including the outliers, will be found to lie within three sample standard deviations of the regression line! In general, filtering data using classical estimators to identify outliers works poorly in practice, and it proves far more effective to use estimators that are inherently robust to outliers.

The Theil-Sen estimate of beta can be further adjusted for the effects of nonsynchronous trading using the Scholes-Williams (1977) or Dimson (1979) corrections and can be shrunk cross-sectionally using a Bayesian correction as is done in Vasicek (1973). In each case, the Theil-Sen estimates of beta will provide a more robust point from which to start building an enhanced estimate of beta.

The Dimson correction sums contemporaneous and lagged betas for the asset to create an overall beta that accounts for the fact that an asset may have both a contemporaneous and a lagged response to market shocks, that is,

$$\beta_{Y|X}^{Dimson} = \sum_{i=0}^k \beta_{Y_t|X_{t-i}} \quad (12)$$

When using daily data, k is commonly set to 4 (i.e., one week's data), and when using monthly data, it is most commonly set to 1, so as not to pick up spurious responses from shocks in the distant past.

Vasicek's (1973) correction is a Bayesian correction, which allows the user to reflect information gleaned from the (known) cross-sectional distribution of betas to enhance an unconditional estimate of beta. In particular, Vasicek (1973) assumes that the prior distribution of

betas is normal and shows that the maximum a posteriori estimate of beta is a linear combination of its initial estimate and its cross-sectional mean, that is,

$$\beta_{Y|X}^{Vasicek} = w_Y \times \beta_{Y|X} + (1 - w_Y) \times \beta_{average} \quad (13)$$

where

$$w_Y = \frac{\sigma_{Cross-Sectional}^2}{\sigma_{\beta(Y|X)}^2 + \sigma_{Cross-Sectional}^2} \quad (14)$$

$\sigma_{\beta(Y|X)}^2$ is the variance of $\beta_{Y|X}$, and $\sigma_{Cross-Sectional}^2$ is the cross-sectional variance of the betas of the entire universe of securities under consideration at this point in time. A particularly simple and reasonably effective implementation of this method sets $w_Y = 0.5$ for all assets and at all points in time. Both techniques see use in the enhanced estimation of beta across a wide range of asset classes in Frazzini and Pedersen (2010).

ROBUST ESTIMATES OF CORRELATION

To derive a robust estimate of the correlation coefficient, we rewrite and re-interpret the expression for the correlation coefficient in a novel way, and then show how it can be estimated using two Theil-Sen regressions. Recall the definition of the correlation between two random variables X and Y :

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \times \sigma_Y} \quad (15)$$

where $Cov(X, Y)$ is the covariance between X and Y , and σ_X and σ_Y are the standard deviations of X and Y , respectively. This expression can be rewritten as

$$\begin{aligned} \rho_{X,Y} &= \sqrt{\frac{Cov(X, Y)^2}{\sigma_X^2 \times \sigma_Y^2}} \\ &= \sqrt{\frac{Cov(X, Y)}{\sigma_X^2} \times \frac{Cov(X, Y)}{\sigma_Y^2}} \\ &= \sqrt{\beta_{Y|X} \times \beta_{X|Y}} \end{aligned} \quad (16)$$

Table 3 Distribution of Theil-Sen Estimates of Correlation vs. Standard Maximum Likelihood Estimate: Normally Distributed Random Variables

Percentiles	1	5	10	25	50	75	90	95	99
Theil-Sen correlation	-0.24	-0.17	-0.14	-0.07	0	0.07	0.13	0.18	0.25
Maximum Likelihood correlation	-0.23	-0.17	-0.13	-0.07	0	0.07	0.13	0.16	0.23

Factored in this way, the correlation coefficient stands revealed as the geometric mean of two betas and is interpreted as follows. If a causal linear relationship runs from X to Y , (i.e., if X causes Y), the logical quantity to focus on is $\beta_{Y|X}$. Likewise, if a causal linear relationship runs from Y to X , (i.e., if Y causes X), the logical quantity to focus on is $\beta_{X|Y}$.

But when we don't know which way the causation flows, or even if the relationship is linear, we throw our hands up, take the geometric mean of these two betas, and call this quantity the correlation coefficient! For jointly normally distributed random variables, the correlation coefficient fully captures and encapsulates their covariation. For all other distributions, it serves merely as a useful shortcut that measures their covariation in a standardized way, as evidenced by the fact that its value is bounded between -1 and 1 .

The application of the Theil-Sen regression to the robust estimation of correlation is now obvious. Given two random vectors, X and Y , first regress X on Y , and then regress Y on X , using the Theil-Sen regression both times. The geometric mean of the two slopes is a robust estimate of the correlation coefficient, that is,

$$\rho_{X,Y}^{\text{Robust}} = \sqrt{\beta_{Y|X}^{\text{Theil-Sen}} \times \beta_{X|Y}^{\text{Theil-Sen}}} \quad (17)$$

When the random vectors are drawn from a normal distribution and are not corrupted by noise, we expect that this approach will work

just as well as the standard maximum likelihood estimator. In the presence of outliers, or if distribution of X and Y is highly skewed, it ought to do much better. And so it is, as the data in Tables 3 and 4 demonstrate.

Table 3 compares the performance of equation (16) to the standard maximum likelihood estimator when X and Y are drawn from a normal distribution, while Table 4 performs an identical comparison for Pareto(2) random variables. Both tables are created by extending the simulations used to illuminate the performance of the Theil-Sen regression to compute correlations as well.

The results follow the pattern established in Tables 1 and 2 for the slope coefficient. When X and Y are drawn from a normal distribution, the distribution of the Theil-Sen estimate of correlation is essentially identical to that of the maximum likelihood estimate; and when they are drawn from a Pareto(2) distribution, the Theil-Sen estimate of correlation is far more stable than the maximum likelihood estimate. Similar results are obtained when either X or Y (or both) are contaminated with noise (i.e., with outliers).

It is a short step from estimating individual correlations to estimating correlation matrices, and the repeated use of the Theil-Sen estimator across a set of random variables gives us a computationally inefficient but robust estimate of a correlation matrix $\hat{\rho}$, whose i,j th element is denoted by $\hat{\rho}_{ij}$, and whose diagonal elements

Table 4 Distribution of Theil-Sen Estimates of Correlation vs. Standard Maximum Likelihood Estimate: Pareto(2) Distributed Random Variables

Percentiles	1	5	10	25	50	75	90	95	99
Theil-Sen correlation	-0.11	-0.07	-0.05	-0.03	0	0.03	0.06	0.08	0.13
Maximum Likelihood correlation	-0.15	-0.11	-0.1	-0.06	-0.02	0.04	0.12	0.19	0.37

are all 1. Unfortunately, there is no guarantee that this correlation matrix will be nonnegative definite.

This, however, is no cause for alarm. It is relatively easy to transform this matrix into a nearby nonnegative definite correlation matrix ρ^* . Ideally, the transformation will minimally distort $\hat{\rho}$, and the many available solutions to this problem differ largely in the metric (or norm) that they use to measure the distance between $\hat{\rho}$ and ρ^* . In general, they solve the following optimization problem:

$$\text{Minimize } \|\rho^* - \hat{\rho}\|, \text{ s.t. } \rho^* \text{ is a nonnegative definite correlation matrix.} \tag{18}$$

Lindskog (2000), Rousseeuw and Molenberghs (1993), and Higham (2002) describe a number of different ways in which the nearest correlation matrix can be identified using both linear and nonlinear transformations of $\hat{\rho}$. The method that seems to work best in practice is the iterative method described by Higham (2002), which iteratively identifies the closest valid correlation matrix under a Frobenius norm (the sum of squared element by element differences) by factoring the correlation matrix in a particular way, forcing its negative eigenvalues to 0, then recombining its constituent pieces and forcing its diagonal elements to 1. The algorithm is described here for the sake of completeness and can be found in the NAG Fortran and C Libraries, as well as the NAG Toolbox for Matlab.

We first define two operators, $P_S(A)$ and $P_U(A)$ that can be applied to any symmetric matrix A . As A is symmetric, it admits a spectral decomposition $A = QDQ^T$, where Q is orthogonal, and $D = \text{diag}(\lambda_i)$ is a square matrix whose diagonal elements are the eigenvalues of A , and whose off-diagonal elements are 0. $P_S(A)$ and $P_U(A)$ are defined via

$$P_S(A) = QD^*Q^T, D_{ij}^* = \max(D_{ij}, 0), \text{ and} \tag{19}$$

$$P_U(A) = \text{Set } D_{ii} = 1, \text{ i.e. replace all diagonal elements of } D \text{ by } 1. \tag{20}$$

The algorithm proceeds as follows, with both X_k and Y_k converging linearly to ρ^* :

Algorithm H (Higham, 2002)

1. $\Delta S_0 = 0, X_0 = I, Y_0 = \hat{\rho}, k = 0$
2. While $\|Y_k - X_k\| > \varepsilon$, Do
 - a. $k = k + 1$
 - b. $R_k = Y_{k-1} - \Delta S_{k-1}$ (Dykstra's correction to speed convergence)
 - c. $X_k = P_S(R_k)$
 - d. $\Delta S_k = X_k - R_k$
 - e. $Y_k = P_U(X_k)$
3. $\rho^* = Y_k$

It is but a short step from estimating a robust nonnegative definite correlation matrix to estimating a similarly robust nonnegative *definite covariance matrix*. Given robust estimates of the volatility of each security, say σ_i^{Robust} , we can form a matrix whose diagonal elements are the robust volatilities of the assets, and whose off-diagonal elements are all 0, that is,

$$\Sigma = \begin{bmatrix} \sigma_1^{Robust} & 0 & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & 0 & \sigma_N^{Robust} \end{bmatrix} \tag{21}$$

and we can then define a robust nonnegative definite covariance matrix \hat{C} via:

$$\hat{C} = \Sigma \rho^* \Sigma \tag{22}$$

If the correlation matrix is nonnegative definite, the covariance matrix described in equation (22) is nonnegative definite as well. Rousseeuw and Croux (1993) describe a number of robust estimators of volatility, their preferred one being $Q_N(X)$, which is defined to be 2.222 times the 25th percentile of the set of distances $\{|x_i - x_j|, i < j\}$. They explore the properties of this estimator, which is similar in spirit to the Hodges-Lehmann (1963) estimate of the mean, show that its efficiency for the normal distribution is high (82%), and that it is robust to errors of arbitrary size in approximately half the points.

The *robust covariance matrix* defined by equation (22) can be used in a variety of applications such as mean-variance portfolio analysis and risk budgeting. It proves remarkably useful in practice, as it reduces and often completely eliminates the need for various constraints to ensure positive solutions that accord with a thoughtful portfolio manager's intuition.

KEY POINTS

- The Theil-Sen regression algorithm is an extraordinarily simple, intuitive, and robust algorithm for performing univariate regressions.
- The Theil-Sen estimator should be used routinely in place of OLS when performing univariate regressions, and in place of the standard maximum likelihood estimator when estimating correlations.
- The fact that the Theil-Sen estimator does not generalize naturally to multivariate regression should not be held against it—the vast majority of regressions that are carried out in practice are univariate, and a wide range of robust algorithms that work well with multivariate data are known.
- The Theil-Sen regression algorithm can be used to obtain robust estimates of beta, which can be further enhanced by the application of Dimson's correction for nonsynchronous trading and Vasicek's Bayesian adjustment.
- The robust estimates of correlation obtained from the Theil-Sen regression algorithm can be used as inputs to Higham's projection algorithm to estimate a nonnegative definite correlation matrix. This nonnegative definite correlation matrix can be combined with Rousseeuw and Croux's robust estimator of volatility to estimate a nonnegative definite covariance matrix. This nonnegative definite covariance matrix is of particular use in a wide range of mean-variance portfolio optimization and risk budgeting applications, including, but not limited to, the construction of minimum variance portfolios.

REFERENCES

- Dimson, E. (1979). Risk measurement when shares are subject to infrequent trading. *Journal of Financial Economics* 7, 2: 197–226.
- Frazzini, A., and Pedersen, L. (2010). Betting against beta. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1723048
- Hastie, T. J., and Tibshirani, R. J. (1990). *Generalized Additive Models*. New York: Chapman and Hall.
- Heitman, G., and Ord, K. J. (1985). An interpretation of the least squares regression surface. *The American Statistician* 39, 2: 120–123.
- Higham, N. (2002). Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis*, 22, 3: 329–343.
- Hodges, J. L., Jr., and Lehmann, E. L. (1963). Estimates of location based on rank tests. *Annals of Mathematical Statistics* 34, 2: 598–611.
- Linkdskog, F. (2000). Linear correlation estimation. <http://www.risklab.ch/ftp/papers/LinearCorrelationEstimation.pdf>
- Lintner, J. (1965). The valuation of risky assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* 47, 1: 13–37.
- Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica* 34, 4: 768–783.
- Oja, H., and Niinimaa, A. (1984). *On Robust Estimation of Regression Coefficients*. Research Report, Department of Applied Mathematics and Statistics, University of Oulu, Finland.
- Peng, H., Wang, S., and Wang, X. (2008). Consistency and asymptotic distribution of the Theil-Sen estimator. *Journal of Statistical Planning and Inference* 138, 6: 1836–1850.
- Rousseeuw, P. J., and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association* 88, 424: 1273–1283.
- Rousseeuw, P. J., and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: John Wiley and Sons.
- Rousseeuw, P. J., and Molenberghs, G. (1993). Transformation of nonpositive semidefinite correlation matrices. *Communications in Statistics—Theory and Methods* 22, 4: 965–984.
- Scholes, M., and Williams, J. (1977). Estimating betas from nonsynchronous data. *Journal of Financial Economics* 5, 3: 309–327.
- Sen, P. K. (1968). Estimate of the regression coefficient based on Kendall's Tau. *Journal of the American Statistical Association* 63, December: 1379–1389.

- Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19, 3: 425–442.
- Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes Mathematicae* 12: 85–91.
- Treynor, J. L. (1961). Towards a theory of market value of risky assets. Unpublished manuscript.
- Vasicek, O. A. (1973). A note on using cross-sectional information in Bayesian estimation on security betas. *Journal of Finance* 28, 5: 1233–1239.

Working with High-Frequency Data

IRENE ALDRIDGE

Managing Partner, Able Alpha Trading

Abstract: High-frequency trading (HFT) has exploded into the popular press as a major development affecting securities markets around the world. Unlike more established trading approaches that examine daily data and tactically rebalance portfolios every month or quarter, HFT parses trade-by-trade data at the highest speeds available. This typically implies that high-frequency traders monitor every tick of many securities concurrently and make their portfolio allocation decisions at lightning speeds with ultra-short investment horizons in mind. In fact, hedge fund managers consider strategies to be high frequency when their holding periods range from microseconds to several hours, without any positions held overnight. To process reams of data and make informed and rational decisions at such high speeds would be difficult even for the most accomplished traders. Thankfully, computer technology has evolved to become robust and inexpensive enough to aid any willing portfolio manager to take up the high-frequency craft.

This entry examines high-frequency data, the particularities and opportunities they bring, and compares these data with their low-frequency counterparts, wherever appropriate. *High-frequency trading* (HFT) strategies by their nature use a different population of data, and the traditional methods of data analysis need to be adjusted accordingly. Specifically, this entry examines the topics of volume, time-spacing, and bid-ask-bounce inherent in the high-frequency data.

WHAT ARE HIGH-FREQUENCY DATA?

High-frequency data, also known as “tick data,” are a record of live market activity. Every time a customer, a dealer, or another entity posts a so-called limit order to buy s units of a specific

security with ticker X at price q , a bid quote $q_{t_b}^b$ is logged at time t_b to buy $s_{t_b}^b$ units of X . (Market orders are incorporated into tick data in a different way as discussed below.) When the newly arrived bid quote $q_{t_b}^b$ has the highest price relative to all other previously arrived bid quotes in force, $q_{t_b}^b$ becomes known as “the best bid” available at time t_b . Similarly, when a trading entity posts a limit order to sell s units of X at price q , an ask quote $q_{t_a}^a$ is logged at time t_a to sell $s_{t_a}^a$ units of X . If the latest $q_{t_a}^a$ is lower than all other available ask quotes for security X , $q_{t_a}^a$ becomes known as “the best ask” at time t_a .

What happens to quotes from the moment they arrive largely depends on the venue where the orders are posted. Best bids and asks posted directly on an exchange will be broadcast to all exchange participants and other parties tracking quote data. In situations when the new best

bid exceeds the best ask already in force on the exchange, $q_{t_b}^b \geq q_{t_a}^a$, most exchanges will immediately “match” such quotes, executing a trade at the preexisting best ask, $q_{t_a}^a$ at time t_b . Conversely, should the newly arrived best ask fall below the current best bid, $q_{t_a}^a \leq q_{t_b}^b$, the trade is executed at the preexisting best bid, $q_{t_b}^b$ at time t_a .

Most *dark pools* match bids and asks “crossing the spread,” but may not broadcast the newly arrived quotes (hence the mysterious moniker, the “dark pools”). Similarly, quotes destined for the interdealer networks may or may not be disseminated to other market participants, depending on the venue.

Market orders contribute to high-frequency data in the form of “last trade” information. Unlike a limit order that is an order to buy a specified quantity of a security at a certain price, a market order is an order to buy a specified quantity of a security at the best price available at the moment the order is “posted” on the trading venue. As such, market orders are executed immediately at the best available bid or best ask prices, with each market buy order executed at the best ask and each market sell matched with the best bid, and the transaction is recorded in the quote data as the “last trade price” and the “last trade size.”

A large market order may need to be matched with one or several best quotes, generating several “last trade” data points. For example, if the newly arrived market buy order is smaller in size than that of the best ask, the best ask quote may still remain in force on most trading venues, but the best ask size will be reduced to reflect that the portion of the best ask quote has been matched with the market order. When the size of the incoming market buy order is bigger than the size of the corresponding best ask, the market order consumes the best ask in its entirety, and then proceeds to be matched sequentially with the next available best ask until the size of the market order is fulfilled. The remaining lowest-priced ask quote becomes the best ask available on the trading venue.

Most limit and market orders are placed in so-called “lot sizes”: increments of a certain number of units, known as a lot. In foreign exchange, a standard trading lot today is US\$5 million, a considerable reduction from a minimum of \$25 million entertained by high-profile brokers just a few years ago. On equity exchanges, a lot can be as low as one share, but dark pools may still enforce a 100 share minimum requirement for orders. An order for the amount other than an integer increment of a lot size is called an “odd lot.”

Small limit and market “odd lot” orders posted through a broker-dealer may be aggregated, or “packaged,” by the broker-dealer into larger-size orders in order to obtain volume discounts at the orders’ execution venue. In the process, the brokers may “sit” on quotes without transmitting them to an executing venue, delaying execution of customers’ orders.

HOW ARE HIGH-FREQUENCY DATA RECORDED?

The highest-frequency data are a collection of sequential “ticks,” arrivals of the latest quote, trade, price, order size, and volume information. Tick data usually have the following properties:

- A timestamp
- A financial security identification code
- An indicator of what information it carries
- Bid price
- Ask price
- Available bid size
- Available ask size
- Last trade price
- Last trade size
- Security-specific data, such as implied volatility for options
- The market value information, such as the actual numerical value of the price, available volume, or size

A timestamp records the date and time at which the quote originated. It may be the time

at which the exchange or the broker-dealer released the quote, or the time when the trading system has received the quote. At the time this entry is written, the standard “round-trip” travel time of an order quote from the ordering customer to the exchange and back to the customer with the acknowledgement of order receipt is 15 milliseconds or less in New York. Brokers have been known to be fired by their customers if they are unable to process orders at this now standard speed. Sophisticated quotation systems, therefore, include milliseconds and even microseconds as part of their timestamps.

Another part of the quote is an identifier of the financial security. In equities, the identification code can be a ticker, or, for tickers simultaneously traded on multiple exchanges, a ticker followed by the exchange symbol. For futures, the identification code can consist of the underlying security, futures expiration date, and exchange code.

The last trade price shows the price at which the last trade in the security cleared. Last trade price can differ from the bid and ask. The differences can arise when a customer posts a favorable limit order that is immediately matched by the broker without broadcasting the customer’s quote. Last trade size shows the actual size of the last executed trade.

The best bid is the highest price available for sale of the security in the market. The best ask is the lowest price entered for buying the security at any particular time. In addition to the best bid and best ask, quotation systems may disseminate “market depth” information: the bid and ask quotes entered posted on the trading venue at prices worse than the best bid and ask, as well as aggregate order sizes corresponding to each bid and ask recorded on the trading venue’s “books.” Market depth information is sometimes referred to as the Level II data and may be disseminated as the premium subscription service only. In contrast, the best bid, best ask, last trade price, and size information (“Level I data”) is often available for a small nominal fee.

Panels (a) and (b) of Figure 1 illustrate a 30-second log of Level I high-frequency data recorded by NYSE Arca for SPDR S&P 500 ETF (ticker SPY) from 14:00:16:400 to 14:02:00:000 GMT on November 9, 2009. Panel (a) shows quote data: best bid, best ask, and last trade information, while panel (b) displays corresponding position sizes (best bid size, best ask size, and last trade size).

PROPERTIES OF HIGH-FREQUENCY DATA

High-frequency securities data have been studied for many years. Yet, the concept is still something of a novelty to many academics and practitioners. Unlike daily or monthly data sets commonly used in much of financial research and related applications, high-frequency data have distinct properties, which simultaneously can be advantageous and intimidating to researchers. Table 1 summarizes the properties of high-frequency data. Each property, its advantages, and its disadvantages are discussed in detail later in the entry.

HIGH-FREQUENCY DATA ARE VOLUMINOUS

The nearly two-minute sample of tick data for SPDR S&P 500 ETF (ticker SPY) shown in Figure 1 contained over 100 observations of Level I data: best bid quotes and sizes, best ask quotes and sizes, and last trade prices and sizes. Table 2 summarizes the breakdown of the data points provided by NYSE Arca for SPY from 14:00:16:400 to 14:02:00:000 GMT on November 9, 2009, and SPY, Japanese yen futures, and a euro call option throughout the day on November 9, 2009. Other Level I data omitted from Table 2 include cumulative daily trade volume for SPY and Japanese yen futures, and “Greeks” for the euro call option. The number of quotes observed on November 9, 2009, for SPY alone would comprise over 160 years of daily open,

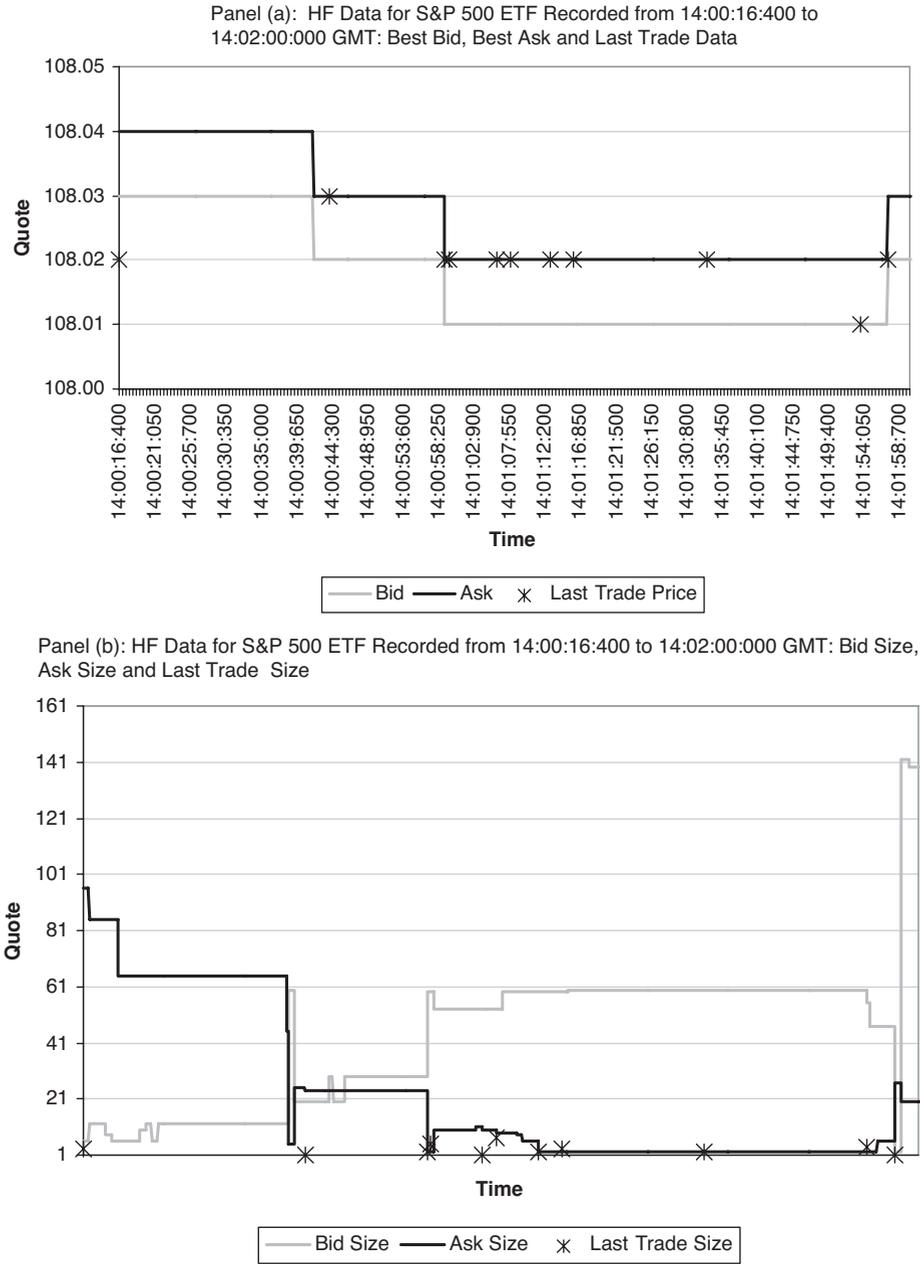


Figure 1 Level I High-Frequency Data Recorded by NYSE Arca for SPDR S&P 500 ETF (ticker SPY) from 14:00:16:400 to 14:02:00:000 GMT on November 9, 2009

high, low, close, and volume data points, assuming an average of 252 trading days per year.

The quality of data does not always match its quantity. Centralized exchanges generally provide accurate data on bids, asks, and volume of any. The information on the limit order book is less commonly available. In decentralized mar-

kets, such as foreign exchange and the inter-bank money market, no market-wide quotes are available at any given time. In such markets, participants are aware of the current price levels, but each institution quotes its own prices adjusted for its order book. In decentralized markets, each dealer provides his or her own

Table 1 Summary of Properties of High-Frequency Data

Property of HF Data	Description	Pros	Cons
Voluminous	Each day of high-frequency data contains the number of observations equivalent to 30 years of daily data.	Large numbers of observations carry lots of information.	High-frequency data are difficult to handle manually.
Irregularly spaced in time	Arrival of tick data is asynchronous.	Durations between data arrivals carry information.	Most traditional models require regularly spaced data; need to convert high-frequency data to some regular intervals, or "bars" of data. Converted data is often sparse (populated with zero returns), once again making traditional econometric inferences difficult.
Subject to bid-ask bounce	Unlike traditional data based on just closing prices, tick data carries additional supply and demand information in the form of bid and ask prices and offering sizes.	Bid and ask quotes can carry valuable information about impending market moves and can be harnessed to researcher's advantage.	Bid and ask quotes are separated by a spread. Continuous movement from bid to ask and back introduces a jump process, difficult to deal with through many conventional models.

tick data to clients. As a result, a specific quote on a given financial instrument at any given time may vary from dealer to dealer. Reuters, Telerate, and Knight Ridder, among others, collect quotes from different dealers and disseminate them back, improving the efficiency of the decentralized markets.

There are generally thought to be three anomalies in interdealer quote discrepancies. First, each dealer's quotes reflect that dealer's own inventory. For example, a dealer that has just sold a customer \$100 million of USD/CAD would be eager to diversify the risk of the posi-

tion and avoid selling any more of USD/CAD. Most dealers are, however, obligated to transact with their clients on tradable quotes. To incite clients to place sell orders on USD/CAD, the dealer temporarily raises the bid quote on USD/CAD. At the same time, to encourage clients to withhold placing buy orders, the dealer raises the ask quote on USD/CAD. Thus, dealers tend to raise both bid and ask prices whenever they are short in a particular financial instrument and lower both bid and ask prices whenever they are disproportionally long in a financial instrument.

Table 2 Summary Statistics for Level I Quotes for Selected Securities on November 9, 2009

Quote Type	SPY, 14:00:16:400 to 14:02:00:000 GMT	SPY, all day	USD/JPY Dec 2009 Futures, all day	EUR/USD Call Expiring Dec 2009 with Strike Price of 1.5100, all day
Best Bid Quote	4 (3%)	5,467 (3%)	6,320 (5%)	1,521 (3%)
Best Bid Size	36 (29%)	38,948 (19%)	39,070 (32%)	5,722 (11%)
Best Ask Quote	4 (3%)	4,998 (2%)	6,344 (5%)	1,515 (3%)
Best Ask Size	35 (28%)	38,721 (19%)	38,855 (32%)	5,615 (11%)
Last Trade Price	6 (5%)	9,803 (5%)	3,353 (3%)	14 (0%)
Last Trade Size	20 (16%)	27,750 (14%)	10,178 (8%)	25 (0%)
Total	125	203,792	123,216	49,982

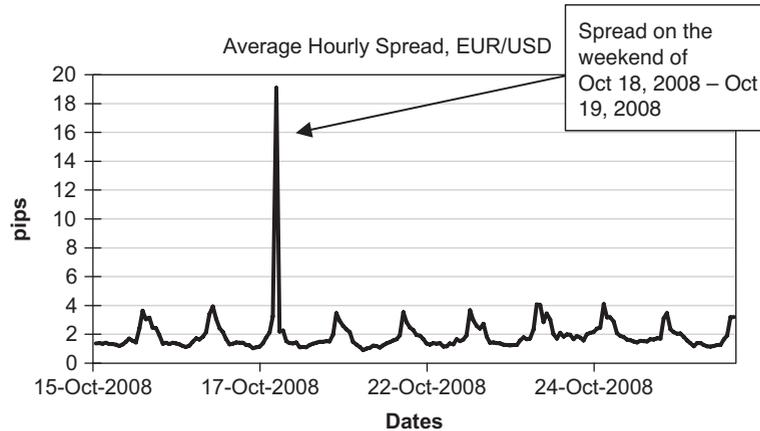


Figure 2 Average Hourly Bid-Ask Spread on EUR/USD Spot for the Last Two Weeks of October 2008 on a Median Transaction Size of USD 5 million
Source: Aldridge (2009).

Second, in an anonymous marketplace, such as a dark pool, dealers as well as other market makers may “fish” for market information by sending indicative quotes that are much off the previously quoted price to assess the available demand or supply.

Third, Dacorogna et al. (2001) note that some dealers’ quotes may lag real market prices. The lag is thought to vary from milliseconds to a minute. Some dealers quote moving averages of quotes of other dealers. The dealers who provide delayed quotes usually do so to advertise their market presence in the data feed. This was particularly true when most order prices were negotiated over the telephone, allowing a considerable delay between quotes and orders. Fast-paced electronic markets discourage lagged quotes, improving the quality of markets.

HIGH-FREQUENCY DATA ARE SUBJECT TO BID-ASK BOUNCE

In addition to trade price and volume data long available in low-frequency formats, high-frequency data comprise bid and ask quotes and the associated order sizes. Bid and ask data

arrive asynchronously and introduce noise in the quote process.

The difference between the bid quote and the ask quote at any given time is known as the bid-ask spread. The bid-ask spread is the cost of instantaneously buying and selling the security. The higher the bid-ask spread, the higher a gain the security must produce in order to cover the spread along with other transaction costs. Most low-frequency price changes are large enough to make the bid-ask spread negligible in comparison. In tick data, on the other hand, incremental price changes can be comparable or smaller than the bid-ask spread.

Bid-ask spreads usually vary throughout the day. Figure 2 illustrates the average bid-ask spread cycles observed in the institutional EUR/USD market for the last two weeks of October 2008. As Figure 2 shows, the average spread increases significantly during Tokyo trading hours when the market is quiet. The spread then reaches its lowest levels during the overlap of the London and New York trading sessions when the market has many active buyers and sellers. The spike in the spread over the weekend of October 18–19, 2008, reflects the market concern over the subpoenas issued on October 17, 2009, to senior Lehman executives

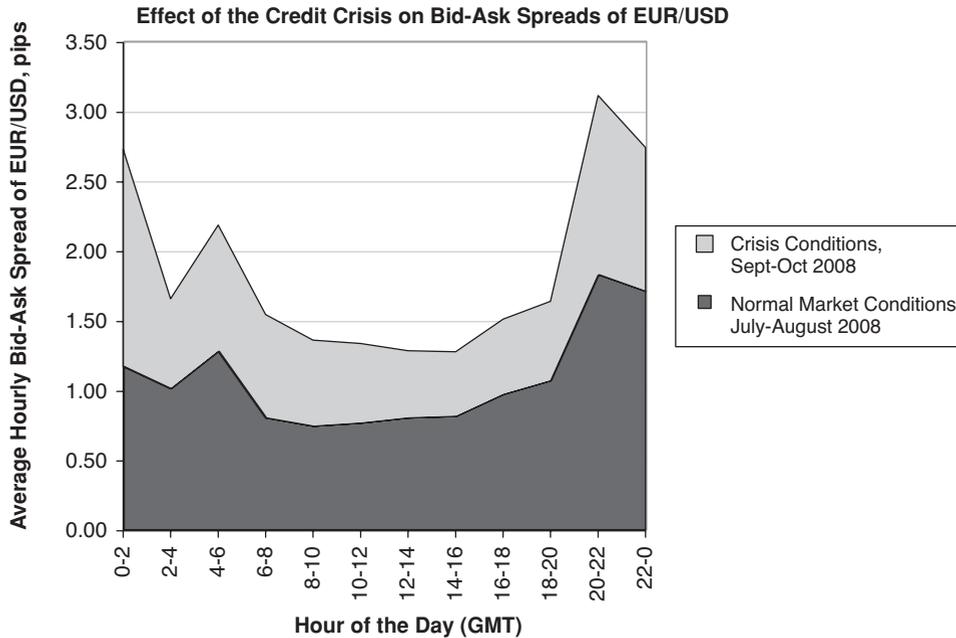


Figure 3 Comparison of Average Bid-Ask Spreads for Different Hours of the Day during Normal Market Conditions and Crisis Conditions

in a case relating to potential securities fraud at Lehman Brothers.

Bid-ask spreads typically increase during periods of market uncertainty or instability. Figure 3, for example, compares average bid-ask spreads on EUR/USD in the stable market conditions of July–August 2008 and the crisis conditions of September–October 2008. As the figure shows, the intraday spread pattern is persistent in both crisis and normal market conditions, but the spreads are significantly higher during crisis months than during normal conditions at all hours of the day. As Figure 3 also shows, the spread increase is not uniform at all hours of the day. The average hourly EUR/USD spreads increased by 0.0048% (0.48 basis points or pips) between the hours of 12 GMT and 16 GMT, when the London and New York trading sessions overlap. From 0 to 2 GMT, during the Tokyo trading hours, the spread increased by 0.0156%, over three times the average increase during the New York/London hours.

As a result of increasing bid-ask spreads during periods of uncertainty and crises, the

profitability of high-frequency strategies decreases during those times. For example, high-frequency EUR/USD strategies running over Asian hours incurred significantly higher costs during September and October 2008 as compared with normal market conditions. A strategy that executed 100 trades during Asian hours alone resulted in 1.56 percent evaporating from daily profits due to the increased spreads, while the same strategy running during London and New York hours resulted in a smaller but still significant daily profit decrease of 0.48%. The situation can be even more severe for high-frequency strategies built for less liquid instruments. For example, bid-ask spreads for NZD/USD (not shown) on average increased thrice during September–October in comparison with market conditions of July–August 2008.

While tick data carries information about market dynamics, it is also distorted by the same processes that make the data so valuable in the first place. Dacorogna et al. (2001) report that sequential trade price bounces between the

bid and ask quotes during market execution of orders introduce significant distortions into estimation of high-frequency parameters. Corsi, Zumbach, Muller, and Dacorogna (2001), for example, show that the bid-ask bounce introduces a considerable bias into volatility estimates. The authors calculate that the bid-ask bounce on average results in -40% negative first-order autocorrelation of tick data. Corsi et al. (2001) as well as Voev and Lunde (2007) propose to remedy the bias by filtering the data from the bid-ask bounce prior to estimation.

To use standard econometric techniques in the presence of the bid-ask bounce, many practitioners convert the tick data to "mid-quote" format: the simple average of the latest bid and ask quotes. The mid-quote is used to approximate the price level at which the market is theoretically willing to trade if buyers and sellers agreed to meet each other halfway on the price spectrum. Mathematically, the mid-quote can be expressed as follows:

$$\hat{q}_{t_m}^m = \frac{1}{2} (q_{t_a}^a + q_{t_b}^b) \text{ where } t_m = \begin{cases} t_a, & \text{if } t_a \geq t_b \\ t_b, & \text{otherwise} \end{cases} \quad (1)$$

The latter condition for t_m reflects the continuous updating of the mid-quote estimate: $\hat{q}_{t_m}^m$ is updated whenever the latest best bid, $q_{t_b}^b$, or best ask quote, $q_{t_a}^a$, arrives, at t_b or t_a respectively.

Another way to sample tick quotes into a cohesive data series is by weighing the latest best bid and best ask quotes by their accompanying order sizes:

$$\tilde{q}_t^s = \frac{q_{t_b}^b s_{t_a}^a + q_{t_a}^a s_{t_b}^b}{s_{t_a}^a + s_{t_b}^b} \quad (2)$$

where $q_{t_b}^b$ and $s_{t_b}^b$ is the best bid quote and the best bid available size recorded at time t_b (when $q_{t_b}^b$ became the best bid), and $q_{t_a}^a$ and $s_{t_a}^a$ is the best ask quote and the best ask available size recorded at time t_a .

Figure 5 compares the histograms of simple returns computed from mid-quote (panel a), size-weighted mid-quote (panel b), and trade-price (panel c) processes for SPDR S&P 500 ETF data recorded as they arrive throughout

November 9, 2009. The data neglect the time difference between the adjacent quotes, treating each sequential quote as an independent observation. Figure 6 contrasts the quantile distribution plots of the same data sets with the quantiles of a standard normal distribution.

As Figures 4 and 5 show, the basic mid-quote distribution is constrained by the minimum "step size": The minimum changes in the mid-quote can occur at half-tick increments (at present, the minimum tick size is \$0.01 in equities). The size-weighted mid-quote forms the most continuous distribution among the three distributions discussed. Figure 6 confirms this notion further and also illustrates the fat tails present in all three types of data distributions.

In addition to real-time adjustments to bid-ask data, researchers deploy forecasting techniques to estimate the impending bid-ask spread and adjust for it in models ahead of time. Future realizations of the bid-ask spread can be estimated using the model suggested by Roll (1984), where the price of an asset at time t , p_t , is assumed to equal an unobservable fundamental value, m_t , offset by a value equal to half of the bid-ask spread, s . The price offset is positive when the next market order is a buy, and negative when the trade is a sell, as shown in equation (3):

$$p_t = m_t + \frac{s}{2} I_t \quad (3)$$

$$\text{where } I_t = \begin{cases} 1, & \text{market buy at ask} \\ -1, & \text{market sell at bid} \end{cases}$$

If either a buy or a sell order can arrive next with equal probability, then $E[I_t] = 0$, and $E[\Delta p_t] = 0$, absent changes in the fundamental asset value, m_t . The covariance of subsequent price changes, however, is different from 0:

$$\text{cov}[\Delta p_t, \Delta p_{t+1}] = E[\Delta p_t \Delta p_{t+1}] = -\frac{s^2}{4} \quad (4)$$

As a result, the future expected spread can be estimated as follows:

$$E[s] = 2\sqrt{-\text{cov}[\Delta p_t, \Delta p_{t+1}]} \text{ whenever } \text{cov}[\Delta p_t, \Delta p_{t+1}] < 0$$

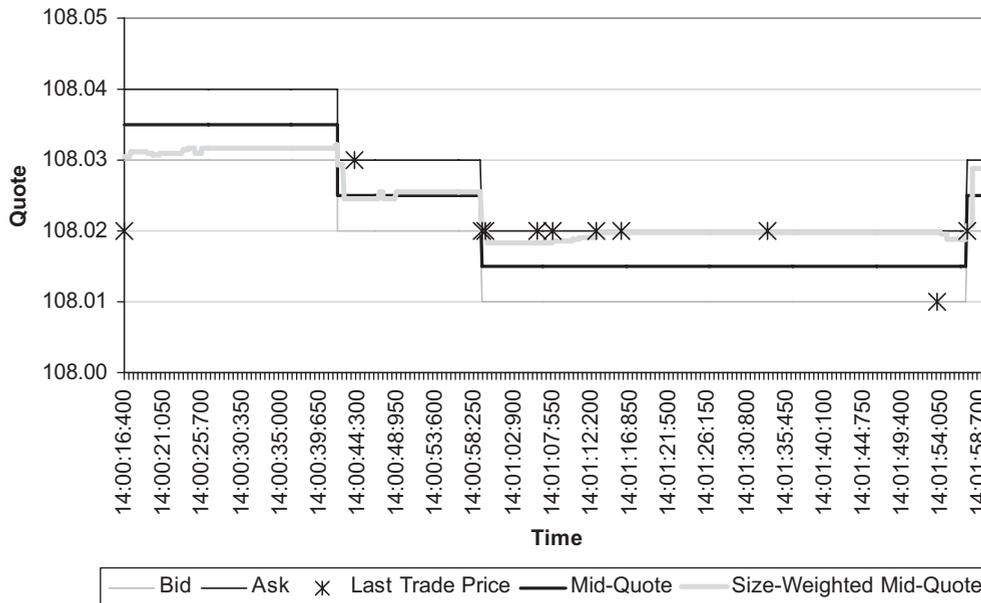


Figure 4 Bid-Ask Aggregation Techniques on Data for SPDR S&P 500 ETF (ticker SPY) Recorded by NYSE Arca on November 9, 2009, from 14:00:16:400 to 14:00:02:000 GMT

Numerous extensions of Roll's model have been developed to account for contemporary market conditions along with numerous other variables. Hasbrouck (2007) provides a good summary of the models.

HIGH-FREQUENCY DATA ARE IRREGULARLY SPACED IN TIME

Most modern computational techniques have been developed to work with regularly spaced data, presented in monthly, weekly, daily, hourly, or other consistent intervals. The traditional reliance of researchers on fixed time intervals is due to:

- Relative availability of daily data (newspapers have published daily quotes since the 1920s).
- Relative ease of processing regularly spaced data.
- An outdated view that “whatever drove security prices and returns, it probably did not vary significantly over short time intervals.” (Goodhart and O’Hara, 1997, pp. 80–81)

In contrast, high-frequency observations are separated by varying time intervals. One way to overcome the irregularities in the data is to sample it at certain predetermined periods of time—for example, every hour or minute. For example, if the data are to be converted from tick data to minute “bars,” then under the traditional approach, the bid or ask price for any given minute would be determined as the last quote that arrived during that particular minute. If no quotes arrived during a certain minute, then the previous minute’s closing prices would be taken as the current minute’s closing prices, and so on. Figure 7, panel (a) illustrates this idea. This approach implicitly assumes that in the absence of new quotes, the prices stay constant, which does not have to be the case.

Dacorogna et al. (2001) propose a potentially more precise way to sample quotes—linear time-weighted interpolation between adjacent quotes. At the core of the interpolation technique is an assumption that at any given time, unobserved quotes lie on a straight line that connects two neighboring observed quotes.

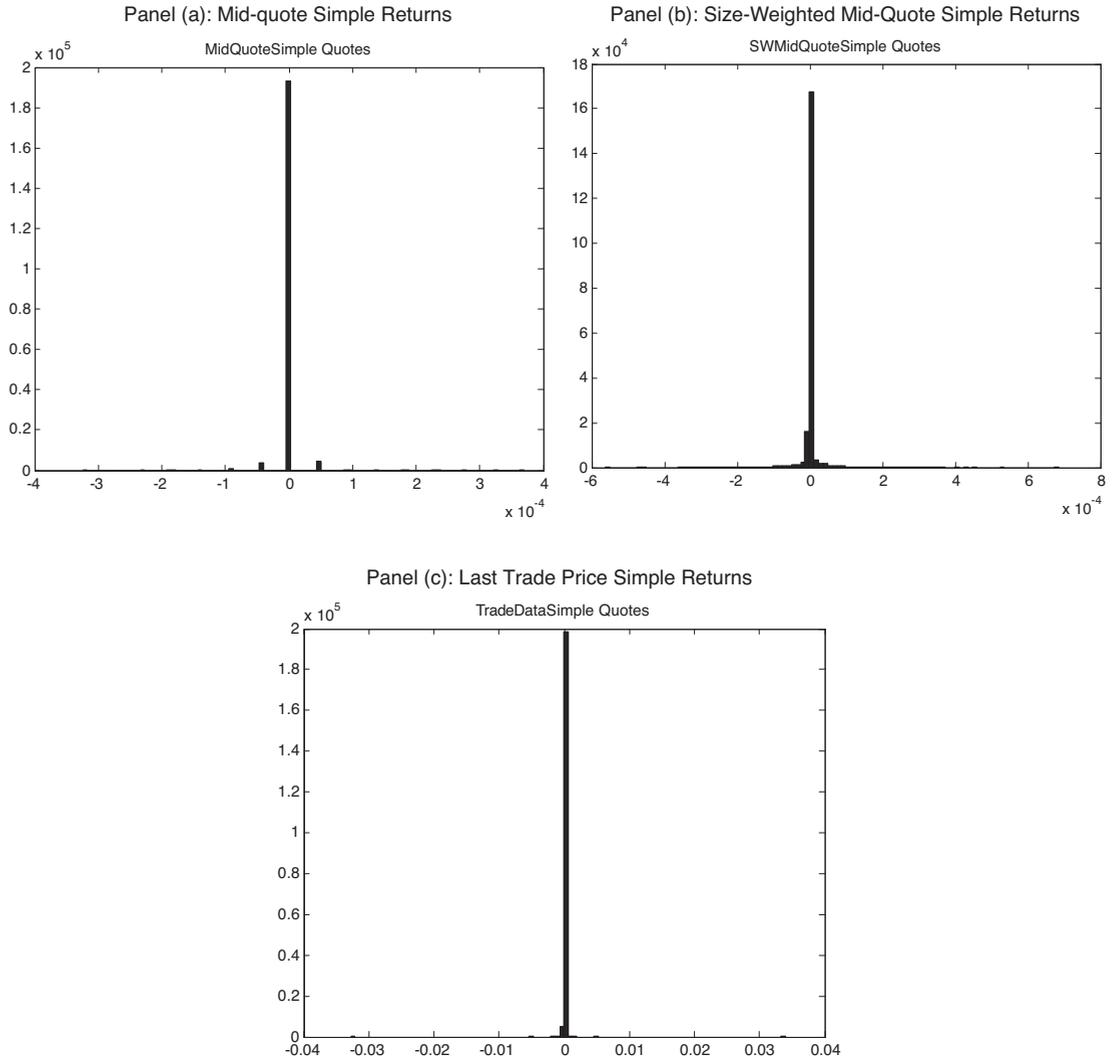


Figure 5 Histograms of Simple Returns Computed from Mid-Quote (panel a), Size-Weighted Mid-Quote (panel b), and Trade-Price (panel c) Processes for SPDR S&P 500 ETF Data Recorded as They Arrive Throughout November 9, 2009

Figure 7, panel (b) illustrates linear interpolation sampling.

As shown in Figure 7, panels (a) and (b), the two quote-sampling methods produce quite different results.

Mathematically, the two sampling methods can be expressed as follows:

Quote sampling using closing prices: $\hat{q}_t = q_{t,last}$ (5)

Quote sampling using linear interpolation:

$$\hat{q}_t = q_{t,last} + (q_{t,next} - q_{t,last}) \frac{t - t_{last}}{t_{next} - t_{last}} \quad (6)$$

where \hat{q}_t is the resulting sampled quote, t is the desired sampling time (start of a new minute, for example), t_{last} is the timestamp of the last observed quote prior to the sampling time t , $q_{t,last}$ is the value of the last quote prior to the sampling time t , t_{next} is the timestamp of the first observed quote after the sampling time t ,

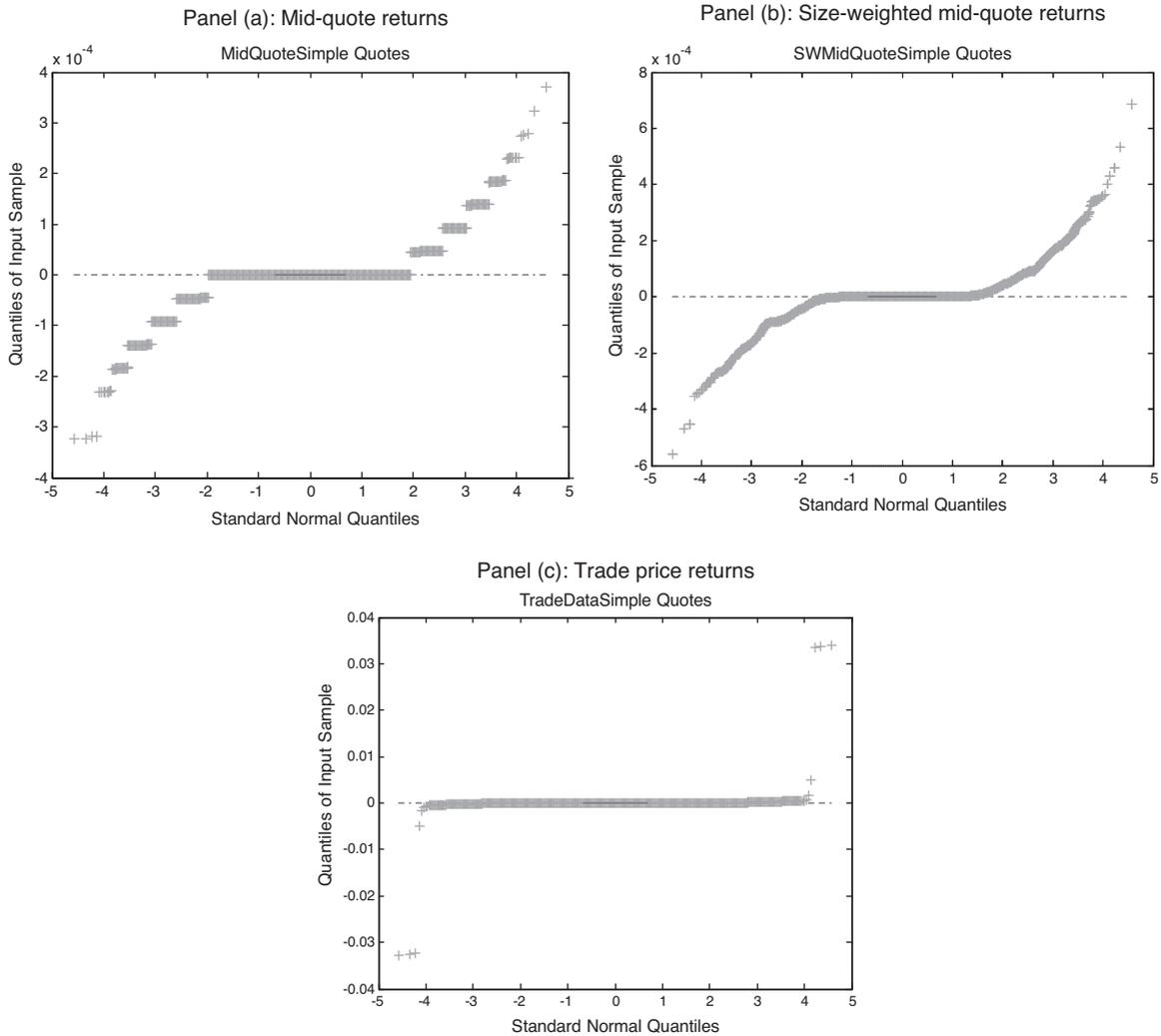


Figure 6 Quantile Plots of Simple Returns of Mid-Quote (panel a), Size-Weighted Mid-Quote (panel b), and Trade-Price (panel c) Processes for SPDR S&P 500 ETF Data Recorded as They Arrive Throughout November 9, 2009

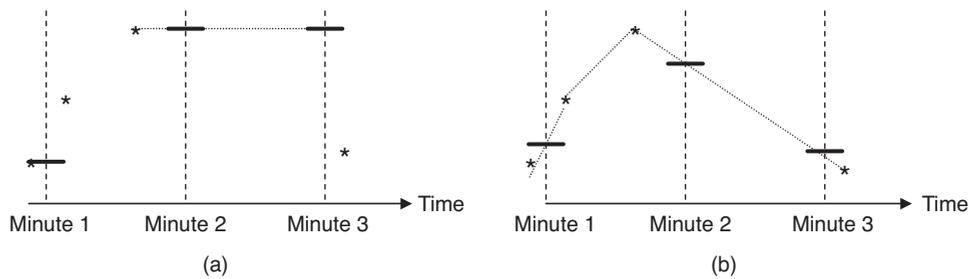


Figure 7 Data-Sampling Methodologies

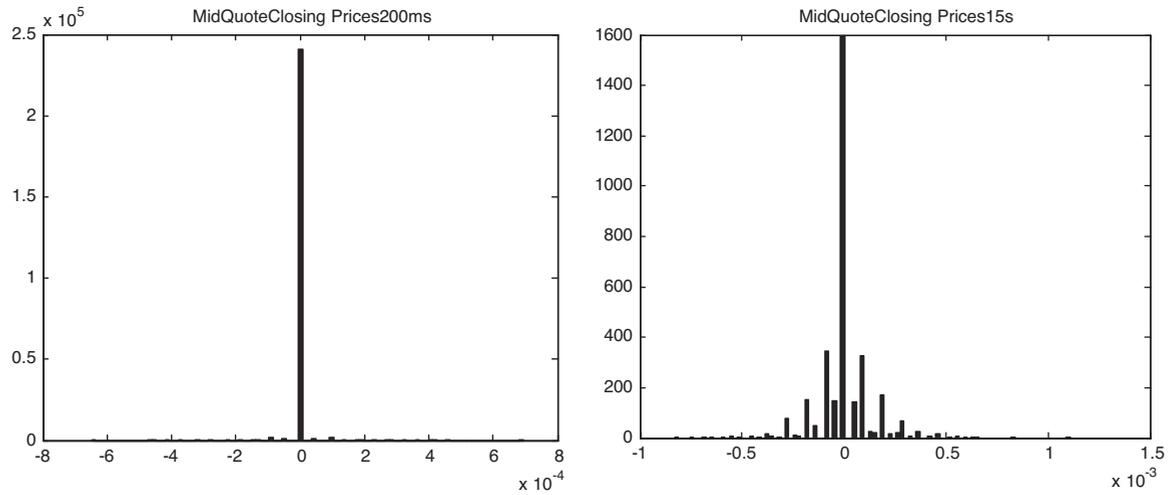


Figure 8 Mid-Quote “Closing Quotes” Sampled at 200 ms (left) and 15s Intervals

and $q_{t,next}$ is the value of the first quote after the sampling time t .

Figures 8 and 9 compare histograms of the mid-quote data sampled as closing prices and interpolated at frequencies of 200 ms and 15s. Figure 10 compares quantile plots of closing prices and interpolated distributions. As Figures 8 and 9 show, often-sampled distributions are sparse, that is, contain more 0 returns than distributions sampled at lower frequencies. At the same time, returns computed from interpo-

lated quotes are more continuous than closing prices, as Figure 10 illustrates.

Instead of manipulating the interquote intervals into the convenient regularly spaced formats, several researchers have studied whether the time distance between subsequent quote arrivals itself carries information. For example, most researchers agree that intertrade intervals indeed carry information on securities for which short sales are disallowed; the lower the *intertrade duration*, the more likely the

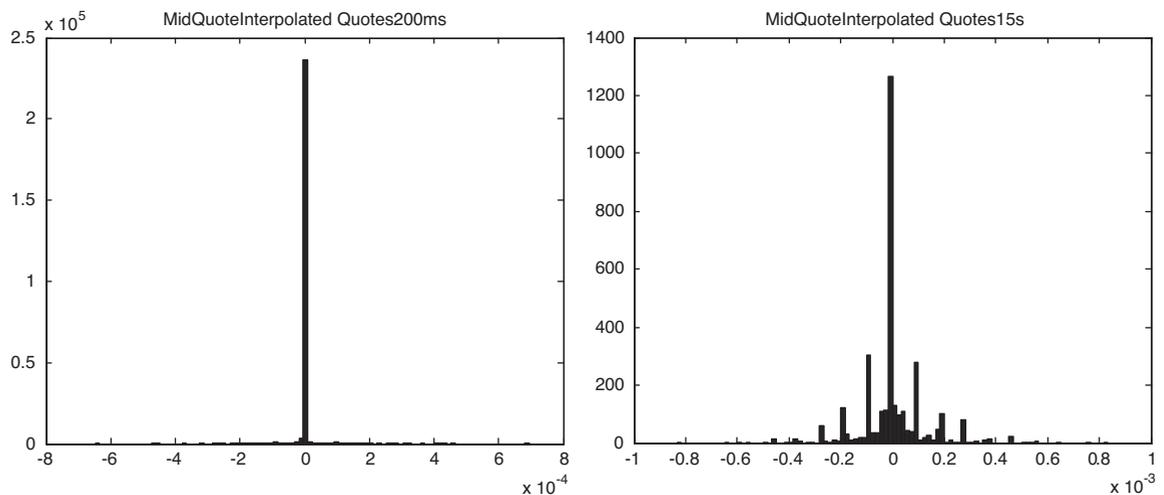


Figure 9 Mid-Quote “Time-Interpolated Quotes” Sampled at 200 ms (left) and 15s Intervals

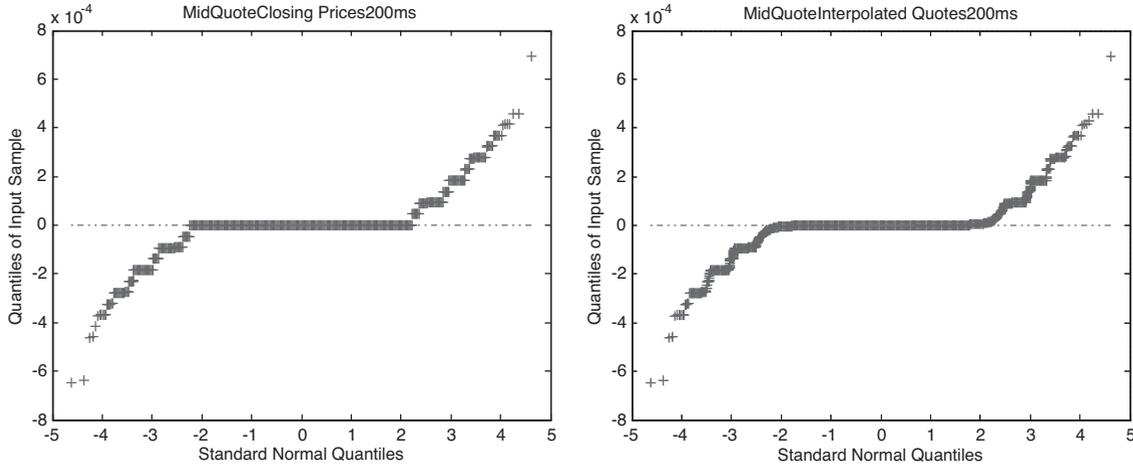


Figure 10 Quantile Plots: Closing Prices vs. Interpolated Mid-Quotes Sampled at 200 ms

yet-to-be-observed good news and the higher the impending price change.

Duration models are used to estimate the factors affecting the time between any two sequential ticks. Such models are known as quote processes and trade processes, respectively. Duration models are also used to measure the time elapsed between price changes of a prespecified size, as well as the time interval between predetermined trade volume increments. The models working with fixed price are known as price processes; the models estimating variation in duration of fixed volume increments are known as volume processes.

Durations are often modeled using *Poisson processes* that assume that sequential events, like quote arrivals, occur independently of one another. The number of arrivals between any two time points t and $(t + \tau)$ is assumed to have a Poisson distribution. In a Poisson process, λ arrivals occur per unit time. In other words, the arrivals occur at an average rate of $(1/\lambda)$. The average arrival rate may be assumed to hold constant, or it may vary with time. If the average arrival rate is constant, the probability of observing exactly k arrivals between times t and $(t + \tau)$ is

$$P[(N(t + \tau) - N(t)) = k] = \frac{1}{k!} e^{-\lambda\tau} (\lambda\tau)^k, \quad k = 0, 1, 2, \dots \quad (7)$$

Diamond and Verrecchia (1987) and Easley and O'Hara (1992) were the first to suggest that the duration between subsequent ticks carries information. Their models posit that in the presence of short-sale constraints, intertrade duration can indicate the presence of good news; in markets of securities where short selling is disallowed, the shorter the intertrade duration, the higher is the likelihood of unobserved good news. The reverse also holds: In markets with limited short selling and normal liquidity levels, the longer the duration between subsequent trade arrivals, the higher the probability of yet-unobserved bad news. A complete absence of trades, however, indicates a lack of news.

Easley and O'Hara (1992) further point out that trades that are separated by a time interval have a much different information content than trades occurring in close proximity. One of the implications of Easley and O'Hara (1992) is that the entire price sequence conveys information and should be used in its entirety whenever possible, strengthening the argument for high-frequency trading.

Table 3 shows summary statistics for a duration measure computed on all trades recorded for S&P 500 Depository Receipts ETF (SPY) on May 13, 2009. As Table 3 illustrates, the average intertrade duration was the longest outside of

Table 3 Hourly Distributions of Intertrade Duration Observed on May 13, 2009 for S&P 500 Depository Receipts ETF (SPY)

Hour (ET)	Intertrade Duration (milliseconds)					
	No. of Trades	Average	Median	Std Dev	Skewness	Kurtosis
4–5 AM	170	19074.58	5998	47985.39	8.430986	91.11571
5–6 AM	306	11556.95	4781.5	18567.83	3.687372	21.92054
6–7 AM	288	12606.81	4251	20524.15	3.208992	16.64422
7–8 AM	514	7096.512	2995	11706.72	4.288352	29.86546
8–9 AM	767	4690.699	1997	7110.478	3.775796	23.56566
9–10 AM	1089	2113.328	1934	24702.9	3.5185	24.6587
10–11 AM	1421	2531.204	1373	3409.889	3.959082	28.53834
11–12 PM	1145	3148.547	1526	4323.262	3.240606	17.24866
12–1 PM	749	4798.666	1882	7272.774	2.961139	13.63373
1–2 PM	982	3668.247	1739.5	5032.795	2.879833	13.82796
2–3 PM	1056	3408.969	1556	4867.061	3.691909	23.90667
3–4 PM	1721	2094.206	1004	2684.231	2.9568	15.03321
4–5 PM	423	8473.593	1500	24718.41	7.264483	69.82157
5–6 PM	47	73579.23	30763	113747.8	2.281743	7.870699
6–7 PM	3	1077663	19241	1849464	0.707025	1.5

regular market hours, and the shortest during the hour preceding the market close (3–4 P.M. ET).

The variation in duration between subsequent trades may be due to several other causes. While the lack of trading may be due to a lack of new information, trading inactivity may also be due to low levels of liquidity, trading halts on exchanges, and strategic motivations of traders. Foucault, Kadan, and Kandel (2005) consider that patiently providing liquidity using limit orders may itself be a profitable trading strategy, as liquidity providers should be compensated for their waiting. The compensation usually comes in the form of a bid-ask spread and is a function of the waiting time until the order limit is “hit” by liquidity takers; lower intertrade durations induce lower spreads. However, Dufour and Engle (2000) and Saar and Hasbrouck (2002) find that spreads are actually higher when traders observe short durations, contrasting the time-based limit order compensation hypothesis.

In addition to durations between subsequent trades and quotes, researchers have also been modeling durations between fixed changes in security prices and volumes. The time interval between subsequent price changes of a spec-

ified magnitude is known as price duration. Price duration has been shown to decrease with increases in volatility. Similarly, the time interval between subsequent volume changes of a prespecified size is known as the volume duration. Volume duration has been shown to decrease with increases in liquidity.

The information content of quote, trade, price, and volume durations introduces biases into the estimation process, however. If the available information determines the time between subsequent trades, time itself ceases to be an independent variable, introducing substantial endogeneity bias into estimation. As a result, traditional estimates of variance of transaction prices are too high in comparison with the true variance of the price series.

KEY POINTS

- High-frequency data are different from daily or lower frequency data. Whereas low frequency data typically comprise regularly spaced open, high, low, close, and volume information for a given financial security recorded during a specific period of time, high-frequency data include bid and ask quotes, sizes, and latest trade characteristics

that are recorded sequentially at irregular time intervals.

- The differences affect trading strategy modeling, introducing new opportunities and pitfalls for researchers.
- Numerous data points allow researchers to deduce statistically significant inferences on even short samples of high-frequency data.
- Different sampling approaches have been developed to convert high-frequency data into a more regular format better familiar to researchers, yet diverse sampling methodologies result in datasets with drastically dissimilar statistical properties.
- Some properties of high-frequency data, like intertrade duration, carry important market information unavailable at lower frequencies.

REFERENCES

- Aldridge, I. (2009). *High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems*. Hoboken, NJ: John Wiley & Sons.
- Corsi, F., Zumbach, G., Müller, U., and Dacorogna, M. (2001). Consistent high-precision volatility from high-frequency data. *Economics Notes* 30: 183–204.
- Dacorogna, M. M., Gencay, R., Muller, U. A., Olsen, R., and Pictet, O. V. (2001). *An Introduction to High-Frequency Finance*. San Diego, CA: Academic Press.
- Diamond, D. W., and Verrecchia, R. E. (1987). Constraints on short-selling and asset price adjustment to private information. *Journal of Financial Economics* 18: 277–311.
- Dufour, A., and Engle, R. F. (2000). Time and the price impact of a trade. *Journal of Finance* 55: 2467–2498.
- Easley, D., and O'Hara, M. (1992). Time and the process of security price adjustment. *Journal of Finance* 47, 2: 557–605.
- Foucault, T., Kadan, O., and Kandel, E. (2005). Limit order book as a market for liquidity. *Review of Financial Studies* 18: 1171–1217.
- Goodhart, C., and O'Hara, M. (1997). High frequency data in financial markets: Issues and applications. *Journal of Empirical Finance* 4: 73–114.
- Hasbrouck, J. (2007). *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*. New York: Oxford University Press.
- Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance* 39, 4: 1127–1240.
- Saar, G., and Hasbrouck, J. (2002). Limit orders and volatility in a hybrid market: The Island ECN. Working paper, New York University.
- Voev, V., and Lunde, A. (2007). Integrated covariance estimation using high-frequency data in the presence of noise. *Journal of Financial Econometrics* 5, 1: 68–104.

Financial Modeling Principles

Milestones in Financial Modeling

SERGIO M. FOCARDI, PhD

Partner, The Intertek Group

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: The origins of financial modeling can be traced back to the development of mathematical equilibrium at the end of the nineteenth century, followed in the beginning of the twentieth century with the introduction of sophisticated mathematical tools for dealing with the uncertainty of prices and returns. In the 1950s and 1960s, financial modelers had tools for dealing with probabilistic models for describing markets, the principles of contingent claims analysis, an optimization framework for portfolio selection based on mean and variance of asset returns, and an equilibrium model for pricing capital assets. The 1970s ushered in models for pricing contingent claims and a new model for pricing capital assets based on arbitrage pricing. Consequently, by the end of the 1970s, the frameworks for financial modeling were well known. It was the advancement of computing power and refinements of the theories to take into account real-world markets starting in the 1980s that facilitated implementation and broader acceptance of mathematical modeling of financial decisions.

The mathematical development of present-day economic and finance theory began in Lausanne, Switzerland at the end of the nineteenth century, with the development of the mathematical equilibrium theory by Leon Walras (1874) and Vilfredo Pareto (1906). Shortly thereafter, at the beginning of the twentieth century, Louis Bachelier (1900) in Paris and Filip Lundberg (1903) in Uppsala (Sweden) made two seminal contributions: They developed sophisticated mathematical tools to describe uncertain price and risk processes. These developments were well in advance of their time. Further progress was to be made only much later in the twentieth century, thanks to the de-

velopment of digital computers. By making it possible to compute approximate solutions to complex problems, digital computers enabled the large-scale application of mathematics to business problems.

A first round of innovation occurred in the 1950s and 1960s. Kenneth Arrow and Georges Debreu (1954) introduced a probabilistic model of markets and the notion of contingent claims. Harry Markowitz (1952) described mathematically the principles of the investment process in terms of utility optimization. In 1961, Franco Modigliani and Merton Miller (1961) clarified the nature of economic value, working out the implications of absence of arbitrage. Between

1964 and 1966, William Sharpe (1964), John Lintner (1965), and Jan Mossin (1966) developed a theoretical model of market prices based on the principles of financial decision making laid down by Markowitz. The notion of efficient markets was introduced by Paul Samuelson (1965), and five years later, further developed by Eugene Fama (1970).

The second round of innovation started at the end of the 1970s. In 1973, Fischer Black, Myron Scholes (1973), and Robert Merton (1973a) discovered how to determine option prices using continuous hedging. Three years later, Stephen Ross (1976) introduced arbitrage pricing theory (APT). Both were major developments that were to result in a comprehensive mathematical methodology for investment management and the valuation of derivative financial products. At about the same time, Merton introduced a continuous-time intertemporal, dynamic optimization model of asset allocation. Major refinements in the methodology of mathematical optimization and new econometric tools were to change the way investments are managed.

More recently, the diffusion of electronic transactions has made available a huge amount of empirical data. The availability of this data created the hope that economics could be given a more solid scientific grounding. A new field—econophysics—opened with the expectation that the proven methods of the physical sciences and the newly born science of complex systems could be applied with benefit to economics. It was hypothesized that economic systems could be studied as physical systems with only minimal a priori economic assumptions. Classical econometrics is based on a similar approach; but while the scope of classical econometrics is limited to dynamic models of time series, econophysics uses all the tools of statistical physics and complex systems analysis, including the theory of interacting multi-agent systems.

In this entry, we will describe the milestones in financial modeling.

THE PRECURSORS: PARETO, WALRAS, AND THE LAUSANNE SCHOOL

The idea of formulating quantitative laws of economic behavior in ways similar to the physical sciences started in earnest at the end of the 19th century. Though quite accurate economic accounting on a large scale dates back to Assyro-Babylonian times, a scientific approach to economics is a recent endeavor.

Leon Walras and Wilfredo Pareto, founders of the so-called Lausanne School at the University of Lausanne in Switzerland, were among the first to explicitly formulate quantitative principles of market economies, stating the principle of economic equilibrium as a mathematical theory. Both worked at a time of great social and economic change. In Pareto's work in particular, pure economics and political science occupy a central place.

Convinced that economics should become a mathematical science, Walras set himself the task of writing the first mathematical general equilibrium system. The British economist Stanley Jevons and the Austrian economist Carl Menger had already formulated the idea of economic equilibrium as a situation where supply and demand match in interrelated markets. Walras's objective—to prove that equilibrium was indeed possible—required the explicit formulation of the equations of supply-and-demand equilibrium.

Walras introduced the idea of *tatonnement* (French for groping) as a process of exploration by which a central auctioneer determines equilibrium prices. A century before, in 1776, Adam Smith had introduced the notion of the "invisible hand" that coordinates the activity of independent competitive agents to achieve desirable global goals. In the modern parlance of complex systems, the "invisible hand" would be called an "emerging property" of competitive markets. Much recent work on complex systems and artificial life has focused

on understanding how the local interaction of individuals might result in complex and purposeful global behavior. Walras was to make the hand “visible” by defining the process of price discovery.

Pareto followed Walras in the Chair of Economics at the University of Lausanne. Pareto’s focus was the process of economic decision making. He replaced the idea of supply-and-demand equilibrium with a more general idea of the ordering of preferences through utility functions. (Pareto used the word “ophelimity” to designate what we would now call utility. The concept of ophelimity is slightly different from the concept of utility insofar as ophelimity includes constraints on people’s preferences.) Equilibrium is reached where marginal utilities are zero. The Pareto system hypothesized that agents are able to order their preferences and take into account constraints in such a way that a numerical index—“utility” in today’s terminology—can be associated with each choice. Note that it was not until 1944 that utility theory was formalized in a set of necessary and sufficient axioms by von Neumann and Morgenstern and applied to decision making under risk and uncertainty.

Economic decision making is therefore based on the maximization of utility. As Pareto assumed utility to be a differentiable function, global equilibrium is reached where marginal utilities (i.e., the partial derivatives of utility) vanish. Pareto was especially interested in the problem of the global optimum of utility. The Pareto optimum is a state in which nobody can be better off without making others worse off. A Pareto optimum does not imply the equal division of resources; quite the contrary, a Pareto optimum might be a maximally unequal distribution of wealth.

A lasting contribution of Pareto is the formulation of a law of income distribution. Known as the *Pareto law*, this law states that there is a linear relationship between the logarithm of the income I and the number N of people that earn

more than this income:

$$\text{Log } N = A + s \log I$$

where A and s are appropriate constants.

The importance of the works of Walras and Pareto were not appreciated at the time. Without digital computers, the equilibrium systems they conceived were purely abstract: There was no way to compute solutions to economic equilibrium problems. In addition, the climate at the turn of the century did not allow a serene evaluation of the scientific merit of their work. The idea of free markets was at the center of heated political debates; competing systems included mercantile economies based on trade restrictions and privileges as well as the emerging centrally planned Marxist economies.

PRICE DIFFUSION: BACHELIER

In 1900, the Sorbonne University student Louis Bachelier presented a doctoral dissertation, *Théorie de la Spéculation*, that was to anticipate much of today’s work in finance theory. Bachelier’s advisor was the great French mathematician Henri Poincaré. There were three notable aspects in Bachelier’s thesis: (1) He argued that in a purely speculative market stock prices should be random; (2) he developed the mathematics of Brownian motion; and (3) he computed the prices of several options.

To appreciate the importance of Bachelier’s work, it should be remarked that at the beginning of the 20th century, the notion of probability was not yet rigorous; the formal mathematical theory of probability was developed only in the 1930s. In particular, the precise notion of the propagation of information essential for the definition of conditional probabilities in continuous time had not yet been formulated.

Anticipating the development of the theory of efficient markets 60 years later, the key economic idea of Bachelier was that asset prices in a speculative market should be a fair game, that

is, a martingale process such that the expected return is zero. According to Bachelier, "The expectation of the speculator is zero." The formal concept of a martingale (i.e., of a process such that its expected value at any moment coincides with the present value) had not yet been introduced in probability theory. In fact, the rigorous notion of conditional probability and filtration were developed only in the 1930s. In formulating his hypothesis on market behavior, Bachelier relied on intuition.

Bachelier actually went much further. He assumed that stock prices evolve as a continuous-time Markov process. This was a brilliant intuition: Markov was to start working on these problems only in 1906. Bachelier established the differential equation for the time evolution of the probability distribution of prices, noting that this equation was the same as the heat diffusion equation. Five years later, in 1905, Albert Einstein used the same diffusion equation for the Brownian motion (i.e., the motion of a small particle suspended in a fluid). Bachelier also made the connection with the continuous limit of random walks, thus anticipating the work of the Japanese mathematician Kiyosi Ito at the end of the 1940s and the Russian mathematician and physicist Ruslan Stratonovich on stochastic integrals at the end of the 1950s.

By computing the extremes of Brownian motion, Bachelier computed the price of several options. He also computed the distributions of a number of functionals of Brownian motion. These were remarkable mathematical results in themselves. Formal proof was given only much later. Even more remarkable, Bachelier established option pricing formulas well before the formal notion of absence of arbitrage was formulated.

Bachelier's work was outside the mainstream of contemporary mathematics but was too mathematically complex for the economists of his time. It wasn't until the formal development of probability theory in 1930s that his ideas became mainstream mathematics and only in the 1960s, with the development of the theory of

efficient markets, that his ideas became part of mainstream finance theory. In an efficient market, asset prices should, in each instant, reflect all the information available at the time, and any event that causes prices to move must be unexpected (i.e., a random disturbance). As a consequence, prices move as martingales, as argued by Bachelier. Bachelier was, in fact, the first to give a precise mathematical structure in continuous time to price processes subject to competitive pressure by many agents.

THE RUIN PROBLEM IN INSURANCE: LUNDBERG

In Uppsala, Sweden, in 1903, three years after Bachelier defended his doctoral dissertation in Paris, Filip Lundberg defended a thesis that was to become a milestone in actuarial mathematics: He was the first to define a collective theory of risk and to apply a sophisticated probabilistic formulation to the insurance ruin problem. The ruin problem of an insurance company in a non-life sector can be defined as follows. Suppose that an insurance company receives a stream of sure payments (premiums) and is subject to claims of random size that occur at random times. What is the probability that the insurer will not be able to meet its obligations (i.e., the probability of ruin)?

Lundberg solved the problem as a collective risk problem, pooling together the risk of claims. To define collective risk processes, he introduced marked Poisson processes. Marked Poisson processes are processes where the random time between two events is exponentially distributed. The magnitude of events is random with a distribution independent of the time of the event. Based on this representation, Lundberg computed an estimate of the probability of ruin.

Lundberg's work anticipated many future developments of probability theory, including what was later to be known as the *theory of point processes*. In the 1930s, the Swedish mathematician and probabilist Harald Cramer

gave a rigorous mathematical formulation to Lundberg's work. A more comprehensive formal theory of insurance risk was later developed. This theory now includes Cox processes—point processes more general than Poisson processes—and fat-tailed distributions of claim size.

A strong connection between actuarial mathematics and asset pricing theory has since been established. (See, for example, Embrechts, Klüppelberg, and Mikosch, 1996). In well-behaved, complete markets, establishing insurance premiums entails principles that mirror asset prices. In the presence of complete markets, insurance would be a risk-free business: There is always the possibility of reinsurance. In markets that are not complete—essentially because they make unpredictable jumps—hedging is not possible; risk can only be diversified and options are inherently risky. Option pricing theory again mirrors the setting of insurance premiums.

Lundberg's work went unnoticed by the actuarial community for nearly 30 years, though this did not stop him from enjoying a successful career as an insurer. Both Bachelier and Lundberg were in advance of their time; they anticipated, and probably inspired, the subsequent development of probability theory. But the type of mathematics implied by their work could not be employed in full earnest prior to the development of digital computers. It was only with digital computers that we were able to tackle complex mathematical problems whose solutions go beyond closed-form formulas.

THE PRINCIPLES OF INVESTMENT: MARKOWITZ

Just how an investor should allocate his resources has long been debated. Classical wisdom suggested that investments should be allocated to those assets yielding the highest returns, without the consideration of correlations. Before the modern formulation of efficient markets, speculators widely acted on the

belief that positions should be taken only if they had a competitive advantage in terms of information. A large amount of resources were therefore spent on analyzing financial information. John Maynard Keynes suggested that investors should carefully evaluate all available information and then make a calculated bet. The idea of diversification was anathema to Keynes, who was actually quite a successful investor.

In 1952, Harry Markowitz, then a graduate student at the University of Chicago, published a seminal article on optimal portfolio selection that upset established wisdom. He advocated that, being risk adverse, investors should diversify their portfolios. (The principles in his article were developed further in his 1959 book.) The idea of making risk bearable through risk diversification was not new: It was widely used by medieval merchants. Markowitz understood that the risk-return trade-off of investments could be improved by diversification and cast diversification in the framework of optimization.

Markowitz was interested in the investment decision-making process. Along the lines set forth by Pareto 60 years earlier, Markowitz assumed that investors order their preferences according to a *utility index*, with utility as a convex function that takes into account investors' risk-return preferences. Markowitz assumed that stock returns are jointly normal. As a consequence, the return of any portfolio is a normal distribution, which can be characterized by two parameters: the mean and the variance. Utility functions are therefore defined on two variables—mean and variance—and the Markowitz framework for portfolio selection is commonly referred to as *mean-variance analysis*. The mean and variance of portfolio returns are in turn a function of a portfolio's weights. Given the variance-covariance matrix, utility is a function of portfolio weights. The investment decision-making process involves maximizing utility in the space of portfolio weights.

The inputs to the mean-variance analysis include expected returns, variance of returns, and

either covariance or correlation of returns between each pair of securities. For example, an analysis that allows 200 securities as possible candidates for portfolio selection requires 200 expected returns, 200 variances of return, and 19,900 correlations or covariances. An investment team tracking 200 securities may reasonably be expected to summarize their analyses in terms of 200 means and variances, but it is clearly unreasonable for them to produce 19,900 carefully considered correlation coefficients or covariances. It was clear to Markowitz that some kind of model of the covariance structure was needed for the practical application of the model. He did little more than point out the problem and suggest some possible models of covariance for research to large portfolios. In 1963, William Sharpe suggested the single index market model as a proxy for the covariance structure of security returns.

Markowitz joined the Rand Corporation, where he met George Dantzig, who introduced him to computer-based optimization technology. Markowitz was quick to appreciate the role that computers would have in bringing mathematics to bear on business problems. Optimization and simulation were on the way to becoming the tools of the future, replacing the quest for closed-form solutions of mathematical problems.

In the following years, Markowitz developed a full theory of the investment management process based on optimization. His optimization theory had the merit of being applicable to practical problems, even outside of the realm of finance. With the progressive diffusion of high-speed computers, the practice of financial optimization has found broad application.

UNDERSTANDING VALUE: MODIGLIANI AND MILLER

At about the same time that Markowitz was tackling the problem of how investors should behave, taking asset price processes as a given, other economists were trying to understand

how markets determine value. Adam Smith had introduced the notion of perfect competition (and therefore perfect markets) in the second half of the 18th century. In a perfect market, there are no impediments to trading: Agents are price takers who can buy or sell as many units as they wish. The neoclassical economists of the 1960s took the idea of perfect markets as a useful idealization of real free markets. In particular, they argued that financial markets are very close to being perfect markets. The theory of asset pricing was subsequently developed to explain how prices are set in a perfect market.

In general, a *perfect market* results when the number of buyers and sellers is sufficiently large, and all participants are small enough relative to the market so that no individual market agent can influence a commodity's price. Consequently, all buyers and sellers are price takers, and the market price is determined where there is equality of supply and demand. This condition is more likely to be satisfied if the commodity traded is fairly homogeneous (for example, corn or wheat).

There is more to a perfect market than market agents being price takers. It is also required that there are no transaction costs or impediments that interfere with the supply and demand of the commodity. Economists refer to these various costs and impediments as "frictions."

The costs associated with frictions generally result in buyers paying more than in the absence of frictions, and/or sellers receiving less. In the case of financial markets, frictions include:

- Commissions charged by brokers.
- Bid-ask spreads charged by dealers.
- Order handling and clearance charges.
- Taxes (notably on capital gains) and government-imposed transfer fees.
- Costs of acquiring information about the financial asset.
- Trading restrictions, such as exchange-imposed restrictions on the size of a position in the financial asset that a buyer or seller may take.

- Restrictions on market makers.
- Halts to trading that may be imposed by regulators where the financial asset is traded.

Modigliani-Miller Irrelevance Theorems and the Absence of Arbitrage

A major step was taken in 1958 when Franco Modigliani and Merton Miller published a then-controversial article in which they maintained that the value of a company does not depend on the capital structure of the firm. (In a 1963 article, they corrected their analysis for the impact of corporate taxes.) The capital structure is the mix of debt and equity used to finance the firm. The traditional view prior to the publication of the article by Modigliani and Miller was that there existed a capital structure that maximized the value of the firm (i.e., there is an optimal capital structure). Modigliani and Miller demonstrated that in the absence of taxes and in a perfect capital market, the capital structure was irrelevant (i.e., the capital structure does not affect the value of a firm). By extension, the irrelevance principle applies to the type of debt a firm may select (e.g., senior, subordinated, secured, and unsecured).

In 1961, Modigliani and Miller published yet another controversial article in which they argued that the value of a company does not depend on the dividends it pays but on its earnings. The basis for valuing a firm—earnings or dividends—had always attracted considerable attention. Because dividends provide the hard cash that remunerates investors, they were considered by many as key to a firm's value.

Modigliani and Miller's challenge to the traditional view that capital structure and dividends matter when determining a firm's value was founded on the principle that the traditional views were inconsistent with the workings of competitive markets where securities are freely traded. In their view, the value of a company is independent of its financial structure: From a valuation standpoint, it does not mat-

ter whether the firm keeps its earnings or distributes them to shareholders.

Known as the *Modigliani-Miller theorems*, these theorems paved the way for the development of arbitrage pricing theory. In fact, to establish their theorems, Modigliani and Miller made use of the notion of *absence of arbitrage*. Absence of arbitrage means that there is no possibility of making a risk-free profit without an investment. This implies that the same stream of cash flows should be priced in the same way across different markets. Absence of arbitrage is the fundamental principle for relative asset pricing; it is the pillar on which derivative pricing rests.

EFFICIENT MARKETS: FAMA AND SAMUELSON

Absence of arbitrage entails market efficiency. Shortly after the Modigliani-Miller theorems had been established, Paul Samuelson in 1965 and Eugene Fama in 1970 developed the notion of efficient markets: A market is efficient if prices reflect all available information. Bachelier had argued that prices in a competitive market should be random conditionally to the present state of affairs. Fama and Samuelson put this concept into a theoretical framework, linking prices to information.

In general, an *efficient market* refers to a market where prices at all times fully reflect all available information that is relevant to the valuation of securities. That is, relevant information about the security is quickly impounded into the price of securities.

Fama and Samuelson define "fully reflects" in terms of the expected return from holding a security. The expected return over some holding period is equal to expected cash distributions plus the expected price change, all divided by the initial price. The price formation process defined by Fama and Samuelson is that the expected return one period from now is a stochastic variable that already takes into account the "relevant" information set. They argued that in a market where information is shared by

all market participants, prices should fluctuate randomly.

A price-efficient market has implications for the investment strategy that investors may wish to pursue. In an active strategy, investors seek to capitalize on what they perceive to be the mispricing of financial instruments (cash instruments or derivative instruments). In a market that is price efficient, active strategies will not consistently generate a return after taking into consideration transaction costs and the risks associated with a strategy that is greater than simply buying and holding securities. This has led investors in certain sectors of the capital market where empirical evidence suggests the sector is price efficient to pursue a strategy of indexing, which simply seeks to match the performance of some financial index. However Samuelson was careful to remark that the notion of efficient markets does not make investment analysis useless; rather, it is a condition for efficient markets.

Another facet in this apparent contradiction of the pursuit of active strategies despite empirical evidence on market efficiency was soon to be clarified. Agents optimize a risk-return trade-off based on the stochastic features of price processes. Price processes are not simply random but exhibit a rich stochastic behavior. The objective of investment analysis is to reveal this behavior.

CAPITAL ASSET PRICING MODEL: SHARPE, LINTNER, AND MOSSIN

Absence of arbitrage is a powerful economic principle for establishing relative pricing. In itself, however, it is not a market equilibrium model. William Sharpe (1964), John Lintner (1965), and Jan Mossin (1966) developed a theoretical equilibrium model of market prices called the *capital asset pricing model* (CAPM). As anticipated 60 years earlier by Walras and Pareto, Sharpe, Lintner, and Mossin developed

the consequences of Markowitz's portfolio selection into a full-fledged stochastic general equilibrium theory.

Asset pricing models categorize risk factors into two types. The first type is risk factors that cannot be diversified away via the Markowitz framework. That is, no matter what the investor does, the investor cannot eliminate these risk factors. These risk factors are referred to as *systematic risk factors* or *nondiversifiable risk factors*. The second type is risk factors that can be eliminated via diversification. These risk factors are unique to the asset and are referred to as *unsystematic risk factors* or *diversifiable risk factors*.

The CAPM has only one systematic risk factor—the risk of the overall movement of the market. This risk factor is referred to as “market risk.” This is the risk associated with holding a portfolio consisting of all assets, called the “market portfolio.” In the market portfolio, an asset is held in proportion to its market value. So, for example, if the total market value of all assets is \$ X and the market value of asset j is \$ Y , then asset j will comprise \$ Y /\$ X of the market portfolio.

The expected return for an asset i according to the CAPM is equal to the risk-free rate plus a risk premium. The risk premium is the product of (1) the sensitivity of the return of asset i to the return of the market portfolio, and (2) the difference between the expected return on the market portfolio and the risk-free rate. It measures the potential reward for taking on the risk of the market above what can be earned by investing in an asset that offers a risk-free rate. Taken together, the risk premium is a product of the quantity of market risk and the potential compensation of taking on market risk (as measured by the second component).

The CAPM was highly appealing from the theoretical point of view. It was the first general-equilibrium model of a market that admitted testing with econometric tools. A critical challenge to the empirical testing of the CAPM as pointed out by Roll (1977) is the identification of the market portfolio.

THE MULTIFACTOR CAPM: MERTON

The CAPM assumes that the only risk that an investor is concerned with is uncertainty about the future price of a security. Investors, however, are usually concerned with other risks that will affect their ability to consume goods and services in the future. Three examples would be the risks associated with future labor income, the future relative prices of consumer goods, and future investment opportunities.

Recognizing these other risks that investors face, Robert Merton (1973b) extended the CAPM based on consumers deriving their optimal lifetime consumption when they face these “extramarket” sources of risk. These extramarket sources of risk are also referred to as “factors,” hence the model derived by Merton is called a multifactor CAPM.

The multifactor CAPM says that investors want to be compensated for the risk associated with each source of extramarket risk, in addition to market risk. In the case of the CAPM, investors hedge the uncertainty associated with future security prices by diversifying. This is done by holding the market portfolio. In the multifactor CAPM, in addition to investing in the market portfolio, investors will also allocate funds to something equivalent to a mutual fund that hedges a particular extramarket risk. While not all investors are concerned with the same sources of extramarket risk, those that are concerned with a specific extramarket risk will basically hedge them in the same way.

The multifactor CAPM is an attractive model because it recognizes nonmarket risks. The pricing of an asset by the marketplace, then, must reflect risk premiums to compensate for these extramarket risks. Unfortunately, it may be difficult to identify all the extramarket risks and to value each of these risks empirically. Furthermore, when these risks are taken together, the multifactor CAPM begins to resemble the arbitrage pricing theory model described next.

ARBITRAGE PRICING THEORY: ROSS

An alternative to the equilibrium asset pricing model just discussed, an asset pricing model based purely on arbitrage arguments, was derived by Stephen Ross (1976). The model, called the *arbitrage pricing theory (APT) model*, postulates that an asset’s expected return is influenced by a variety of risk factors, as opposed to just market risk as assumed by the CAPM. The APT model states that the return on a security is linearly related to H systematic risk factors. However, the APT model does not specify what the systematic risk factors are, but it is assumed that the relationship between asset returns and the risk factors is linear.

The APT model as given asserts that investors want to be compensated for all the risk factors that systematically affect the return of a security. The compensation is the sum of the products of each risk factor’s systematic risk and the risk premium assigned to it by the capital market.

Proponents of the APT model argue that it has several major advantages over the CAPM. First, it makes less restrictive assumptions about investor preferences toward risk and return. As explained earlier, the CAPM theory assumes investors trade off between risk and return solely on the basis of the expected returns and standard deviations of prospective investments. The APT model, in contrast, simply requires that some rather unobtrusive bounds be placed on potential investor utility functions. Second, no assumptions are made about the distribution of asset returns. Finally, since the APT model does not rely on the identification of the true market portfolio, the theory is potentially testable. The model simply assumes that no arbitrage is possible. That is, using no additional funds (wealth) and without increasing risk, it is not possible for an investor to create a portfolio to increase return.

The APT model provides theoretical support for an asset pricing model where there is more than one risk factor. Consequently, models of

this type are referred to as multifactor risk models. These models are applied to portfolio management.

ARBITRAGE, HEDGING, AND OPTION THEORY: BLACK, SCHOLES, AND MERTON

The idea of arbitrage pricing can be extended to any price process. A general model of asset pricing will include a number of independent price processes plus a number of price processes that depend on the first process by arbitrage. The entire pricing structure may or may not be cast in a general equilibrium framework.

Arbitrage pricing allowed derivative pricing. With the development of derivatives trading, the requirement of a derivative valuation and pricing model made itself felt. The first formal solution of the option pricing model was developed independently by Fisher Black and Myron Scholes (1973), working together, and in the same year by Robert Merton (1973a).

The solution of the option pricing problem proposed by Black, Scholes, and Merton was simple and elegant. Suppose that a market contains a risk-free bond, a stock, and an option. Suppose also that the market is arbitrage-free and that stock price processes follow a continuous-time geometric Brownian motion. Black, Scholes, and Merton demonstrated that it is possible to construct a portfolio made up of the stock plus the bond that perfectly replicates the option. The replicating portfolio can be exactly determined, without anticipation, solving a partial differential equation.

The idea of replicating portfolios has important consequences. Whenever a financial instrument (security or derivative instrument) process can be exactly replicated by a portfolio of other securities, absence of arbitrage requires that the price of the original financial instrument coincide with the price of the replicating portfolio. Most derivative pricing algorithms are based on this principle: To price a deriva-

tive instrument, one must identify a replicating portfolio whose price is known.

Pricing by portfolio replication received a powerful boost with the discovery that calculations can be performed in a risk-neutral probability space where processes assume a simplified form. The foundation was thus laid for the notion of *equivalent martingales*, developed by Michael Harrison and David Kreps (1979) and Michael Harrison and Stanley Pliska (1981). Not all price processes can be reduced in this way: If price processes do not behave sufficiently well (i.e., if the risk does not vanish with the vanishing time interval), then replicating portfolios cannot be found. In these cases, risk can be minimized but not hedged.

KEY POINTS

- The development of mathematical finance began at the end of the nineteenth century with work on general equilibrium theory by Walras and Pareto.
- At the beginning of the twentieth century, Bachelier and Lundberg made a seminal contribution, introducing respectively Brownian motion price processes and Markov Poisson processes for collective risk events.
- The advent of digital computers enabled the large-scale application of advanced mathematics to finance theory, ushering in optimization and simulation.
- In 1952, Markowitz introduced the theory of portfolio optimization, which advocates the strategy of portfolio diversification.
- In 1961, Modigliani and Miller argued that the value of a company is based not on its dividends and capital structure, but on its earnings; their formulation was to be called the Modigliani-Miller theorem.
- In the 1960s, major developments included the efficient market hypothesis (Samuelson and Fama), the capital asset pricing model (Sharpe, Lintner, and Mossin), and the multifactor CAPM (Merton).

- In the 1970s, major developments included the arbitrage pricing theory (Ross) that led to multifactor models and option pricing formulas (Black, Scholes, and Merton) based on replicating portfolios, which are used to price derivatives if the underlying price processes are known.

REFERENCES

- Arrow, K. J., and Debreu, G. (1954). The existence of an equilibrium for a competitive economy. *Econometrica* 22: 265–290.
- Bachelier, L. [1900] 2006. *Louis Bachelier's Theory of Speculation: The Origins of Modern Finance*. Translated by Mark Davis and Alison Etheridge. Princeton, NJ: Princeton University Press.
- Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 3: 637–654.
- Embrechts, P., Klüppelberg, C., and Mikosch, T. (1996). *Modelling Extremal Events for Insurance and Finance*. Berlin: Springer.
- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25, 2: 383–417.
- Harrison, J. M., and Kreps, D. M. (1979). Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory* 20: 381–408.
- Harrison, J. M., and Pliska, S. (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and Their Applications* 11: 313–316.
- Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolio and capital budgets. *Review of Economics and Statistics* 47, 1: 13–37.
- Lundberg, F. (1903). *Approximerad framställning af sannolikhetsfunktionerna. II. Aterförsäkring af kollektivrisken*. Uppsala: Almqvist & Wiksell.
- Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance* 7, 1: 77–91.
- Markowitz, H. M. (1959). *Portfolio Selection: Efficient Diversification of Investments*. New York: John Wiley.
- Merton, R. C. (1973a). Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4: 141–183.
- Merton, R. C. (1973b). An intertemporal capital asset pricing model. *Econometrica* 41, 5: 867–888.
- Miller, M. H., and Modigliani, F. (1961). Dividend policy, growth, and the valuation of shares. *Journal of Business* 3: 411–433.
- Modigliani, F., and Miller, M. H. (1958). The cost of capital, corporation finance, and the theory of investment. *American Economic Review* 48, 3: 261–297.
- Modigliani, F., and Miller, M. H. (1963). Corporate income taxes and the cost of capital: A correction. *American Economic Review* 53, 3: 433–443.
- Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica* 34, 4: 768–783.
- Pareto, V. (1906). *Manuel d'économie Politique* (Manual of Political Economy). Translated by A. S. Schwier from the 1906 edition. New York: A. M. Kelley.
- Roll, R. R. (1977). A critique of the asset pricing theory's tests. *Journal of Financial Economics* 4: 129–176.
- Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory* 13: 343–362.
- Samuelson, P. A. (1965). Proof that the properly anticipated prices fluctuate randomly. *Industrial Management Review* 6, 2: 41–50.
- Sharpe, W. F. (1963). A simplified model for portfolio analysis. *Management Science* 9, 2: 277–293.
- Sharpe, W. F. (1964). Capital asset prices. *Journal of Finance* 19, 3: 425–442.
- Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*. Reprinted, University of Chicago Press, 1976.
- Von Neumann, J., and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.
- Walras, L. (1874). *Elements of Pure Economics*. Reprinted, Harvard University Press, 1954.

From Art to Financial Modeling

SERGIO M. FOCARDI, PhD

Partner, The Intertek Group

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: It is often said that investment management is an art, not a science. However, since the early 1990s the market has witnessed a progressive shift toward a more industrial view of the investment management process. There are several reasons for this change. First, with globalization the universe of investable assets has grown many times over. Asset managers might have to choose from among several thousand possible investments from around the globe. The S&P 500 index is itself chosen from a pool of 8,000 investable U.S. stocks. Second, institutional investors, often together with their investment consultants, have encouraged asset management firms to adopt an increasingly structured process with documented steps and measurable results. Pressure from regulators and the media is another factor. Lastly, the sheer size of the markets makes it imperative to adopt safe and repeatable methodologies. The volumes are staggering.

In its modern sense, financial modeling is the design (or engineering) of contracts and portfolios of contracts that result in predetermined cash flows contingent on different events. Broadly speaking, financial models are employed to manage investment portfolios and risk. The objective is the transfer of risk from one entity to another via appropriate contracts. Though the aggregate risk is a quantity that cannot be altered, risk can be transferred if there is a willing counterparty.

Financial modeling came to the forefront of finance in the 1980s with the broad diffusion of derivative instruments. However, the concept and practice of financial modeling are quite old. Evidence of the use of sophisticated cross-border instruments of credit and pay-

ment dating from the time of the First Crusade (1095–1099) has come down to us from the letters of Jewish merchants in Cairo. The notion of the diversification of risk (central to modern risk management) and the quantification of insurance risk (a requisite for pricing insurance policies) were already understood, at least in practical terms, in the 14th century. The rich epistolary of Francesco Datini, a 14th-century merchant, banker, and insurer from Prato (Tuscany, Italy), contains detailed instructions to his agents on how to diversify risk and insure cargo. It also gives us an idea of insurance costs: Datini charged 3.5% to insure a cargo of wool from Malaga to Pisa and 8% to insure a cargo of malmsey (sweet wine) from Genoa to Southampton, England. These, according to

one of Datini's agents, were low rates: He considered 12–15% a fair insurance premium for similar cargo.

What is specific to modern financial modeling is the quantitative management of risk. Both the pricing of contracts and the optimization of investments require some basic capabilities of statistical modeling of financial contingencies. It is the size, diversity, and efficiency of modern competitive markets that makes the use of modeling imperative.

THE ROLE OF INFORMATION TECHNOLOGY

Advances in information technology are behind the widespread adoption of modeling in finance. The most important advance has been the enormous increase in the amount of computing power, concurrent with a steep fall in prices. Government agencies have long been using computers for economic modeling, but private firms found it economically justifiable only as of the 1980s. Back then, economic modeling was considered one of the "Grand Challenges" of computational science (a term coined by Kenneth Wilson [1989], recipient of the 1982 Nobel Prize in Physics, and later adopted by the U.S. Department of Energy in its High Performance Communications and Computing Program, which included economic modeling among the grand challenges).

In the late 1980s, firms such as Merrill Lynch began to acquire supercomputers to perform derivative pricing computations. The overall cost of these supercomputing facilities, in the range of several million dollars, limited their diffusion to the largest firms. Today, computational facilities ten times more powerful cost only a few thousand dollars. To place today's computing power in perspective, consider that a 1990 run-of-the-mill Cray supercomputer cost several million U.S. dollars and had a clock cycle of 4 nanoseconds (i.e., 4 billionths of a second or 250 million cycles per second, notated as 250 MHz). Today's fast laptop

computers are 10 times faster with a clock cycle of 2.5 GHz and, at a few thousand dollars, cost only a fraction of the price. Supercomputer performance has itself improved significantly, with top computing speed in the range of several teraflops compared to the several megaflops of a Cray supercomputer in the 1990s. (Flops, which stands for floating point operations per second, is a measure of computational speed. A teraflop computer is a computer able to perform a trillion floating point operations per second.) In the space of 15 years, sheer performance has increased 1,000 times while the price-performance ratio has decreased by a factor of 10,000. Storage capacity has followed similar dynamics.

The diffusion of low-cost, high-performance computers has allowed the broad use of numerical methods. Computations that were once performed by supercomputers in air-conditioned rooms are now routinely performed on desktop machines. This has changed the landscape of financial modeling. The importance of finding closed-form solutions and the consequent search for simple models has been dramatically reduced. Computationally intensive methods such as Monte Carlo simulations and the numerical solution of differential equations are now widely used. As a consequence, it has become feasible to represent prices and returns with relatively complex models. *Non-normal probability distributions* have become commonplace in many sectors of financial modeling. It is fair to say that the key limitation of financial econometrics is now the size of available data samples or training sets, not the computations; it is the data that limit the complexity of estimates.

Mathematical modeling has also undergone major changes. Techniques such as equivalent martingale methods are being used in derivative pricing and cointegration, the theory of fat-tailed processes, and state-space modeling (including ARCH/GARCH and stochastic volatility models) are being used in econometrics.

Powerful specialized mathematical languages and vast statistical software libraries have been developed. The ability to program sequences of statistical operations within a single programming language has been a big step forward. Software firms such as Mathematica and Mathworks, and major suppliers of statistical tools such as SAS, have created simple computer languages for the programming of complex sequences of statistical operations. This ability is key to financial econometrics, which entails the analysis of large portfolios. (Note that although a number of highly sophisticated statistical packages are available to economists, these packages do not serve the needs of the financial econometrician who has to analyze a large number of time series.)

Presently only large or specialized firms write complex applications from scratch; this is typically done to solve specific problems, often in the derivatives area. The majority of financial modelers make use of high-level software programming tools and statistical libraries. It is difficult to overestimate the advantage brought by these software tools; they cut development time and costs by orders of magnitude.

In addition, there is a wide range of off-the-shelf financial applications that can be used directly by operators who have a general understanding of the problem but no advanced statistical or mathematical training. For example, powerful complete applications from firms such as MSCI Barra and component applications from firms such as FEA make sophisticated analytical methods available to a large number of professionals.

Data have, however, remained a significant expense. The diffusion of electronic transactions has made available large amounts of data, including high-frequency data (HFD), which gives us information at the transaction level. As a result, in budgeting for financial modeling, data have become an important factor in deciding whether to undertake a new modeling effort.

A lot of data are now available free on the Internet. If the required granularity of data is not high, these data allow one to study the viability of models and to perform rough tuning. However, real-life applications, especially applications based on finely grained data, require data streams of a higher quality than those typically available free on the Internet.

INTEGRATING QUALITATIVE AND QUANTITATIVE INFORMATION

Textual information has remained largely outside the domain of *quantitative modeling*, having long been considered the domain of judgment. This is now changing as financial firms begin to tackle the problem of what is commonly called *information overload*; advances in computer technology are again behind the change (see Jonas and Focardi, 2002). Reuters publishes the equivalent of three bibles of (mostly financial) news daily; it is estimated that five new research documents come out of Wall Street every minute; asset managers at medium-sized firms report receiving up to 1,000 e-mails daily and work with as many as five screens on their desk. Conversely, there is also a lack of “digested” information. It has been estimated that only one third of the roughly 10,000 U.S. public companies are covered by meaningful Wall Street research; there are thousands of companies quoted on the U.S. exchanges with no Wall Street research at all. It is unlikely the situation is better relative to the tens of thousands of firms quoted on other exchanges throughout the world. Yet increasingly companies are providing information, including press releases and financial results, on their Web sites.

Such *unstructured (textual) information* is progressively being transformed into self-describing, *semistructured information* that can be automatically categorized and searched by

computers. A number of developments are making this possible. These include:

- The development of XML (eXtensible Markup Language) standards for tagging textual data. This is taking us from free text search to queries on semistructured data.
- The development of RDF (Resource Description Framework) standards for appending metadata. This provides a description of the content of documents.
- The development of algorithms and software that generate taxonomies and perform automatic categorization and indexation.
- The development of database query functions with a high level of expressive power.
- The development of high-level text mining functionality that allows “discovery.”

The emergence of standards for the handling of “meaning” is a major development. It implies that unstructured textual information, which some estimates put at 80% of all content stored in computers, will be largely replaced by semistructured information ready for machine handling at a semantic level. Today’s standard structured databases store data in a prespecified format so that the position of all elementary information is known. For example, in a trading transaction, the date, the amount exchanged, the names of the stocks traded, and so on are all stored in predefined fields. However, textual data such as news or research reports do not allow such a strict structuring. To enable the computer to handle such information, a descriptive metafile is appended to each unstructured file. The descriptive metafile is a structured file that contains the description of the key information stored in the unstructured data. The result is a semistructured database made up of unstructured data plus descriptive metafiles.

Industry-specific and application-specific standards are being developed around the general-purpose XML. At the time of this writing, there are numerous initiatives established with the objective of defining XML standards for applications in finance, from time series to

analyst and corporate reports and news. While it is not yet clear which of the competing efforts will emerge as the de facto standards, attempts are now being made to coordinate standardization efforts, eventually adopting the ISO 15022 central data repository as an integration point.

Technology for handling unstructured data has already made its way into the industry. Factiva, a Dow Jones-Reuters company, uses commercially available text mining software to automatically code and categorize more than 400,000 news items daily, in real time (prior to adopting the software, they manually coded and categorized some 50,000 news articles daily). Users can search the Factiva database, which covers 118 countries and includes some 8,000 publications and more than 30,000 company reports with simple intuitive queries expressed in a language close to the natural language. Several firms use text mining technology in their Web-based research portals for clients on the buy and sell sides. Such services typically offer classification, indexation, tagging, filtering, navigation, and search.

These technologies are helping to organize research flows. They allow us to automatically aggregate, sort, and simplify information and provide the tools to compare and analyze the information. In serving to pull together material from myriad sources, these technologies will not only form the basis of an internal knowledge management system but allow us to better structure the whole investment management process. Ultimately, the goal is to integrate data and text mining in applications such as fundamental research and event analysis, linking news, and financial time series.

PRINCIPLES FOR ENGINEERING A SUITE OF MODELS

Creating a suite of models to satisfy the needs of a financial firm is engineering in full earnest. It begins with a clear statement of the objectives.

In the case of financial modeling, the objective is identified by the type of decision-making process that a firm wants to implement. The engineering of a suite of financial models requires that the process on which decisions are made is fully specified and that the appropriate information is supplied at every step. This statement is not as banal as it might seem.

We have now reached the stage where, in some markets, financial decision making can be completely automated through optimizers. As we will see in the following entries, one can define models able to construct a conditional probability distribution of returns. An optimizer will then translate the forecast into a tradable portfolio. The manager becomes a kind of high-level supervisor of an otherwise automated process.

However, not all financial decision-making applications are, or can be, fully automated. In many cases, it is the human operator who makes the decision, with models supplying the information needed to arrive at the decision. Building an effective suite of financial models requires explicit decisions as to (1) what level of automation is feasible and desirable, and (2) what information or knowledge is required.

The integration of different models and of qualitative and quantitative information is a fundamental need. This calls for integration of different statistical measures and points of view. For example, an asset management firm might want to complement a portfolio optimization methodology based on Gaussian forecasting with a risk management process based on extreme value theory. The two processes offer complementary views. In many cases, however, different methodologies give different

results though they work on similar principles and use the same data. In these cases, integration is delicate and might run against statistical principles.

In deciding which modeling efforts to invest in, many firms have in place a sophisticated evaluation system. Firms evaluate a model's return on investment and how much it will cost to buy the data necessary to run the model.

KEY POINTS

- Key to a quantitative framework is the measurement and management of uncertainty (i.e., risk) and financial modeling.
- Modeling is the tool to achieve these objectives; advances in information technology are the enabler.
- Unstructured textual information is progressively being transformed into self-describing, semistructured information, allowing a better structuring of the research process.
- After nearly two decades of experience with quantitative methods, market participants now more clearly perceive the benefits and the limits of modeling; given today's technology and markets, the need to better integrate qualitative and quantitative information is clearly felt.

REFERENCES

- Jonas, C., and Focardi, S. M. (2002). *Leveraging Unstructured Data in Investment Management*. Paris: The Intertek Group.
- Wilson, K. (1989). Grand challenges to computational science. *Future Generation Computer Systems* 5: 171.

Basic Data Description for Financial Modeling and Analysis

MARKUS HÖCHSTÖTTER, PhD

Assistant Professor, University of Karlsruhe

SVETLOZAR T. RACHEV, PhD, DrSci

Frey Family Foundation Chair-Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: We are confronted with data every day, constantly. Daily newspapers contain information on stock prices, economic figures, quarterly business reports on earnings and revenues, and much more. These data offer observed values of given quantities. The basic data types can be qualitative, ordinal, or quantitative.

In this entry, we will present the first essentials of data description. We describe all data types and levels. We explain and illustrate why one has to be careful about the permissible computations concerning each data level.¹

We will restrict ourselves to *univariate data*, that is, data of only one dimension. For example, if you follow the daily returns of one particular stock, you obtain a one-dimensional series of observations. If you had observed two stocks, then you would have obtained a two-dimensional series of data, and so on. Moreover, the notion of frequency distributions, empirical frequency distributions, and cumulative frequency distributions is introduced. The goal of this entry is to provide the first methods

necessary to begin data analysis. After reading this entry you will learn how to formalize the first impression you obtain from the data in order to retrieve the most basic structure inherent in the data. That is essential for any subsequent tasks you may undertake with the data. Above all, though, you will have to be fully aware of what you want to learn from the data. That step is maybe the most important task before getting started in investigating the data. For example, you may just want to know what the minimum return has been of your favorite stock during the last year before you decide to purchase. Or you are interested in all returns from last year to learn how this stock typically performs, that is, which returns occur more often than others,

and how often. In the latter case, you definitely have to be more involved to obtain the necessary information than in the first case.

DATA TYPES

Data are gathered by several methods. In the financial industry, we have market data based on regular trades recorded by the exchanges. These data are directly observable. Aside from the regular trading process, there is so-called over-the-counter (OTC) business whose data are less accessible. Annual reports and quarterly reports, on the other hand, are published by companies themselves in print or electronically. These data are available also in the business and finance sections of most major business oriented print media and the Internet. The fields of marketing and the social sciences know additional forms of data collection methods. There are telephone surveys, mail questionnaires, and even experiments.

If one does research on certain financial quantities of interest, one might find the data available from either free or commercial databases. Hence, one must be concerned with the quality of the data. Unfortunately, very often databases of unrestricted access such as those available on the Internet may be of limited credibility. In contrast, there are many commercial purveyors of financial data who are generally acknowledged as providing accurate data. But, as always, quality may have its price.

Information Contained in the Data

Once the data are gathered, it is the objective of descriptive statistics to visually and computationally convert the amount of information given into quantities revealing the essentials in which we are interested. Commonly in this context, visual support is added since very often that allows for a much easier grasp of the information.

The field of descriptive statistics discerns different types of data. Very generally, there are two types: qualitative and quantitative data.

If certain attributes of an item can only be assigned to categories, these data are referred to as qualitative. For example, stocks listed on the New York Stock Exchange (NYSE) can be categorized as belonging to a specific industry sector such as “banking,” “energy,” “media and telecommunications,” and so on. That way, we assign the item stock as its attribute sector one or possibly more values from the set containing banking, energy, media and telecommunications, and so on. (Instead of attribute, we will most of the time use the term “variable.”) Another example would be the credit ratings assigned to debt obligations by commercial rating companies such as Standard & Poor’s, Moody’s, and Fitch Ratings. Except for retrieving the value of an attribute, nothing more can be done with *qualitative data*. One may use a numerical code to indicate the different sectors (e.g., 1 = banking, 2 = energy, and so on). However, we are not allowed to perform any computation with these figures since they are simply proxies of the underlying attribute sector.

However, if an item is assigned a quantitative variable, the value of this variable is numerical. Generally, all real numbers are eligible. Depending on the case, however, one will use discrete values only, such as integers. Stock prices or dividends, for example, are *quantitative data* drawing from—up to some digits—positive real numbers. Quantitative data have the feature that one can perform transformations and computations with them. One can easily think of the average price of all companies comprising some index on a certain day, while it would make absolutely no sense to do the same with qualitative data.

Data Levels and Scale

In descriptive statistics, we group data according to measurement levels. The measurement level gives an indication as to the sophistication of the analysis techniques that one can

apply to the data collected. Typically, a hierarchy with five levels of measurement—nominal, ordinal, interval, ratio, and absolute—is used to group data. The latter three form the set of quantitative data. If the data are of a certain measurement level, they are said to be scaled accordingly. That is, the data are referred to as nominally scaled, and so on.

Nominally scaled data are on the bottom of the hierarchy. Despite the low level of sophistication, this type of data are commonly used. An example is the attribute sector of stocks. We already learned that, even though we can assign numbers as proxies to nominal values, these numbers have no numerical meaning whatsoever. We might just as well assign letters to the individual nominal values, for example, “B = banking,” “E = energy,” and so on.

Ordinally scaled data are one step higher in the hierarchy. We also refer to this type as “rank data,” since we can already perform a ranking within the set of values. We can make use of a relationship among the different values by treating them as quality grades. For example, we can divide the stocks listed in a particular stock index according to their market capitalization into five groups of equal size. Let “A” denote the top 20% of the stocks. Also, let “B” denote the next 20% below, and so on, until we obtain the five groups: A, B, C, D, and E. After ordinal scaling, we can make statements such as “Group A is better than group C.” Hence, we have a natural ranking or order among the values. However, we cannot quantify the difference between them. Also, the credit rating of debt obligations is ordinarily scaled.

Until now, we can summarize that while we can test the relationship between nominal data for equality only, we can additionally determine a greater or less than relationship between *ordinal data*.

Data on an *interval scale* are given if they can be reasonably transformed by a linear equation. Suppose we are given values x . It is now feasible to express a new variable y by the relationship $y = a^* x + b$, where the x 's are our original

data. If x has a meaning, then so does y . It is obvious that data have to possess a numerical meaning and therefore be quantitative in order to be measured on an interval scale. For example, consider the temperature F given in degrees Fahrenheit. Then, the corresponding temperature in degrees Celsius, C , will result from the equation $C = (F - 32)/ 1.8$. Equivalently, if one is familiar with physics, the same temperature measured in degrees Kelvin, K , will result from $K = C + 273.15$. So, say it is 55° Fahrenheit for Americans, the same temperature will mean approximately 13° Celsius for Europeans, and they will not feel any cooler. Generally, interval data allow for the calculation of differences. For example, 70° – 60° Fahrenheit = 10° Fahrenheit may reasonably express the difference in temperature between Los Angeles and San Francisco. But be careful—the difference in temperature measured in Celsius between the two cities is not the same. How much is it?

Data measured on a *ratio scale* share all the properties of interval data. In addition, ratio data have a fixed or true zero point. This is not the case with interval data. Their intercept, b , can be arbitrarily changed through transformation. Since the zero point of ratio data is invariable, one can only transform the slope, a . So, for example, $y = a^* x$ is always a multiple of x . In other words, there is a relationship between y and x given by the ratio a , hence the name used to describe this type of data. One would not have this feature if one would permit some b different from zero in the transformation. Consider, for example, the stock price, E , of some European stock given in euro units. The same price in U.S. dollars, D , would be D equals E times the exchange rate between euros and U.S. dollars. But if the company's price after bankruptcy went to zero, the price in either currency would be zero, even at different rates determined by the ratio of U.S. dollar per euro. This is a result of the invariant zero point.

Absolute data are given by quantitative data measured on a scale even stricter than for

ratio data. Here, along with the zero point, the units are invariant as well. Data measured on an absolute scale occur when transformation would be mathematically feasible but lacks any interpretational implication. A common example is provided by counting numbers. Anybody would agree on the number of stocks listed in a certain stock index. There is no ambiguity as to the zero point and the count increments. If one stock is added to the index, it is immediately clear that the difference to the content of the old index is exactly one unit of stock, assuming that no stock is deleted. This absolute scale is the most intuitive and needs no further discussion.

Cross-Sectional and Time Series Data

There is another way of classifying data. Imagine collecting data from one and the same quantity of interest or variable. A variable is some quantity that can assume values from a value set. For example, the variable “stock price” can technically assume any nonnegative real number of currency but only one value at a time. Each day, it assumes a certain value, which is the day’s stock price. As another example, a variable could be the dividend payments from a specific company over some period of time. In the case of dividends, the observations are made each quarter. The accumulated data then form what is called *time series data*. In contrast, one could pick a particular time period of interest such as the first quarter of the current year and observe the dividend payments of all companies listed in the Standard & Poor’s 500 index. By doing so, one would obtain *cross-sectional data* of the universe of stocks in the S&P 500 index at that particular time.

Summarizing, time series data are data related to a variable successively observed at a sequence of points in time. Cross-sectional data are values of a particular variable across some universe of items observed at a unique point in time. This is visualized in Figure 1.

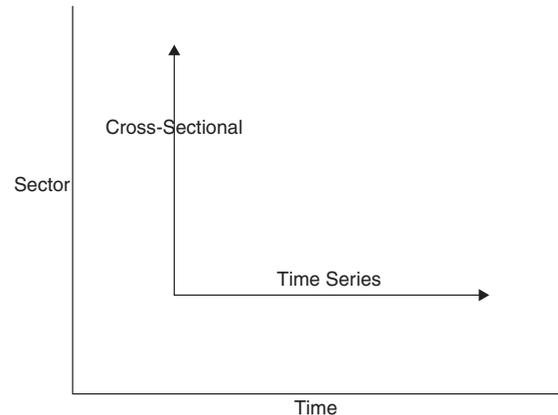


Figure 1 Relationship between Cross-Sectional and Time Series Data

FREQUENCY DISTRIBUTIONS

Sorting and Counting Data

One of the most important aspects when dealing with data is that they are effectively organized and transformed in order to convey the essential information contained in them. This processing of the original data helps to display the inherent meaning in a way that is more accessible for intuition. But before advancing to the graphical presentation of the data, we will first describe the methods of structuring data.

Suppose that we are interested in a particular variable that can assume a set of either finite or infinitely many values. These values may be qualitative or quantitative by nature. In either case, the initial step when obtaining a data sample for some variable is to sort the values of each observation and then to determine the frequency distribution of the dataset. This is done simply by counting the number of observations for each possible value of the variable. Alternatively, if the variable can assume values on all or part of the real line, the frequency can be determined by counting the number of observations that fall into nonoverlapping intervals partitioning the real line.

In our illustration, we will begin with qualitative data first and then move on to the

Table 1 DJIA Components as of December 12, 2006

Company	Industrial Classification Benchmark (ICB) Subsector
3M Co.	Diversified Industrials
Alcoa Inc.	Aluminum
Altria Group Inc.	Tobacco
American Express Co.	Consumer Finance
American International Group Inc.	Full Line Insurance
AT&T Inc.	Fixed Line Telecommunications
Boeing Co.	Aerospace
Caterpillar Inc.	Commercial Vehicles & Trucks
Citigroup Inc.	Banks
Coca-Cola Co.	Soft Drinks
E.I. DuPont de Nemours & Co.	Commodity Chemicals
Exxon Mobil Corp.	Integrated Oil & Gas
General Electric Co.	Diversified Industrials
General Motors Corp.	Automobiles
Hewlett-Packard Co.	Computer Hardware
Home Depot Inc.	Home Improvement Retailers
Honeywell International Inc.	Diversified Industrials
Intel Corp.	Semiconductors
International Business Machines Corp.	Computer Services
Johnson & Johnson	Pharmaceuticals
JPMorgan Chase & Co.	Banks
McDonald's Corp.	Restaurants & Bars
Merck & Co. Inc.	Pharmaceuticals
Microsoft Corp.	Software
Pfizer Inc.	Pharmaceuticals
Procter & Gamble Co.	Nondurable Household Products
United Technologies Corp.	Aerospace
Verizon Communications Inc.	Fixed Line Telecommunications
Wal-Mart Stores Inc.	Broadline Retailers
Walt Disney Co.	Broadcasting & Entertainment

quantitative aspects in the sequel. For example, suppose we want to analyze the frequency of the industry subsectors of the components listed in the Dow Jones Industrial Average (DJIA), an index comprised of 30 U.S. stocks. Table 1 displays the 30 companies in the index along with their respective industry sectors as of December 12, 2006. By counting the observed number of each possible Industry Clas-

Table 2 Frequency Distribution of the Industry Subsectors

ICB Subsector	Frequency a_i
Aerospace	2
Aluminum	1
Automobiles	1
Banks	2
Broadcasting & Entertainment	1
Broadline Retailers	1
Commercial Vehicles & Trucks	1
Commodity Chemicals	1
Computer Hardware	1
Computer Services	1
Consumer Finance	1
Diversified Industrials	3
Fixed Line Telecommunications	2
Full Line Insurance	1
Home Improvement Retailers	1
Integrated Oil & Gas	1
Nondurable Household Products	1
Pharmaceuticals	3
Restaurants & Bars	1
Semiconductors	1
Soft Drinks	1
Software	1
Tobacco	1

sification Benchmark (ICB) subsector, we obtain Table 2, which shows the frequency distribution of the variable subsector. Note in the table that many subsector values appear only once. Hence, this might suggest employing a coarser set for the ICB subsector values in order to reduce the amount of information in the data to a necessary minimum.

Now suppose you would like to compare this to the Dow Jones Global Titans 50 Index (DJGTI). This index includes the 50 largest-capitalization and best-known blue-chip companies listed on the NYSE. The companies contained in this index are listed in Table 3 along with their respective ICB subsectors. The next step would also be to sort the data according to their values and count each hit of a value, finally listing the respective count numbers for each value. A problem arises now, however, when you want to directly compare the numbers with those obtained for the DJIA because the number of stocks contained in each index is not the same. Hence, we cannot compare the respective

Table 3 Dow Jones Global Titans 50 Index as of December 12, 2006

Company Name	ICB Subsector
Abbott Laboratories	Pharmaceuticals
Altria Group Inc.	Tobacco
American International Group Inc.	Full Line Insurance
Astrazeneca PLC	Pharmaceuticals
AT&T Inc.	Fixed Line Telecommunications
Bank of America Corp.	Banks
Barclays PLC	Banks
BP PLC	Integrated Oil & Gas
Chevron Corp.	Integrated Oil & Gas
Cisco Systems Inc.	Telecommunications Equipment
Citigroup Inc.	Banks
Coca-Cola Co.	Soft Drinks
ConocoPhillips	Integrated Oil & Gas
Dell Inc.	Computer Hardware
ENI S.p.A.	Integrated Oil & Gas
Exxon Mobil Corp.	Integrated Oil & Gas
General Electric Co.	Diversified Industrials
GlaxoSmithKline PLC	Pharmaceuticals
HBOS PLC	Banks
Hewlett-Packard Co.	Computer Hardware
HSBC Holdings PLC (UK Reg)	Banks
ING Groep N.V.	Life Insurance
Intel Corp.	Semiconductors
International Business Machines Corp.	Computer Services
Johnson & Johnson	Pharmaceuticals
JPMorgan Chase & Co.	Banks
Merck & Co. Inc.	Pharmaceuticals
Microsoft Corp.	Software
Mitsubishi UFJ Financial Group Inc.	Banks
Morgan Stanley	Investment Services
Nestle S.A.	Food Products
Nokia Corp.	Telecommunications Equipment
Novartis AG	Pharmaceuticals
PepsiCo Inc.	Soft Drinks
Pfizer Inc.	Pharmaceuticals
Procter & Gamble Co.	Nondurable Household Products
Roche Holding AG Part. Cert.	Pharmaceuticals
Royal Bank of Scotland Group PLC	Banks
Royal Dutch Shell PLC A	Integrated Oil & Gas
Samsung Electronics Co. Ltd.	Semiconductors
Siemens AG	Electronic Equipment
Telefonica S.A.	Fixed Line Telecommunications
Time Warner Inc.	Broadcasting & Entertainment
Total S.A.	Integrated Oil & Gas
Toyota Motor Corp.	Automobiles
UBS AG	Banks
Verizon Communications Inc.	Fixed Line Telecommunications
Vodafone Group PLC	Mobile Telecommunications
Wal-Mart Stores Inc.	Broadline Retailers
Wyeth	Pharmaceuticals

Table 4 Comparison of Relative Frequencies of DJIA and DJGTI

ICB Subsector	Relative Frequencies	
	DJIA	DJGTI
Aerospace	0.067	0.000
Aluminum	0.033	0.000
Automobiles	0.033	0.020
Banks	0.067	0.180
Broadcasting & Entertainment	0.033	0.020
Broadline Retailers	0.033	0.020
Commercial Vehicles & Trucks	0.033	0.000
Commodity Chemicals	0.033	0.000
Computer Hardware	0.033	0.040
Computer Services	0.033	0.020
Consumer Finance	0.033	0.000
Diversified Industrials	0.100	0.020
Electronic Equipment	0.000	0.020
Fixed Line Telecommunications	0.067	0.060
Food Products	0.000	0.020
Full Line Insurance	0.033	0.020
Home Improvement Retailers	0.033	0.000
Integrated Oil & Gas	0.033	0.140
Investment Services	0.000	0.020
Life Insurance	0.000	0.020
Mobile Telecommunications	0.000	0.020
Nondurable Household Products	0.033	0.020
Pharmaceuticals	0.100	0.180
Restaurants & Bars	0.033	0.000
Semiconductors	0.033	0.040
Soft Drinks	0.033	0.040
Software	0.033	0.020
Telecommunications Equipment	0.000	0.040
Tobacco	0.033	0.020

absolute frequencies. Instead, we have to resort to something that creates comparability of the two datasets. This is done by expressing the number of observations of a particular value as the proportion of the total number of observations in a specific dataset. That means we have to compute the relative frequency. See Table 4.

Formal Presentation of Frequency

For a better formal presentation, we denote the (absolute) frequency by a and, in particular, by a_i for the i th value of the variable. Formally, the relative frequency f_i of the i th value is, then, defined by

$$f_i = \frac{a_i}{n}$$

where n is the total number of observations. With k being the number of the different values, the following holds:

$$n = \sum_{i=1}^k f_i$$

In our illustration, let $n_1 = 30$ be the number of total observations in the DJIA and $n_2 = 50$ the total number of observations in the DJGTI. Table 4 shows the relative frequencies for all possible values. Notice that each index has some values that were observed with zero frequency, which still have to be listed for comparison. When we look at the DJIA, we find out that the sectors Diversified Industrials and Pharmaceuticals each account for 10% of all sectors and therefore are the sectors with the highest frequencies. Comparing these two sectors to the DJGTI, we find out that Pharmaceuticals play as important a role as a sector with an 18% share, while Diversified Industrials are of minor importance. In this index, Banks are a very important sector with 18% also. A comparison of this sort can now be carried through for all subsectors thanks to the relative frequencies.

Naturally, frequency (absolute and relative) distributions can be computed for all types of data since they do not require that the data have a numerical value.

EMPIRICAL CUMULATIVE FREQUENCY DISTRIBUTION

Accumulating Frequencies

In addition to the frequency distribution, there is another quantity of interest for comparing data that is closely related to the absolute or relative frequency distribution. Suppose that one is interested in the percentage of all large-capitalization stocks in the DJIA with closing prices of at most US \$50 on a specific day. One can sort the observed closing prices by their numerical values in ascending order to obtain something like the array shown in Table 5 for market prices as of December 15, 2006. Note that

Table 5 DJIA Stocks by Share Price in Ascending Order as of December 15, 2006

Company	Share Price
Intel Corp.	20.77
Pfizer Inc.	25.56
General Motors Corp.	29.77
Microsoft Corp.	30.07
Alcoa Inc.	30.76
Walt Disney Co.	34.72
AT&T Inc.	35.66
Verizon Communications Inc.	36.09
General Electric Co.	36.21
Hewlett-Packard Co.	39.91
Home Depot Inc.	39.97
Honeywell International Inc.	42.69
Merck & Co. Inc.	43.60
McDonald's Corp.	43.69
Wal-Mart Stores Inc.	46.52
JPMorgan Chase & Co.	47.95
E.I. DuPont de Nemours & Co.	48.40
Coca-Cola Co.	49.00
Citigroup Inc.	53.11
American Express Co.	61.90
United Technologies Corp.	62.06
Caterpillar Inc.	62.12
Procter & Gamble Co.	63.35
Johnson & Johnson	66.25
American International Group Inc.	72.03
Exxon Mobil Corp.	78.73
3M Co.	78.77
Altria Group Inc.	84.97
Boeing Co.	89.93
International Business Machines Corp.	95.36

Source: www.dj.com/TheCompany/FactSheets.htm, December 15, 2006.

since each value occurs once only, we have to assign each value an absolute frequency of 1 or a relative frequency of $1/30$, respectively, since there are 30 component stocks in the DJIA. We start with the lowest entry (\$20.77) and advance up to the largest value still less than \$50, which is \$49 (Coca-Cola). Each time we observe less than or equal to \$50, we add $1/30$, accounting for the frequency of each company to obtain an accumulated frequency of $18/30$ representing the total share of closing prices below \$50. This accumulated frequency is called the “empirical cumulative frequency” at the value \$50. If one computes this for all values, one obtains the empirical cumulative frequency distribution. The term “empirical” is used because the distribution is computed from observed data.

Formal Presentation of Cumulative Frequency Distributions

Formally, the empirical cumulative frequency distribution F_{emp} is defined as

$$F_{emp}(x) = \sum_{i=1}^k a_i$$

where k is the index of the largest value observed that is still less than x . In our example, k is 18. When we use relative frequencies, we obtain the empirical relative cumulative frequency distribution defined analogously to the empirical cumulative frequency distribution, this time using relative frequencies. Hence, we have

$$F_{emp}^f(x) = \sum_{i=1}^k f_i$$

In our example, $F_{emp}^f(50) = 18/30 = 0.6 = 60\%$.

Note that the empirical cumulative frequency distribution can be evaluated at any real x even though x need not be an observation. For any value x between two successive observations $x_{(i)}$ and $x_{(i+1)}$, the empirical cumulative frequency distribution as well as the empirical cumulative relative frequency distribution remain at their respective levels at $x_{(i)}$; that is, they are of constant level $F_{emp}(x_{(i)})$ and $F_{emp}^f(x_{(i)})$, respectively. For example, consider the empirical relative cumulative frequency distribution for the data shown in Table 5. We can extend the distribution to a function that determines the value of the distribution at each possible value of the share price. The function is given in Table 6. Notice that if no value is observed more than once, then the empirical relative cumulative frequency distribution jumps by $1/N$ at each observed value. In our illustration, the jump size is $1/30$.

In Figure 2 the empirical relative cumulative frequency distribution is shown as a graph. Note that the values of the function are constant on the extended line between two successive observations, indicated by the solid point to the

Table 6 Empirical Relative Cumulative Frequency Distribution of DJIA Stocks from Table 5

$F_{emp}^f(x)$			
0.00		$x <$	20.77
0.03	20.77	$\leq x <$	25.56
0.07	25.56	$\leq x <$	29.77
0.10	29.77	$\leq x <$	30.07
0.13	30.07	$\leq x <$	30.76
0.17	30.76	$\leq x <$	34.72
0.20	34.72	$\leq x <$	35.66
0.23	35.66	$\leq x <$	36.09
0.27	36.09	$\leq x <$	36.21
0.30	36.21	$\leq x <$	39.91
0.33	39.91	$\leq x <$	39.97
0.37	39.97	$\leq x <$	42.69
0.40	42.69	$\leq x <$	43.60
0.43	43.60	$\leq x <$	43.69
0.47	43.69	$\leq x <$	46.52
0.50	46.52	$\leq x <$	47.95
0.53	47.95	$\leq x <$	48.40
0.57	48.40	$\leq x <$	49.00
0.60	49.00	$\leq x <$	53.11
0.63	53.11	$\leq x <$	61.90
0.67	61.90	$\leq x <$	62.06
0.70	62.06	$\leq x <$	62.12
0.73	62.12	$\leq x <$	63.35
0.77	63.35	$\leq x <$	66.25
0.80	66.25	$\leq x <$	72.03
0.83	72.03	$\leq x <$	78.73
0.87	78.73	$\leq x <$	78.77
0.90	78.77	$\leq x <$	84.97
0.93	84.97	$\leq x <$	89.93
0.97	89.93	$\leq x <$	95.36
1.00	95.36	$\leq x$	

left of each horizontal line. At each observation, the vertical distance between the horizontal line extending to the right from the preceding observation and the value of the function is exactly the increment $1/30$.

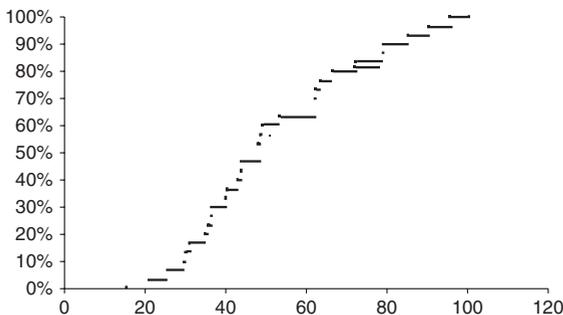


Figure 2 Empirical Relative Cumulative Frequency Distribution of DJIA Stocks from Table 5

The computation of either form of empirical cumulative distribution function is obviously not intuitive for categorical data unless we assign some meaningless numerical proxy to each value such as “Sector A” = 1, “Sector B” = 2, and so on.

DATA CLASSES

Reasons for Classifying

When quantitative variables are such that the set of values—whether observed or theoretically possible—includes intervals or the entire real numbers, then the variable is continuous. This is in contrast to discrete variables, which assume values only from a limited or countable set. Variables on a nominal scale cannot be considered in this context. And because of the difficulties with interpreting the results, we will not attempt to explain the issue of classes for rank data either.

When one counts the frequency of observed values of a continuous variable, one notices that hardly any value occurs more than once. (Naturally, the precision given by the number of digits rounded may result in higher occurrences of certain values.) Theoretically, with 100% chance, all observations will yield different values. Thus, the method of counting the frequency of each value is not feasible. Instead, the continuous set of values is divided into mutually exclusive intervals. Then, for each such interval, the number of values falling within that interval can be counted again. In other words, one groups the data into classes for which the frequencies can be computed. Classes should be such that their respective lower and upper bounds are real numbers. Also, whether the class bounds are elements of the classes or not must be specified. The class bounds of a class must be bounds of the respective adjacent classes as well, such that the classes seamlessly cover the entire data. The width should be the same for all classes. However, if there are areas where the data are very intensely dense in contrast to areas of lesser density, then the

class width can vary according to significant changes in value density. In certain cases, most of the data are relatively evenly scattered within some range, while there are extreme values that are located in isolated areas on either end of the data array. Then, it is sometimes advisable to specify no lower bound to the lowest class and no upper bound to the uppermost class. Classes of this sort are called “open classes.” Moreover, one should consider the precision to which the data are given. If values are rounded to the first decimal but there is the chance that the exact value might vary within half a decimal about the value given, class bounds have to consider this lack of certainty by admitting plus half a decimal on either end of the class.

Formal Procedure of Classifying

Formally, there are four criteria that the classes need to meet:

Criterion 1: *Mutual Exclusiveness*: Each value can be placed in only one class.

Criterion 2: *Completeness*: The set of classes needs to cover all values.

Criterion 3: *Equidistance*: If possible, form classes of equal width.

Criterion 4: *Nonemptiness*: If possible, avoid forming empty classes.

It is intuitive that the number of classes should increase with an increasing range of values and increasing number of data. Though there are no stringent rules, two rules of thumb are given here with respect to the advised number of classes (first rule) and the best class width (second rule). The first, the so-called Sturge’s rule, states that for a given set of continuous data of size n , one should use the nearest integer figure to

$$1 + \log_2 n = 1 + 3.222 \log_{10} n.$$

Here, $\log_a n$ denotes the logarithm of n to the base a , with a being either 2 or 10.

The second guideline is the so-called *Freedman-Diaconis rule* for the appropriate class width or bin size. Before turning to the second

rule of thumb in more detail, we have to introduce the notion of the *inner quartile range (IQR)*. This quantity measures the distance between the value where F_{emp}^f is closest to 0.25 (that is, the so-called 0.25-quantile), and the value where F_{emp}^f is closest to 0.75 (that is, the so-called 0.75-quantile). (The term “percentile” is used interchangeably with “quantile.”) So the IQR range states how remote the lowest 25% of the observations are from the highest 25%.² As a consequence, the IQR comprises the central 50% of a data sample. A little more attention will be given to the determination of the above-mentioned quantiles when we discuss sample moments and quantiles, since formally there might arise some ambiguity when computing them. (Note that the IQR cannot be computed for nominal or categorical data in a natural way.)

Now we can return to the Freedman-Diaconis rule. It states that a good class width is given by the nearest integer to

$$2 \times IQR \times N^{-1/3}$$

where N is the number of observations in the dataset. Note that there is an inverse relationship between the class width and the number of classes for each set of data. That is, given that the partitioning of the values into classes covers all observations, the number of classes n has to be equal to the difference between largest and smallest value divided by the class width, if classes are all of equal size w . Mathematically, that means

$$n = (x_{\max} - x_{\min})/w$$

where x_{\max} denotes the largest value and x_{\min} denotes the smallest value considered, respectively.

One should not be intimidated by all these rules. Generally, by mere ordering of the data in an array, intuition produces quite a good feeling of what the classes should look like. Some thought can be given to the timing of the formation of the classes. That is, when classes are formed prior to the data-gathering process, one does not have to store the specific values but

rather count only the number of hits within each class.

Example of Classing Procedures

Let's illustrate these rules. Table 7 gives the 12-month returns (in percent) of the 235 Franklin Templeton Investments Funds on January 11, 2007. With this many data, it becomes obvious that it cannot be helpful to anyone to know the relative performance for the 235 funds. To obtain an overall impression of the distribution of the data without getting lost in detail, one has to aggregate the information given by classifying the data.

For the sake of a better overview, the ordered array is given in Table 8. A quick glance at the data sorted in ascending order gives us the lowest (minimum) and largest (maximum) return, respectively. Here, we have $x_{\min} = -18.3\%$ and $x_{\max} = 41.3\%$, respectively, yielding a range of 59.6% to cover.

We first classify the data according to Sturge's rule. For the number of classes, n , we obtain the nearest integer to $1 + \log_2 235 = 8.877$, which is 9. The class width is then determined by the range divided by the number of classes, $59.6\%/9$, yielding a width of roughly 6.62%. This is not a nice number to deal with, so we may choose 7% instead without deviating noticeably from the exact numbers given by Sturge's rule. We now cover a range of $9 \times 7\% = 63\%$, which is slightly larger than the original range of the data.

Selecting a value for the lower class bound of the lowest class slightly below our minimum, say -20.0% , and an upper class bound of the highest class, say 43.0% , we spread the surplus of the range (3.4%) evenly. The resulting classes can be viewed in Table 9, where in the first row the index of the respective class is given. The second row contains the class bounds. Brackets indicate that the value belongs to the class, whereas parentheses exclude given values. So, we obtain a half-open interval for each class containing all real numbers between the lower bound and just below the upper bound, thus

excluding that value. In row three, we have the number of observations that fall into the respective classes.

We can check for the compliance with the four criteria given earlier. Because we use half-open intervals, we guarantee that Criterion 1 is fulfilled. Since the lowest class starts at -20% , and the highest class ends at 43% , Criterion 2 is satisfied. All nine classes are of width 7%, which complies with Criterion 3. Finally, the compliance with Criterion 4 can be checked easily.

Next, we apply the Freedman-Diaconis rule. With our ordered array of data, we can determine the 0.25 quartile by selecting the observation whose index is the first to exceed $0.25 \times N = .25 \times 235 = 58.75$. This yields the value of observation 59, which is 4.2%. Accordingly, the 0.75-quartile is given by the value whose index is the first to exceed $0.75 \times 235 = 176.25$. For our return data, it is x_{177} , which is 18.9%. The *IQR* is computed as

$$18.9\% - 4.2\% = 14.7\%$$

such that the bin size of the classes (or class width) is now determined according to $w = 2 \times IQR \times \frac{1}{\sqrt[3]{235}} = 4.764\%$. Taking the data range of 59.6% from the previous calculation, we obtain as the suggested number of classes $59.6\%/4.764 = 12.511$. Once again, this is not a neat-looking figure. We stick with the initial class width of $w = 4.764\%$ as closely as possible by selecting the next integer, say 5%. And, without any loss, we extend the range artificially to 60%. So, we obtain for the number of classes $60\%/5 = 12$, which is close to our original real number, 12.511, computed according to the Freedman-Diaconis rule but much nicer to handle. We again spread the range surplus of 0.4% ($60\% - 59.6\%$) evenly across either end of the range such that we begin our lowest class at -18.5% and end our highest class at 41.5% . The classes are given in Table 10. The first row of the table indicates the index of the respective class, while the second row gives the class bounds. The number of observations that fall into each class is shown in the last row. (One can easily

Table 7 12-Month Returns (in %) for the 235 Franklin Templeton Investment Funds (Luxembourg) on January 11, 2007

Aggr Growth A Acc	1.9	Mut Gb Discov A Acc EUR	8.1	Asian Grth A Dis USD	21.3	Glbl Bd A Dis GBP	-0.7
Aggr Growth A Dis	-6.9	Mut Gb Discov A Acc USD	16.4	Asian Grth C Acc	20.6	Glbl Bd B Dis	7.5
Aggr Growth B Acc	0.9	Mut Gb Discov A Dis GBP	5.9	Asian Grth I Acc EUR	14.0	Glbl Bd C Dis	8.2
Aggr Growth I Acc	2.9	Mut Gb Discov B Acc USD	14.9	Asian Grth I Acc USD	22.6	Glbl Bd I Acc EUR	1.4
Biotech Disc A Acc	-0.4	Mut Gb Discov C Acc USD	15.7	BRIC A Acc EUR	25.3	Glbl Bd I Acc USD	9.9
Biotech Disc B Acc	-1.8	Mut Gb Discov I Acc EUR	9.1	BRIC A Acc USD	34.8	Glbl Bd(Euro) A Acc	1.5
Biotech Disc I Acc	0.6	Mut Gb Discov I Acc USD	17.4	BRIC A Dis GBP	22.7	Glbl Bd(Euro) A Dis	1.4
Europ Growth A Acc	20.0	T Japan A Acc EUR	-18.3	BRIC B Acc USD	33.2	Glbl Bd(Euro) I Acc	2.0
Europ Growth I Acc	21.3	T Japan A Acc JPY	-8.3	BRIC C Acc USD	34.1	Global Euro A Acc	11.1
EurSMidCapGr A Acc	33.1	T Japan A Acc USD	-12.2	BRIC I Acc USD	36.4	Global Euro A Dis	11.1
EurSMidCapGr I Acc	33.3	T Japan C Acc USD	-12.6	China A Acc	36.1	Global Euro I Acc	12.1
EurSMidCapGrBAccUSD	41.3	T Japan I Acc EUR	-17.6	China A Dis	23.7	Global A Acc	20.5
Global Growth A Acc	15.5	T Japan I Acc USD	-11.4	China I Acc	37.6	Global A Dis	20.4
GlblMidCapGr A Acc	10.7	T Glb Gr&Val A Acc	16.7	Eastern Europ A Acc EUR	13.3	Global B Acc	18.9
GlblMidCapGr B Acc	9.3	T Glb Gr&Val B Acc	15.3	Eastern Europ A Acc USD	21.9	Global C Acc	19.7
GlblRealEst A Acc EUR	19.7	T Glb Gr&Val C Acc	16.0	Eastern Europ A Dis EUR	13.3	Global I Acc	21.5
GlblRealEst I Acc EUR	20.7	T Glb Gr&Val I Acc	17.8	Eastern Europ A Dis GBP	11.0	Glb Eq Inc A Acc EUR	11.4
GlblRealEst A Dis GBP	17.1	Technology A Acc	-0.4	Eastern Europ C Acc EUR	12.6	Glb Eq Inc A Acc USD	19.9
GlblRealEst A Acc USD	22.1	Technology B Acc	-1.4	Eastern Europ C Acc USD	21.2	Glb Eq Inc A Dis	19.9
GlblRealEst A Dis USD	22.1	US Eqty A Acc EUR	0.2	Eastern Europ I Acc	14.7	Glb Eq Inc B Dis	18.4
GlblRealEst B Dis USD	20.5	US Eqty A Acc EUR Hdg	4.9	Emg Mkt A Acc	14.4	Glb Eq Inc C Dis	19.3
GlblRealEst C Dis USD	21.4	US Eqty A Acc USD	7.7	Emg Mkt A Dis	14.4	Glb Eq Inc I Acc	20.5
GlblRealEst I Acc USD	23.1	US Eqty B Acc	6.4	Emg Mkt B Acc	13.0	Glb Inc A Acc EUR	10.4
GlblRealEst I Dis USD	23.1	US Eqty C Acc	7.1	Emg Mkt C Acc	13.7	Glb Inc A Acc USD	18.7
High Yield A Acc	6.9	US Eqty I Acc EUR	-5.9	Emg Mkt I Acc	15.8	Glb Inc A Dis	18.7
High Yield A Dis	7.1	US Eqty I Acc USD	8.9	EmMktBd A Dis EUR	5.2	Glb Inc B Dis	17.2
High Yield B Dis	5.6	US Gov A Dis	3.1	EmMktBd A Dis USD	13.2	Glb Inc C Dis	17.9
High Yield C Acc	6.2	US Gov B Dis	1.8	Emg Mkt Bd B Dis	11.9	Glb Inc I Acc	19.4
High Yield I Dis	7.8	US Gov B Acc	1.9	Emg Mkt Bd C Acc	12.6	Glbl Sm Co A Acc	21.3
High Yld Eur A Acc	8.3	US Gov C Acc	2.2	Emg Mkt Bd I Acc	14.3	Glbl Sm Co A Dis	21.3
High Yld Eur A Dis	8.3	US Gov I Dis	3.8	Euro Liq Res A Acc	1.9	Glbl Sm Co C Acc	12.1
High Yld Eur I Acc	9.1	US Growth A Acc	3.8	Euro Liq Res A Dis	1.9	Glbl Sm Co I Acc	22.4
High Yld Eur I Dis	9.1	US Growth B Acc	2.5	Euroland Bd A Dis	-1.8	Dlbl Tot Ret A Acc	12.6
Income A Dis	12.8	US Growth C Acc	3.3	Euroland Bd I Acc	-1.2	Dlbl Tot Ret A Dis	12.6
Income B Dis	11.4	US Growth I Acc	6.4	Euroland A Acc	18.5	Dlbl Tot Ret B Acc	10.9
Income C Acc	12.1	US Ultra Sh Bd A Dis	3.7	Euroland A Dis	19.8	Dlbl Tot Ret B Dis	10.9
Income C Dis	12.1	US Ultra Sh Bd B Acc	2.5	Euroland C Acc	17.8	Dlbl Tot Ret C Dis	11.8
Income I Acc	13.7	US Ultra Sh Bd B Dis	2.5	Euroland I Acc	19.6	Dlbl Tot Ret I Acc	13.1
India A Acc EUR	29.0	US Ultra Sh Bd C Dis	2.6	Europan A Acc USD	24.0	Dlbl Tot Ret I Dis	10.0
India A Acc USD	38.7	US Ultra Sh Bd I Acc	4.2	European A Acc EUR	15.3	Growth(Euro) A Acc	7.5
India A Dis GBP	26.2	US SmMidCapGro A Acc	2.5	European A Dis EUR	15.2	Growth(Euro) A Dis	7.4
India B Acc USD	36.9	US SmMidCapGro B Acc	1.2	European A Dis USD	24.0	Growth(Euro) I Acc	8.4
India C Acc USD	37.9	US SmMidCapGro C Acc	2.0	European C Acc EUR	14.6	Growth(Euro) I Dis	8.4
India I Acc EUR	30.2	US Tot Rtn A Acc	4.1	European I Acc	16.4	Japan A Acc	-8.0
India I Acc USD	40.0	US Tot Rtn A Dis	4.2	Euro Tot Ret A Acc	-0.4	Korea A Acc	-3.8
Mut Beacon AAccEUR	7.4	US Tot Rtn B Acc	2.6	Euro Tot Ret A Dis EUR	-0.5	Latin Amer A Acc	35.9
Mut Beacon AAccUSD	15.5	US Tot Rtn B Dis	2.7	Euro Tot Ret A Dis GBP	-2.6	Latin Amer A Dis GBP	23.6
Mut Beacon ADisUSD	15.5	US Tot Rtn C Dis	3.1	Euro Tot Ret A Dis USD	7.1	Latin Amer A Dis USD	35.9
Mut Beacon Bacc	14.0	US Tot Rtn I Acc	4.8	Euro Tot Ret C Acc EUR	-1.3	Latin Amer I Acc USD	37.4
Mut Beacon Cacc	14.8	Asian Bond A Acc EUR	5.9	Euro Tot Ret C Dis USD	6.2	Thailand A Acc	-11.0
Mut Beacon IAcc	16.6	Asian Bond A Acc USD	14.1	Euro Tot Ret I Acc	-0.3	US\$ Liq Res A Acc	4.2
Mut Europ AAcc EUR	15.9	Asian Bond A Dis USD	14.0	Glbl Bal A Acc EUR	6.5	US\$ Liq Res A Dis	4.1
Mut Europ AAcc USD	24.7	Asian Bond B Dis USD	12.4	Glbl Bal A Acc USD	14.6	US\$ Liq Res B Dis	3.1
Mut Europ ADis EUR	15.9	Asian Bond C Dis USD	13.0	Glbl Bal A Dis	14.6	US\$ Liq Res C Acc	3.2
Mut Europ ADis GBP	14.0	Asian Bond I Acc USD	14.6	Glbl Bal B Acc	13.1	US Value A Acc	14.5
Mut Europ B Acc	23.1	Asian Grth A Acc EUR	12.7	Glbl Bal C Dis	13.9	US Value B Acc	13.0
Mut Europ C Acc USD	23.9	Asian Grth A Acc USD	21.4	Glbl Bd A Dis USD	9.2	US Value C Acc	13.8
Mut Europ C Acc EUR	15.2	Asian Grth A Dis EUR	12.8	Glbl Bd A Acc EUR	1.5	US Value I Acc	15.6
Mut Europ I Acc	16.9	Asian Grth A Dis GBP	10.4	Glbl Bd A Dis EUR	1.5		

Table 8 Ordered Array of the 235 12-month Returns for the Franklin Templeton Investment Funds (Luxembourg)

Obs. (<i>i</i>)	Value
x(1)	-18.3
x(2)	-17.6
x(3)	-12.6
x(4)	-12.2
x(5)	-11.4
x(6)	-11
x(7)	-8.3
x(8)	-8
x(9)	-6.9
x(10)	-5.9
x(11)	-3.8
x(12)	-2.6
x(13)	-1.8
x(14)	-1.8
x(15)	-1.4
x(16)	-1.3
x(17)	-1.2
x(18)	-0.7
x(19)	-0.5
x(20)	-0.4
x(21)	-0.4
x(22)	-0.4
x(23)	-0.3
x(24)	0.2
x(25)	0.6
x(26)	0.9
x(27)	1.2
x(28)	1.4
x(29)	1.4
x(30)	1.5
x(31)	1.5
x(32)	1.5
x(33)	1.8
x(34)	1.9
x(35)	1.9
x(36)	1.9
x(37)	1.9
x(38)	2
x(39)	2
x(40)	2.2
x(41)	2.5
x(42)	2.5
x(43)	2.5
x(44)	2.5
x(45)	2.6
x(46)	2.6
x(47)	2.7
x(48)	2.9
x(49)	3.1
x(50)	3.1
x(51)	3.1
x(52)	3.2
x(53)	3.3
x(54)	3.7
x(55)	3.8
x(56)	3.8
x(57)	4.1
x(58)	4.1
x(59)	4.2
x(60)	4.2
x(61)	4.2
x(62)	4.8
x(63)	4.9
x(64)	5.2
x(65)	5.6
x(66)	5.9
x(67)	5.9
x(68)	6.2
x(69)	6.2
x(70)	6.4
x(71)	6.4
x(72)	6.5
x(73)	6.9
x(74)	7.1
x(75)	7.1
x(76)	7.1
x(77)	7.4
x(78)	7.4
x(79)	7.5
x(80)	7.5
x(81)	7.7
x(82)	7.8
x(83)	8.1
x(84)	8.2
x(85)	8.3
x(86)	8.3
x(87)	8.4
x(88)	8.4
x(89)	8.9
x(90)	9.1
x(91)	9.1
x(92)	9.1
x(93)	9.2
x(94)	9.3
x(95)	9.9
x(96)	10
x(97)	10.4
x(98)	10.4
x(99)	10.7
x(100)	10.9
x(101)	10.9
x(102)	11
x(103)	11.1
x(104)	11.1
x(105)	11.4
x(106)	11.4
x(107)	11.8
x(108)	11.9
x(109)	12.1
x(110)	12.1
x(111)	12.1
x(112)	12.1
x(113)	12.4
x(114)	12.6
x(115)	12.6
x(116)	12.6
x(117)	12.6
x(118)	12.7
x(119)	12.8
x(120)	12.8
x(121)	13
x(122)	13
x(123)	13
x(124)	13.1
x(125)	13.1
x(126)	13.2
x(127)	13.3
x(128)	13.3
x(129)	13.7
x(130)	13.7
x(131)	13.8
x(132)	13.9
x(133)	14
x(134)	14
x(135)	14
x(136)	14
x(137)	14.1
x(138)	14.3
x(139)	14.4
x(140)	14.4
x(141)	14.5
x(142)	14.6
x(143)	14.6
x(144)	14.6
x(145)	14.6
x(146)	14.7
x(147)	14.8
x(148)	14.9
x(149)	15.2
x(150)	15.2
x(151)	15.3
x(152)	15.3
x(153)	15.5
x(154)	15.5
x(155)	15.5
x(156)	15.6
x(157)	15.7
x(158)	15.8
x(159)	15.9
x(160)	15.9
x(161)	16
x(162)	16.4
x(163)	16.4
x(164)	16.6
x(165)	16.7
x(166)	16.9
x(167)	17.1
x(168)	17.2
x(169)	17.4
x(170)	17.8
x(171)	17.8
x(172)	17.9
x(173)	18.4
x(174)	18.5
x(175)	18.7
x(176)	18.7
x(177)	18.9
x(178)	19.3
x(179)	19.4
x(180)	19.6
x(181)	19.7
x(182)	19.7
x(183)	19.8
x(184)	19.9
x(185)	19.9
x(186)	20
x(187)	20.4
x(188)	20.5
x(189)	20.5
x(190)	20.5
x(191)	20.6
x(192)	20.7
x(193)	21.2
x(194)	21.3
x(195)	21.3
x(196)	21.3
x(197)	21.3
x(198)	21.4
x(199)	21.4
x(200)	21.5
x(201)	21.9
x(202)	22.1
x(203)	22.1
x(204)	22.4
x(205)	22.6
x(206)	22.7
x(207)	23.1
x(208)	23.1
x(209)	23.1
x(210)	23.6
x(211)	23.7
x(212)	23.9
x(213)	24
x(214)	24
x(215)	24.7
x(216)	25.3
x(217)	26.2
x(218)	29
x(219)	30.2
x(220)	33.1
x(221)	33.2
x(222)	33.3
x(223)	34.1
x(224)	34.8
x(225)	35.9
x(226)	35.9
x(227)	36.1
x(228)	36.4
x(229)	36.9
x(230)	37.4
x(231)	37.6
x(232)	37.9
x(233)	38.7
x(234)	40
x(235)	41.3

check that the four requirements for the classes are met again.)

Let us next compare Tables 9 and 10. We observe a finer distribution when the Freedman-Diaconis rule is employed because this rule generates more classes for the same data. However, it is generally difficult to judge which rule provides us with the better information because, as is seen, the two rules set up completely different classes. But the choice of class bounds is essential. By just slightly shifting the

bounds between two adjacent classes, many observations may fall from one class into the other due to this alteration. As a result, this might produce a totally different picture about the data distribution. So, we have to be very careful when we interpret the two different results.

For example, class 7, that is, [22,29) in Table 9 contains 16 observations. Classes 9 and 10 of Table 10 cover approximately the same range, [21.5,31.5). Together they account for 20 observations. We could now easily present two

Table 9 Classes for the 235 Fund Returns According to Sturge's Rule

Class Index									
<i>I</i>	1	2	3	4	5	6	7	8	9
$[a_i, b_i)$	[-20, -13)	[-13, -6)	[-6, 1)	[1, 8)	[8, 15)	[15, 22)	[22, 29)	[29, 36)	[36, 43)
a_i	2	7	17	56	66	53	16	9	9

scenarios that would provide rather different conceptions about the frequency. In scenario one, suppose one assumes that two observations are between 21.5 and 22.0. Then, there would have to be 16 observations between 22.0 and 26.5 to add up to 18 observations in class 9 of Table 10. This, in return, would mean that the 16 observations of class 7 from Table 9 would all have to lie between 22.0 and 26.5 as well. Then, the two observations from class 10 of Table 10 must lie beyond 29.0. The other scenario could assume that we have four observations between 21.5 and 22.0. Then, for similar reasons as before, we would have 14 observations between 22.0 and 26.5. The two observations from class 10 of Table 10 would now have to be between 26.5 and 29.0, so that the total of 16 observations in class 7 of Table 9 is met. See how easily slightly different classes can lead to ambiguous interpretation? Looking at all classes at once, many of these puzzles can be solved. However, some uncertainty remains. As can be seen, the choice of the number of classes and thus the class bounds can have a significant impact on the information that the data conveys when condensed into classes.

CUMULATIVE FREQUENCY DISTRIBUTIONS

In contrast to the empirical cumulative frequency distributions, in this section we will introduce functions that convey basically the same information, that is, the *frequency distribu-*

tion, but rely on a few more assumptions. These cumulative frequency distributions introduced here, however, should not be confused with the theoretical definitions given in probability theory even though the notion is akin to both.

The absolute cumulative frequency at each class bound states how many observations have been counted up to this particular class bound. However, we do not exactly know how the data are distributed within the classes. When relative frequencies are used, though, the cumulative relative frequency distribution states the overall proportion of all values up to a certain lower or upper bound of some class.

So far, things are not much different from the definition of the empirical cumulative frequency distribution and *empirical cumulative relative frequency distribution*. At each bound, the empirical cumulative frequency distribution and cumulative frequency coincide. However, an additional assumption is made regarding the distribution of the values between bounds of each class when computing the cumulative frequency distribution. The data are thought of as being continuously distributed and equally spread between the particular bounds. (This type of assumed behavior is defined as a "uniform distribution of data.") Hence, both forms (absolute and relative) of the cumulative frequency distributions increase in a linear fashion between the two class bounds. So, for both forms of cumulative distribution functions, one can compute the accumulated frequencies at values inside of classes.

Table 10 Classes for the 235 Fund Returns According to the Freedman-Diaconis Rule

<i>I</i>	1	2	3	4	5	6	7	8	9	10	11	12
$[a_i, b_i)$	[-18.5; -13.5)	[-13.5; -8.5)	[-8.5; -3.5)	[-3.5; 1.5)	[1.5; 6.5)	[6.5; 11.5)	[11.5; 16.5)	[16.5; 21.5)	[21.5; 26.5)	[26.5; 31.5)	[31.5; 36.5)	[36.5; 41.5)
a_i	2	4	5	18	42	35	57	36	18	2	9	7

For a more thorough summary of this, let's use a more formal presentation. Let I denote the set of all class index i with i being some integer value between 1 and $n_I = |I|$ (that is, the number of classes). Moreover, let a_j and f_j denote the (absolute) frequency and relative frequency of some class j , respectively. The cumulative frequency distribution at some upper bound, x_u^i , of a given class i is computed as

$$F(x_u^i) = \sum_{j:x_j^l \leq x_u^i} a_j = \sum_{j:x_j^l \leq x_u^i} a_j + a_i \quad (1)$$

In words, this means that we sum up the frequencies of all classes whose upper bound is less than x_u^i plus the frequency of class i itself. The corresponding cumulative relative frequency distribution at the same value is then

$$F^f(x_u^i) = \sum_{j:x_j^l \leq x_u^i} f_j = \sum_{j:x_j^l \leq x_u^i} f_j + f_i \quad (2)$$

This describes the same procedure as in equation (1) using relative frequencies instead of frequencies. For any value x in between the boundaries of, say, class i , x_l^i and x_u^i , the cumulative relative frequency distribution is defined by

$$F^f(x) = F^f(x_l^i) + \frac{x - x_l^i}{x_u^i - x_l^i} f_i \quad (3)$$

In words, this means that we compute the cumulative relative frequency distribution at value x as the sum of two things. First, we take the cumulative relative frequency distribution at the lower bound of class i . Second, we add that share of the relative frequency of class i that is determined by the part of the whole interval of class i that is covered by x .

Figure 3 might appeal more to intuition. At the bounds of class i , we have values of the cumulative relative frequency given by $F^f(x_l^i)$ and $F^f(x_u^i)$ respectively. We assume that the cumulative relative frequency increases linearly along the line connecting $F^f(x_l^i)$ and $F^f(x_u^i)$. Then, at any value x^* inside of class i , we find the corresponding value $F^f(x^*)$ by the intersection of the dashed line and the vertical axis as shown. The dashed line is obtained by ex-

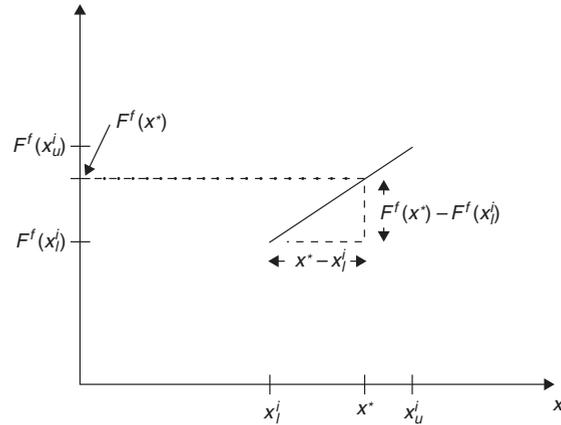


Figure 3 Determination of Frequency Distribution within Class Bounds

tending a horizontal line through the intersection of the vertical line through x^* and the line connecting $F^f(x_l^i)$ and $F^f(x_u^i)$ with slope $F^f(x^*) - F^f(x_l^i)/x^* - x_l^i$.

KEY POINTS

- The field of descriptive statistics discerns different types of data. Very generally, there are two types: qualitative and quantitative data. If certain attributes of an item can only be assigned to categories, these data are referred to as qualitative. However, if an item is assigned a quantitative variable, the value of this variable is numerical. Generally, all real numbers are eligible.
- In descriptive statistics, data are grouped according to measurement levels. The measurement level gives an indication as to the sophistication of the analysis techniques that can be applied to the data collected. Typically, a hierarchy with five levels of measurement—nominal, ordinal, interval, ratio, and absolute data—are used to group data. The latter three form the set of quantitative data. If the data are of a certain measurement level, they are said to be scaled accordingly. That is, the data are referred to as nominally scaled, and so on.
- Another way of classifying data is in terms of cross-sectional and time series data.

Cross-sectional data are values of a particular variable across some universe of items observed at a unique point in time. Time series data are data related to a variable successively observed at a sequence of points in time.

- Frequency (absolute and relative) distributions can be computed for all types of data since they do not require that the data have a numerical value. The cumulative frequency distribution is another quantity of interest for comparing data that is closely related to the absolute or relative frequency distribution.
- Four criteria that data classes need to satisfy are (1) each value can be placed in only one class (mutual exclusiveness), (2) the set of classes needs to cover all values (completeness), (3) if possible, form classes of equal width (equidistance), and

(4) if possible, avoid forming empty classes (nonemptiness).

NOTES

1. For a more detailed discussion, see Rachev et al. (2010).
2. The 0.75-quantile divides the data into the lowest 75% and the highest 25%.

REFERENCES

- Rachev, S. T., Hoechstetter, S., Fabozzi, F. J., and Focardi, S. M. (2010). *Probability and Statistics for Finance*. Hoboken, NJ: John Wiley & Sons.
- Rachev, S. T., Mittnik, S., Fabozzi, F. J., Focardi, S. M., and Jasic, R. (2007). *Financial Econometrics: From Basics to Advanced Modeling Techniques*. Hoboken, NJ: John Wiley & Sons.

Time Series Concepts, Representations, and Models

SERGIO M. FOCARDI, PhD
Partner, The Intertek Group

FRANK J. FABOZZI, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

Abstract: A stochastic process is a time-dependent random variable. Stochastic processes such as Brownian motion and Ito processes develop in continuous time. This means that time is a real variable that can assume any real value. In many financial modeling applications, however, it is convenient to constrain time to assume only discrete values. A time series is a discrete-time stochastic process; that is, it is a collection of random variables X_t indexed with the integers $\dots, -n, \dots, -2, -1, 0, 1, 2, \dots, n, \dots$

In this entry, we introduce models of *discrete-time stochastic processes* (that is, time series). In financial modeling, both continuous-time and discrete-time models are used. In many instances, continuous-time models allow simpler and more concise expressions as well as more general conclusions, though at the expense of conceptual complication. For instance, in the limit of continuous time, apparently simple processes such as white noise cannot be meaningfully defined. The mathematics of asset management tends to prefer discrete-time processes, while the mathematics of derivatives tends to prefer continuous-time processes.

The first issue to address in financial econometrics is the spacing of discrete points of time. An obvious choice is regular, constant spacing. In this case, the time points are placed at multiples of a single time interval: $t = i \Delta t$. For

instance, one might consider the closing prices at the end of each day. The use of fixed spacing is appropriate in many applications. Spacing of time points might also be irregular but deterministic. For instance, weekends introduce irregular spacing in a sequence of daily closing prices. These questions can be easily handled within the context of discrete-time series.

In this entry, we discuss only time series at discrete and fixed intervals of time, introducing concepts, representations, and models of time series.¹

CONCEPTS OF TIME SERIES

A time series is a collection of random variables X_t indexed with a discrete time index $t = \dots, -2, -1, 0, 1, 2, \dots$. The variables X_t are defined over a

probability space $(\Omega, P, \mathfrak{J})$, where Ω is the set of states, P is a probability measure, and \mathfrak{J} is the σ -algebra of events, equipped with a discrete *filtration* $\{\mathfrak{J}_t\}$ that determines the propagation of information (see the Appendix). A realization of a time series is a countable sequence of real numbers, one for each time point.

The variables X_t are characterized by *finite-dimensional distributions* as well as by conditional distributions, $F_s(x_s/\mathfrak{J}_t)$, $s > t$. The latter are the distributions of the variable x at time s given the σ -algebra $\{\mathfrak{J}_t\}$ at time t . Note that conditioning is always conditioning with respect to a σ -algebra though we will not always strictly use this notation and will condition with respect to the value of variables, for instance:

$$F_s(x_s/x_t), \quad s > t$$

If the series starts from a given point, initial conditions must be fixed. Initial conditions might be a set of fixed values or a set of random variables. If the initial conditions are not fixed values but random variables, one has to consider the correlation between the initial values and the random shocks of the series. A usual assumption is that the initial conditions and the random shocks of the series are statistically independent.

How do we describe a time series? One way to describe a time series is to determine the mathematical form of the conditional distribution. This description is called an *autopredictive model* because the model predicts future values of the series from past values. However, we can also describe a time series as a function of another time series. This is called an explanatory model as one variable is explained by another. The simplest example is a regression model where a variable is proportional to another exogenously given variable plus a constant term. Time series can also be described as random fluctuations or adjustments around a deterministic path. These models are called *adjustment models*. Explanatory, autopredictive, and adjustment models can be mixed in a single model. The *data generation process* (DGP) of a series is a mathemat-

ical process that computes the future values of the variables given all information known at time t .

An important concept is that of a *stationary time series*. A series is stationary in the “strict sense” if all finite dimensional distributions are invariant with respect to a time shift. A series is stationary in a “weak sense” if only the moments up to a given order are invariant with respect to a time shift. In this entry, time series will be considered (weakly) stationary if the first two moments are time-independent. Note that a stationary series cannot have a starting point but must extend over the entire infinite time axis. Note also that a series can be strictly stationary (i.e., have all distributions time-independent, but the moments might not exist). Thus a strictly stationary series is not necessarily weakly stationary.

A time series can be univariate or multivariate. A *multivariate time series* is a time-dependent random vector. The principles of modeling remain the same but the problem of estimation might become very difficult given the large numbers of parameters to be estimated.

Models of time series are essential building blocks for financial forecasting and, therefore, for financial decision-making. In particular asset allocation and portfolio optimization, when performed quantitatively, are based on some model of financial prices and returns. This entry lays down the basic financial econometric theory for financial forecasting. We will introduce a number of specific models of time series and of multivariate time series, presenting the basic facts about the theory of these processes. We will consider primarily models of financial assets, though most theoretical considerations apply to macroeconomic variables as well. These models include:

- **Correlated random walks.** The simplest model of multiple financial assets is that of *correlated random walks*. This model is only a rough approximation of equity price processes and presents serious problems of

estimation in the case of a large number of processes.

- **Factor models.** *Factor models* address the problem of estimation in the case of a large number of processes. In a factor model there are correlations only among factors and between each factor and each time series. Factors might be exogenous or endogenously modeled.
- **Cointegrated models.** In a *cointegrated model* there are portfolios that are described by autocorrelated, stationary processes. All processes are linear combinations of common trends that are represented by the factors.

The above models are all linear. However, nonlinearities are at work in financial time series. One way to model nonlinearities is to break down models into two components, the first being a linear autoregressive model of the parameters, the second a regressive or autoregressive model of empirical quantities whose parameters are driven by the first. This is the case with most of today's nonlinear models (e.g., ARCH/GARCH models), Hamilton models, and Markov switching models.

There is a coherent modeling landscape, from correlated random walks and factor models to the modeling of factors, and, finally, the modeling of nonlinearities by making the model parameters vary. Before describing models in detail, however, let's present some key empirical facts about financial time series.

STYLIZED FACTS OF FINANCIAL TIME SERIES

Most sciences are stratified in the sense that theories are organized on different levels. The empirical evidence that supports a theory is generally formulated in a lower level theory. In physics, for instance, quantum mechanics cannot be formulated as a stand-alone theory but needs classical physics to give meaning to measurement. Economics is no exception. A basic level of knowledge in economics is represented

by the so-called stylized facts. Stylized facts are statistical findings of a general nature on financial and economic time series; they cannot be considered raw data insofar as they are formulated as statistical hypotheses. On the other hand, they are not full-fledged theories.

Among the most important stylized facts from the point of view of finance theory, we can mention the following:

- Returns of individual stocks exhibit nearly zero autocorrelation at every lag.
- Returns of some equity portfolios exhibit significant autocorrelation.
- The volatility of returns exhibits hyperbolic decay with significant autocorrelation.
- The distribution of stock returns is not normal. The exact shape is difficult to ascertain but power law decay cannot be rejected.
- There are large stock price drops (that is, market crashes) that seem to be outliers with respect to both normal distributions and power law distributions.
- Stock return time series exhibit significant cross-correlation.

These findings are, in a sense, model-dependent. For instance, the distribution of returns, a subject that has received a lot of attention, can be fitted by different distributions. There is no firm evidence on the exact value of the power exponent, with alternative proposals based on variable exponents. The autocorrelation is model-dependent while the exponential decay of return autocorrelation can be interpreted only as absence of linear dependence.

It is fair to say that these stylized facts set the stage for financial modeling but leave ample room for model selection. Financial time series seem to be nearly random processes that exhibit significant cross correlations and, in some instances, cross autocorrelations. The global structure of auto and cross correlations, if it exists at all, must be fairly complex and there is no immediate evidence that financial time series admit a simple DGP.

One more important feature of financial time series is the presence of trends. *Prima facie* trends of economic and financial variables are exponential trends. Trends are not quantities that can be independently measured. Trends characterize an entire stochastic model. Therefore there is no way to arrive at an assessment of trends independent from the model. Exponential trends are a reasonable first approximation.

Given the finite nature of world resources, exponential trends are not sustainable in the long run. However, they might still be a good approximation over limited time horizons. An additional insight into financial time series comes from the consideration of investors' behavior. If investors are risk averse, as required by the theory of investment, then price processes must exhibit a trade-off between risk and returns. The combination of this insight with the assumption of exponential trends yields market models with possibly diverging exponential trends for prices and market capitalization.

Again, diverging exponential trends are difficult to justify in the long run as they would imply that after a while only one entity would dominate the entire market. Some form of reversion to the mean or more disruptive phenomena that prevent time series to diverge exponentially must be at work.

In the following sections we will proceed to describe the theory and the estimation procedures of a number of market models that have been proposed. We will present the multivariate random walk model, introduce cointegration and autoregressive models.

INFINITE MOVING-AVERAGE AND AUTOREGRESSIVE REPRESENTATION OF TIME SERIES

There are several general representations (or models) of time series. This section introduces

representations based on infinite moving averages or infinite autoregressions useful from a theoretical point of view. In the practice of econometrics, however, more parsimonious models such as the ARMA models (described in the next section) are used. Representations are different for stationary and nonstationary time series. Let's start with univariate stationary time series.

Univariate Stationary Series

The most fundamental model of a univariate stationary time series is the infinite moving average of a white noise process. In fact, it can be demonstrated that under mild regularity conditions, any univariate stationary causal time series admits the following *infinite moving-average representation*:

$$x_t = \sum_{i=0}^{\infty} h_i \varepsilon_{t-i} + m$$

where the h_i are coefficients and ε_{t-i} is a one-dimensional zero-mean white-noise process. This is a *causal time series* as the present value of the series depends only on the present and past values of the noise process. A more general infinite moving-average representation would involve a summation that extends from $-\infty$ to $+\infty$. Because this representation would not make sense from an economic point of view, we will restrict ourselves only to causal time series.

A sufficient condition for the above series to be stationary is that the coefficients h_i are absolutely summable:

$$\sum_{i=0}^{\infty} |h_i| < \infty$$

The Lag Operator L

Let's now simplify the notation by introducing the lag operator L . The lag operator L is an operator that acts on an infinite series and produces another infinite series shifted one place to the

left. In other words, the lag operator replaces every element of a series with the one delayed by one time lag:

$$L(x_t) = x_{t-1}$$

The n -th power of the lag operator shifts a series by n places:

$$L^n(x_t) = x_{t-n}$$

Negative powers of the lag operator yield the forward operator F , which shifts places to the right. The lag operator can be multiplied by a scalar and different powers can be added. In this way, linear functions of different powers of the lag operator can be formed as follows:

$$A(L) = \sum_{i=1}^N a_i L^i$$

Note that if the lag operator is applied to a series that starts from a given point, initial conditions must be specified.

Within the domain of stationary series, infinite power series of the lag operator can also be formed. In fact, given a stationary series, if the coefficients h_i are absolutely summable, the series

$$\sum_{i=0}^{\infty} h_i L^i x_t$$

is well defined in the sense that it converges and defines another stationary series. It therefore makes sense to define the operator:

$$A(L) = \sum_{i=0}^{\infty} h_i L^i$$

Now consider the operator $I - \lambda L$. If $|\lambda| < 1$, this operator can be **inverted** and its inverse is given by the infinite power series,

$$(I - \lambda L)^{-1} = \sum_{i=0}^{\infty} \lambda^i L^i$$

as can be seen by multiplying $I - \lambda L$ by the power series $\sum_{i=1}^{\infty} \lambda^i L^i$:

$$(I - \lambda L) \sum_{i=1}^{\infty} \lambda^i L^i = L^0 = I$$

On the basis of this relationship, it can be demonstrated that any operator of the type

$$A(L) = \sum_{i=1}^N a_i L^i$$

can be inverted provided that the solutions of the equation

$$\sum_{i=1}^N a_i z^i = 0$$

have absolute values strictly greater than 1. The inverse operator is an infinite power series

$$A^{-1}(L) = \sum_{i=0}^{\infty} \psi_i L^i$$

Given two linear functions of the operator L , it is possible to define their product

$$A(L) = \sum_{i=1}^M a_i L^i$$

$$B(L) = \sum_{j=1}^N b_j L^j$$

$$P(L) = A(L)B(L) = \sum_{i=1}^{M+N} p_i L^i$$

$$p_i = \sum_{r=1}^i a_r b_{i-r}$$

The convolution product of two infinite series in the lag operator is defined in a similar way

$$A(L) = \sum_{i=0}^{\infty} a_i L^i$$

$$B(L) = \sum_{j=0}^{\infty} b_j L^j$$

$$C(L) = A(L) \times B(L) = \sum_{k=0}^{\infty} c_k L^k$$

$$c_k = \sum_{s=0}^k a_s b_{k-s}$$

We can define the left-inverse (right-inverse) of an infinite series as the operator $A^{-1}(L)$, such that $A^{-1}(L) \times A(L) = I$. The inverse can always be computed solving an infinite set of recursive equations provided that $a_0 \neq 0$. However, the inverse series will not necessarily be stationary. A sufficient condition for stationarity is that the coefficients of the inverse series are absolutely summable.

In general, it is possible to perform on the symbolic series

$$H(L) = \sum_{i=1}^{\infty} h_i L^i$$

the same operations that can be performed on the series

$$H(z) = \sum_{i=1}^{\infty} h_i z^i$$

with z complex variable. However operations performed on a series of lag operators neither assume nor entail convergence properties. In fact, one can think of z simply as a symbol. In particular, the inverse does not necessarily exhibit absolutely summable coefficients.

Stationary Univariate Moving Average

Using the lag operator L notation, the infinite moving-average representation can be written as follows:

$$x_t = \left(\sum_{i=0}^{\infty} h_i L^i \right) \varepsilon_t + m = H(L)\varepsilon_t + m$$

Consider now the inverse series:

$$\Pi(L) = \sum_{i=0}^{\infty} \lambda_i L^i, \quad \Pi(L)H(L) = I$$

If the coefficients λ_i are absolutely summable, we can write

$$\varepsilon_t = \Pi(L)x_t = \sum_{i=0}^{\infty} \lambda_i L^i x_{t-i}$$

and the series is said to be invertible.

Multivariate Stationary Series

The concepts of infinite moving-average representation and of invertibility defined above for univariate series carry over immediately to the multivariate case. In fact, it can be demonstrated that under mild regularity conditions, any multivariate stationary causal time series admits the following infinite moving-average representation:

$$\mathbf{x}_t = \sum_{i=0}^{\infty} \mathbf{H}_i \varepsilon_{t-i} + \mathbf{m}$$

where the \mathbf{H}_i are $n \times n$ matrices, ε_t is an n -dimensional, zero-mean, white noise process with nonsingular variance-covariance matrix Ω , and \mathbf{m} is an n -vector of constants. The coefficients \mathbf{H}_i are called Markov coefficients. This moving-average representation is called the Wold representation. Wold representation states that any series where only the past influences the present can be represented as an infinite moving average of white noise terms. Note that, as in the univariate case, the infinite moving-average representation can be written in more general terms as a sum that extends from $-\infty$ to $+\infty$. However, a series of this type is not suitable for financial modeling as it is not causal (that is, the future influences the present). Therefore we consider only moving averages that extend to past terms.

Suppose that the Markov coefficients are an absolutely summable series:

$$\sum_{i=0}^{\infty} \|\mathbf{H}_i\| < +\infty$$

where $\|\mathbf{H}\|^2$ indicates the largest eigenvalue of the matrix $\mathbf{H}\mathbf{H}'$. Under this assumption, it can be demonstrated that the series is stationary and that the (time-invariant) first two moments can be computed in the following way:

$$\text{cov}(\mathbf{x}_t, \mathbf{x}_{t-h}) = \sum_{i=0}^{\infty} \mathbf{H}_i \Omega \mathbf{H}'_{i-h}$$

$$E[\mathbf{x}_t] = \mathbf{m}$$

with the convention $\mathbf{H}_i = 0$ if $i < 0$. Note that the assumption that the Markov coefficients are an absolutely summable series is essential, otherwise the covariance matrix would not exist. For instance, if the \mathbf{H}_i were identity matrices, the variances of the series would become infinite.

As the second moments are all constants, the series is weakly stationary. We can write the time-independent autocovariance function of the series, which is an $n \times n$ matrix whose entries are a function of the lag h , as

$$\Gamma_x(h) = \sum_{i=0}^{\infty} \mathbf{H}_i \boldsymbol{\Omega} \mathbf{H}'_{i-h}$$

Under the assumption that the Markov coefficients are an absolutely summable series, we can use the lag-operator L representation and write the operator

$$\mathbf{H}(L) = \sum_{i=0}^{\infty} \mathbf{H}_i L^i$$

so that the Wold representation of a series can be written as

$$\mathbf{x}_t = \mathbf{H}(L)\boldsymbol{\varepsilon} + \mathbf{m}$$

The concept of invertibility carries over to the multivariate case. A multivariate stationary time series is said to be invertible if it can be represented in autoregressive form. Invertibility means that the white noise process can be recovered as a function of the series. In order to explain the notion of invertible processes, it is useful to introduce the generating function of the operator \mathbf{H} , defined as the following matrix power series:

$$\mathbf{H}(z) = \sum_{i=0}^{\infty} \mathbf{H}_i z^i$$

It can be demonstrated that, if $\mathbf{H}_0 = \mathbf{I}$, then $\mathbf{H}(0) = \mathbf{H}_0$ and the power series $\mathbf{H}(z)$ is invertible in the sense that it is possible to formally derive the inverse series,

$$\boldsymbol{\Pi}(z) = \sum_{i=0}^{\infty} \boldsymbol{\Pi}_i z^i$$

such that

$$\boldsymbol{\Pi}(z)\mathbf{H}(z) = (\boldsymbol{\Pi} \times \mathbf{H})(z) = \mathbf{I}$$

where the product is intended as a convolution product. If the coefficients $\boldsymbol{\Pi}_i$ are absolutely summable, as the process \mathbf{x}_t is assumed to be stationary, it can be represented in infinite autoregressive form:

$$\boldsymbol{\Pi}(L)(\mathbf{x}_t - \mathbf{m}) = \boldsymbol{\varepsilon}_t$$

In this case the process \mathbf{x}_t is said to be invertible.

From the above, it is clear that the infinite moving average representation is a more general linear representation of a stationary time than the infinite autoregressive form. A process that admits both representations is called invertible.

Nonstationary Series

Let's now look at nonstationary series. As there is no very general model of nonstationary time series valid for all nonstationary series, we have to restrict somehow the family of admissible models. Let's consider a family of *linear, moving-average, nonstationary models* of the following type:

$$\mathbf{x}_t = \sum_{i=0}^t \mathbf{H}_i \boldsymbol{\varepsilon}_{t-i} + \mathbf{h}(t)\mathbf{z}_{-1}$$

where the \mathbf{H}_i are left unrestricted and do not necessarily form an absolutely summable series, $\mathbf{h}(t)$ is deterministic, and \mathbf{z}_{-1} is a random vector called the initial conditions, which is supposed to be uncorrelated with the white noise process. The essential differences of this linear model with respect to the Wold representation of stationary series are:

- The presence of a starting point and of initial conditions.
- The absence of restrictions on the coefficients.
- The index t , which restricts the number of summands.

The first two moments of a linear process are not constant. They can be computed in a

way similar to the infinite moving average case:

$$\text{cov}(\mathbf{x}_t \mathbf{x}_{t-h}') = \sum_{i=0}^t \mathbf{H}_i \boldsymbol{\Omega} \mathbf{H}_{i-h}' + \mathbf{h}(t) \text{var}(\mathbf{z}) \mathbf{h}'$$

$$E[\mathbf{x}_t] = \mathbf{m}_t = \mathbf{h}(t) E[\mathbf{z}]$$

Let's now see how a linear process can be expressed in autoregressive form. To simplify notation let's introduce the processes $\tilde{\boldsymbol{\varepsilon}}_t$ and $\tilde{\mathbf{x}}_t$ and the deterministic series $\tilde{\mathbf{h}}(t)$ defined as follows:

$$\tilde{\boldsymbol{\varepsilon}}_t = \begin{cases} \boldsymbol{\varepsilon}_t & \text{if } t > 0 \\ 0 & \text{if } t < 0 \end{cases} \quad \tilde{\mathbf{x}}_t = \begin{cases} \mathbf{x}_t & \text{if } t > 0 \\ 0 & \text{if } t < 0 \end{cases}$$

$$\tilde{\mathbf{h}}_t = \begin{cases} \mathbf{h}_t & \text{if } t > 0 \\ 0 & \text{if } t < 0 \end{cases}$$

It can be demonstrated that, due to the initial conditions, a linear process always satisfies the following autoregressive equation:

$$\boldsymbol{\Pi}(L) \tilde{\mathbf{x}}_t = \boldsymbol{\varepsilon}_t + \boldsymbol{\Pi}(L) \tilde{\mathbf{h}}_t \times (t) \mathbf{z}_{-1}$$

A random walk model

$$x_t = x_{t-1} + \varepsilon_t = \varepsilon_t + \sum_{i=1}^t \varepsilon_{t-i}$$

is an example of a linear nonstationary model.

The above linear model can also represent processes that are nearly stationary in the sense that they start from initial conditions but then converge to a stationary process. A process that converges to a stationary process is called asymptotically stationary.

We can summarize the previous discussion as follows. Under mild regularity conditions, any causal stationary series can be represented as an infinite moving average of a white noise process. If the series can also be represented in an autoregressive form, then the series is said to be invertible. Nonstationary series do not have corresponding general representations. Linear models are a broad class of nonstationary models and of asymptotically stationary models that provide the theoretical base for ARMA and state-space processes that will be discussed in the following sections.

ARMA REPRESENTATIONS

The infinite moving average or autoregressive representations of the previous section are useful theoretical tools but they cannot be applied to estimate processes. One needs a parsimonious representation with a finite number of coefficients. *Autoregressive moving average (ARMA)* models and state-space models provide such representation; though apparently conceptually different, they are statistically equivalent.

Stationary Univariate ARMA Models

Let's start with univariate stationary processes. An autoregressive process of order p – AR(p) is a process of the form:

$$x_t + a_1 x_{t-1} + \dots + a_p x_{t-p} = \varepsilon_t$$

which can be written using the lag operator as

$$A(L)x_t = (1 + a_1 L + \dots + a_p L^p)x_t$$

$$= x_t + a_1 L x_t + \dots + a_p L^p x_{t-p} = \varepsilon_t$$

Not all processes that can be written in autoregressive form are stationary. In order to study the stationarity of an autoregressive process, consider the following polynomial:

$$A(z) = 1 + a_1 z + \dots + a_p z^p$$

where z is a complex variable.

The equation

$$A(z) = 1 + a_1 z + \dots + a_p z^p = 0$$

is called the inverse characteristic equation. It can be demonstrated that if the roots of this equation, that is, its solutions, are all strictly greater than 1 in modulus (that is, the roots are outside the unit circle), then the operator $A(L)$ is invertible and admits the inverse representation:

$$x_t = A^{-1}(L)\varepsilon_t$$

$$= \sum_{i=0}^{+\infty} \lambda_i \varepsilon_{t-i}, \quad \text{with} \quad \sum_{i=0}^{+\infty} |\lambda_i| < +\infty$$

In order to avoid possible confusion, note that the solutions of the inverse characteristic equation are the reciprocal of the solution of the characteristic equation defined as

$$A(z) = z^p + a_1z^{p-1} + \dots + a_p = 0$$

Therefore an autoregressive process is invertible with an infinite moving average representation that only involves positive powers of the operator L if the solutions of the characteristic equation are all strictly smaller than 1 in absolute value. This is the condition of invertibility often stated in the literature.

Let's now consider finite moving-average representations. A process is called a moving average process of order q – $MA(q)$ if it admits the following representation:

$$\begin{aligned} x_t &= (1 + b_1L + \dots + b_pL^q)\varepsilon_t \\ &= \varepsilon_t + b_1\varepsilon_{t-1} + \dots + b_p\varepsilon_{t-q} \end{aligned}$$

In a way similar to the autoregressive case, if the roots of the equation

$$B(z) = 1 + b_1z + \dots + b_qz^q = 0$$

are all strictly greater than 1 in modulus, then the $MA(q)$ process is invertible and, therefore, admits the infinite autoregressive representation:

$$\begin{aligned} \varepsilon_t &= B^{-1}(L)x_t \\ &= \sum_{i=0}^{+\infty} \pi_i x_{t-i}, \quad \text{with} \quad \sum_{i=0}^{+\infty} |\pi_i| < +\infty \end{aligned}$$

As in the previous case, if one considers the characteristic equation,

$$B(z) = z^q + b_1z^{q-1} + \dots + b_q = 0$$

then the $MA(q)$ process admits a causal autoregressive representation if the roots of the characteristic equation are strictly smaller than 1 in modulus.

Let's now consider, more in general, an ARMA process of order p, q . We say that a stationary process admits a minimal ARMA(p, q) representation if it can be written as

$$x_t + a_1x_{t-1} + a_px_{t-p} = b_1\varepsilon_t + \dots + b_q\varepsilon_{t-q}$$

or equivalently in terms of the lag operator

$$A(L)x_t = B(L)\varepsilon_t$$

where ε_t is a serially uncorrelated white noise with nonzero variance, $a_0 = b_0 = 1$, $a_p \neq 0$, $b_q \neq 0$, the polynomials A and B have roots strictly greater than 1 in modulus and do not have any root in common.

Generalizing the reasoning in the pure MA or AR case, it can be demonstrated that a generic process that admits the ARMA(p, q) representation $A(L)x_t = B(L)\varepsilon_t$ is stationary if both polynomials A and B have roots strictly different from 1. In addition, if all the roots of the polynomial $A(z)$ are strictly greater than 1 in modulus, then the ARMA(p, q) process can be expressed as a moving average process:

$$x_t = \frac{B(L)}{A(L)}\varepsilon_t$$

Conversely, if all the roots of the polynomial $B(z)$ are strictly greater than 1, then the ARMA(p, q) process can be expressed as an autoregressive process:

$$\varepsilon_t = \frac{A(L)}{B(L)}x_t$$

Note that in the above discussions every process was centered—that is, it had zero constant mean. As we were considering stationary processes, this condition is not restrictive as the eventual nonzero mean can be subtracted.

Note also that ARMA stationary processes extend through the entire time axis. An ARMA process, which begins from some initial conditions at starting time $t = 0$, is not stationary even if its roots are strictly outside the unit circle. It can be demonstrated, however, that such a process is asymptotically stationary.

Nonstationary Univariate ARMA Models

So far we have considered only stationary processes. However, ARMA equations can also represent nonstationary processes if some of the

roots of the polynomial $A(z)$ are equal to 1 in modulus. A process defined by the equation

$$A(L)x_t = B(L)\varepsilon_t$$

is called an *autoregressive integrated moving-average (ARIMA) process* if at least one of the roots of the polynomial A is equal to 1 in modulus. Suppose that λ be a root with multiplicity d . In this case the ARMA representation can be written as

$$A'(L)(I - \lambda L)^d x_t = B(L)\varepsilon_t$$

$$A(L) = A'(L)(I - \lambda L)^d$$

However this formulation is not satisfactory as the process A is not invertible if initial conditions are not provided; it is therefore preferable to offer a more rigorous definition, which includes initial conditions. Therefore, we give the following definition of nonstationary integrated ARMA processes.

A process x_t defined for $t \geq 0$ is called an *autoregressive integrated moving-average process—ARIMA(p, d, q)*—if it satisfies a relationship of the type

$$A(L)(I - \lambda L)^d x_t = B(L)\varepsilon_t$$

where:

- The polynomials $A(L)$ and $B(L)$ have roots strictly greater than 1.
- ε_t is a white noise process defined for $t \geq 0$.
- A set of initial conditions $(x_{-1}, \dots, x_{-p-d}, \varepsilon_t, \dots, \varepsilon_{-q})$ independent from the white noise is given.

Later in this entry we discuss the interpretation and further properties of the ARIMA condition.

Stationary Multivariate ARMA Models

Let's now move on to consider stationary multivariate processes. A stationary process that admits an infinite moving-average representation

of the type

$$x_t = \sum_{i=0}^{\infty} H_i \varepsilon_{t-i}$$

where ε_{t-i} is an n -dimensional, zero-mean, white-noise process with nonsingular variance-covariance matrix Ω is called an *autoregressive moving-average—ARMA(p, q)—model*, if it satisfies a difference equation of the type

$$A(L)x_t = B(L)\varepsilon_t$$

where A and B are matrix polynomials in the lag operator L of order p and q respectively:

$$A(L) = \sum_{i=0}^p A_i L^i, \quad A_0 = I, A_p \neq 0$$

$$B(L) = \sum_{j=0}^q B_j L^j, \quad B_0 = I, B_q \neq 0$$

If $q = 0$, the process is purely autoregressive of order p ; if $p = 0$, the process is purely a moving average of order q . Rearranging the terms of the difference equation, it is clear that an ARMA process is a process where the i -th component of the process at time t , $x_{i,t}$ is a linear function of all the components at different lags plus a finite moving average of white noise terms.

It can be demonstrated that the ARMA representation is not unique. The nonuniqueness of the ARMA representation is due to different reasons, such as the existence of a common polynomial factor in the autoregressive and the moving-average part. It entails that the same process can be represented by models with different pairs p, q . For this reason, one would need to determine at least a minimal representation—that is, an ARMA(p, q) representation such that any other ARMA(p', q') representation would have $p' > p, q' > q$. With the exception of the univariate case, these problems are very difficult from a mathematical point of view and we will not examine them in detail.

Let's now explore what restrictions on the polynomials $A(L)$ and $B(L)$ ensure that the relative ARMA process is stationary. Generalizing the univariate case, the mathematical analysis

of stationarity is based on the analysis of the polynomial $\det[\mathbf{A}(z)]$ obtained by formally replacing the lag operator L with a complex variable z in the matrix $\mathbf{A}(L)$ whose entries are finite polynomials in L .

It can be demonstrated that if the complex roots of the polynomial $\det[\mathbf{A}(z)]$, that is, the solutions of the algebraic equation $\det[\mathbf{A}(z)] = 0$, which are in general complex numbers, all lie outside the unit circle, that is, their modulus is strictly greater than one, then the process that satisfies the ARMA conditions,

$$\mathbf{A}(L)\mathbf{x}_t = \mathbf{B}(L)\boldsymbol{\varepsilon}_t$$

is stationary. As in the univariate case, if one would consider the equations in $1/z$, the same reasoning applies but with roots strictly inside the unit circle.

A stationary ARMA(p,q) process is an autocorrelated process. Its time-independent autocorrelation function satisfies a set of linear difference equations. Consider an ARMA(p,q) process that satisfies the following equation:

$$\mathbf{A}_0\mathbf{x}_t + \mathbf{A}_1\mathbf{x}_{t-1} + \dots + \mathbf{A}_p\mathbf{x}_{t-p} = \mathbf{B}_0\boldsymbol{\varepsilon}_t + \mathbf{B}_1\boldsymbol{\varepsilon}_{t-1} + \dots + \mathbf{B}_q\boldsymbol{\varepsilon}_{t-q}$$

where $\mathbf{A}_0 = \mathbf{I}$. By expanding the expression for the autocovariance function, it can be demonstrated that the autocovariance function satisfies the following set of linear difference equations:

$$\mathbf{A}_0\boldsymbol{\Gamma}_h + \mathbf{A}_1\boldsymbol{\Gamma}_{h-1} + \dots + \mathbf{A}_p\boldsymbol{\Gamma}_{h-p} = \begin{cases} 0 & \text{if } h > q \\ \sum_{j=0}^{q-h} \mathbf{B}_{j+h}\boldsymbol{\Omega}\mathbf{H}'_j & \text{if } h \leq q \end{cases}$$

where $\boldsymbol{\Omega}$ and \mathbf{H}_i are, respectively, the covariance matrix and the Markov coefficients of the process in its infinite moving-average representation:

$$\mathbf{x}_t = \sum_{i=0}^{\infty} \mathbf{H}_i \boldsymbol{\varepsilon}_{t-i}$$

From the above representation, it is clear that if the process is purely MA, that is, if $p = 0$,

then the autocovariance function vanishes for lag $h > q$.

It is also possible to demonstrate the converse of this theorem. If a linear stationary process admits an autocovariance function that satisfies the following equations,

$$\mathbf{A}_0\boldsymbol{\Gamma}_h + \mathbf{A}_1\boldsymbol{\Gamma}_{h-1} + \dots + \mathbf{A}_p\boldsymbol{\Gamma}_{h-p} = 0 \quad \text{if } h > q$$

then the process admits an ARMA(p,q) representation. In particular, a stationary process is a purely finite moving-average process MA(q), if and only if its autocovariance functions vanish for $h > q$, where q is an integer.

Nonstationary Multivariate ARMA Models

Let's now consider nonstationary series. Consider a series defined for $t \geq 0$ that satisfies the following set of difference equations:

$$\mathbf{A}_0\mathbf{x}_t + \mathbf{A}_1\mathbf{x}_{t-1} + \dots + \mathbf{A}_p\mathbf{x}_{t-p} = \mathbf{B}_0\boldsymbol{\varepsilon}_t + \mathbf{B}_1\boldsymbol{\varepsilon}_{t-1} + \dots + \mathbf{B}_q\boldsymbol{\varepsilon}_{t-q}$$

where, as in the stationary case, $\boldsymbol{\varepsilon}_{t-i}$ is an n -dimensional zero-mean, white noise process with nonsingular variance-covariance matrix $\boldsymbol{\Omega}$, $\mathbf{A}_0 = \mathbf{I}$, $\mathbf{B}_0 = \mathbf{I}$, $\mathbf{A}_p \neq 0$, $\mathbf{B}_q \neq 0$. Suppose, in addition, that initial conditions ($\mathbf{x}_{-1}, \dots, \mathbf{x}_{-p}, \boldsymbol{\varepsilon}_t, \dots, \boldsymbol{\varepsilon}_{-q}$) are given. Under these conditions, we say that the process \mathbf{x}_t , which is well defined, admits an ARMA representation.

A process \mathbf{x}_t is said to admit an ARIMA representation if, in addition to the above, it satisfies the following two conditions: (1) $\det[\mathbf{B}(z)]$ has all its roots strictly outside of the unit circle, and (2) $\det[\mathbf{A}(z)]$ has all its roots outside the unit circle but with at least one root equal to 1. In other words, an ARIMA process is an ARMA process that satisfies some additional conditions. Later in this entry we will clarify the meaning of integrated processes.

Markov Coefficients and ARMA Models

For the theoretical analysis of ARMA processes, it is useful to state what conditions on the Markov coefficients ensure that the process admits an ARMA representation. Consider a process \mathbf{x}_t , stationary or not, which admits a moving-average representation either as

$$\mathbf{x}_t = \sum_{i=0}^{\infty} \mathbf{H}_i \boldsymbol{\varepsilon}_{t-i}$$

or as a linear model:

$$\mathbf{x}_t = \sum_{i=0}^t \mathbf{H}_i \boldsymbol{\varepsilon}_{t-i} + \mathbf{h}(t)\mathbf{z}$$

The process \mathbf{x}_t admits an ARMA representation if and only if there is an integer q and a set of p matrices $\mathbf{A}_i, i = 0, \dots, p$ such that the Markov coefficients \mathbf{H}_i satisfy the following linear difference equation starting from q :

$$\sum_{j=0}^p \mathbf{A}_j \mathbf{H}_{l-j} = 0, \quad l > q$$

Therefore, any ARMA process admits an infinite moving-average representation whose Markov coefficients satisfy a linear difference equation starting from a certain point. Conversely, any such linear infinite moving-average representation can be expressed parsimoniously in terms of an ARMA process.

Hankel Matrices and ARMA Models

For the theoretical analysis of ARMA processes it is also useful to restate the above conditions in terms of the Hankel infinite matrices. (A Hankel matrix is a matrix where for each antidiagonal the element is the same.) It can be demonstrated that a process, stationary or not, which admits either the infinite moving average representation

$$\mathbf{x}_t = \sum_{i=0}^{\infty} \mathbf{H}_i \boldsymbol{\varepsilon}_{t-i}$$

or a linear moving average model

$$\mathbf{x}_t = \sum_{i=0}^t \mathbf{H}_i \boldsymbol{\varepsilon}_{t-i} + \mathbf{h}(t)\mathbf{z}$$

also admits an ARMA representation if and only if the Hankel matrix formed with the sequence of its Markov coefficients has finite rank or, equivalently, a finite column rank or row rank.

INTEGRATED SERIES AND TRENDS

This section introduces the fundamental notions of trend stationary series, difference stationary series, and integrated series. Consider a one-dimensional time series. A *trend stationary series* is a series formed by a deterministic trend plus a stationary process. It can be written as

$$X_t = f(t) + \varepsilon(t)$$

A trend stationary process can be transformed into a stationary process by subtracting the trend. Removing the deterministic trend entails that the deterministic trend is known. A trend stationary series is an example of an adjustment model.

Consider now a time series X_t . The operation of differencing a series consists of forming a new series $Y_t = \Delta X_t = X_t - X_{t-1}$. The operation of differencing can be repeated an arbitrary number of times. For instance, differencing twice the series X_t yields the following series:

$$\begin{aligned} Z_t &= \Delta Y_t = \Delta(\Delta X_t) \\ &= (X_t - X_{t-1}) - (X_{t-2} - X_{t-3}) \\ &= X_t - X_{t-1} - X_{t-2} + X_{t-3} \end{aligned}$$

Differencing can be written in terms of the lag operator as

$$\Delta X_t^d = (1 - L)^d X_t$$

A difference stationary series is a series that is transformed into a stationary series by differencing. A difference stationary series can be

written as

$$\Delta X_t = \mu + \varepsilon(t)$$

$$X_t = X_{t-1} + \mu + \varepsilon(t)$$

where $\varepsilon(t)$ is a zero-mean stationary process and μ is a constant. A trend stationary series with a linear trend is also difference stationary, if spacings are regular. The opposite is not generally true. A time series is said to be integrated of order n if it can be transformed into a stationary series by differencing n times.

Note that the concept of integrated series as defined above entails that a series extends on the entire time axis. If a series starts from a set of initial conditions, the difference sequence can only be asymptotically stationary.

There are a number of obvious differences between trend stationary and difference stationary series. A trend stationary series experiences stationary fluctuation, with constant variance, around an arbitrary trend. A difference stationary series meanders arbitrarily far from a linear trend, producing fluctuations of growing variance. The simplest example of difference stationary series is the random walk.

An integrated series is characterized by a stochastic trend. In fact, a difference stationary series can be written as

$$X_t = \mu t + \left[\sum_{s=0}^{t-1} \varepsilon(s) \right] + \varepsilon(t)$$

The difference $X_t - X_t^*$ between the value of a process at time t and the best affine prediction at time $t - 1$ is called the innovation of the process. In the above linear equation, the stationary process $\varepsilon(t)$ is the innovation process. A key aspect of integrated processes is that innovations $\varepsilon(t)$ never decay but keep on accumulating. In a trend stationary process, on the other hand, past innovations disappear at every new step.

These considerations carry over immediately in a multidimensional environment. Multidimensional trend stationary series will exhibit multiple trends, in principle one for

each component. Multidimensional difference-stationary series will yield a stationary process after differencing.

Let's now see how these concepts fit into the ARMA framework, starting with univariate ARMA model. Recall that an ARIMA process is defined as an ARMA process in which the polynomial B has all roots outside the unit circle while the polynomial A has one or more roots equal to 1. In the latter case the process can be written as

$$A'(L)\Delta^d x_t = B(L)\varepsilon_t$$

$$A(L) = (1 - L)^d A'(L)$$

and we say that the process is integrated of order n . If initial conditions are supplied, the process can be inverted and the difference sequence is asymptotically stationary.

The notion of integrated processes carries over naturally in the multivariate case but with a subtle difference. Recall from earlier discussion in this entry that an ARIMA model is an ARMA model:

$$\mathbf{A}(L)\mathbf{x}_t = \mathbf{B}(L)\varepsilon_t$$

which satisfies two additional conditions: (1) $\det[\mathbf{B}(z)]$ has all its roots strictly outside of the unit circle, and (2) $\det[\mathbf{A}(z)]$ has all its roots outside the unit circle but with at least one root equal to 1.

Now suppose that, after differencing d times, the multivariate series $\Delta^d \mathbf{x}_t$ can be represented as follows:

$$\mathbf{A}'(L)\mathbf{x}_t = \mathbf{B}'(L)\varepsilon_t, 1 \quad \text{with} \quad \mathbf{A}'(L) = \mathbf{A}(L)\Delta^d$$

In this case, if (1) $\mathbf{B}'(z)$ is of order q and $\det[\mathbf{B}'(z)]$ has all its roots strictly outside of the unit circle and (2) $\mathbf{A}'(z)$ is of order p and $\det[\mathbf{A}'(z)]$ has all its roots outside the unit circle, then the process is called ARIMA(p, d, q). Not all ARIMA models can be put in this framework as different components might have a different order of integration.

Note that in an ARIMA(p, d, q) model each component series of the multivariate model is individually integrated. A multivariate series is

integrated of order d if every component series is integrated of order d .

Note also that ARIMA processes are not invertible as infinite moving averages, but as discussed, they can be inverted in terms of a generic linear moving-average model with stochastic initial conditions. In addition, the process in the d -differences is asymptotically stationary.

In both trend stationary and difference stationary processes, innovations can be serially autocorrelated. In the ARMA representations discussed in the previous section, innovations are serially uncorrelated white noise as all the autocorrelations are assumed to be modeled in the ARMA model. If there is residual autocorrelation, the ARMA or ARIMA model is somehow misspecified.

The notion of an integrated process is essentially linear. A process is integrated if stationary innovations keep on adding indefinitely. Note that innovations could, however, cumulate in ways other than addition, producing essentially nonlinear processes. In ARCH and GARCH processes for instance, innovations do not simply add to past innovations.

The behavior of integrated and nonintegrated time series is quite different and the estimation procedures are different as well. It is therefore important to ascertain if a series is integrated or not. Often a preliminary analysis to ascertain integratedness suggests what type of model should be used.

A number of statistical tests to ascertain if a univariate series is integrated are available. Perhaps the most widely used and known are the Dickey-Fuller (DF) and the Augmented Dickey-Fuller (ADF) tests. The DF test assumes as a null hypothesis that the series is integrated of order 1 with uncorrelated innovations. Under this assumption, the series can be written as a random walk in the following form:

$$\begin{aligned} X_{t+1} &= \rho X_t + b + \varepsilon_t \\ \rho &= 1 \\ \varepsilon_t &\text{ IID} \end{aligned}$$

where IID is an independent and identical sequence.

In a sample generated by a model of this type, the value of ρ estimated on the sample is stochastic. Estimation can be performed with the ordinary least square (OLS) method. Dickey and Fuller determined the theoretical distribution of ρ and computed the critical values of ρ that correspond to different confidence intervals. The theoretical distribution of ρ is determined computing a functional of the Brownian motion.

Given a sample of a series, for instance a series of log prices, application of the DF test entails computing the autoregressive parameter ρ on the given sample and comparing it with the known critical values for different confidence intervals. The strict hypothesis of random walk is too strong for most econometric applications. The DF test was extended to cover the case of correlated residuals that are modeled as a linear model. In the latter case, the DF test is called the Augmented Dickey-Fuller or ADF test. The Phillips and Perron test is the DF test in the general case of autocorrelated residuals.

APPENDIX

We will begin with several concepts from probability theory.

Stochastic Processes

When it is necessary to emphasize the dependence of the random variable from both time t and the element ω , a stochastic process is explicitly written as a function of two variables: $X = X(t, \omega)$. Given ω , the function $X = X_t(\omega)$ is a function of time that is referred to as the path of the stochastic process.

The variable X might be a single random variable or a multidimensional random vector. A stochastic process is therefore a function $X = X(t, \omega)$ from the product space $[0, T] \times \Omega$ into the n -dimensional real space R^n . Because to each ω corresponds a time path of the process—in

general formed by a set of functions $X = X_t(\omega)$ —it is possible to identify the space Ω with a subset of the real functions defined over an interval $[0, T]$.

Let's now discuss how to represent a stochastic process $X = X(t, \omega)$ and the conditions of identity of two *stochastic processes*. As a stochastic process is a function of two variables, we can define equality as pointwise identity for each couple (t, ω) . However, as processes are defined over probability spaces, pointwise identity is seldom used. It is more fruitful to define equality modulo sets of measure zero or equality with respect to probability distributions. In general, two random variables X, Y will be considered equal if the equality $X(\omega) = Y(\omega)$ holds for every ω with the exception of a set of probability zero. In this case, it is said that the equality holds almost everywhere (denoted *a.e.*).

A rather general (but not complete) representation is given by the finite dimensional probability distributions. Given any set of indices t_1, \dots, t_m , consider the distributions

$$\mu_{t_1, \dots, t_m}(H) = P[(X_{t_1}, \dots, X_{t_m}) \in H, H \in \mathfrak{B}^n]$$

These probability measures are, for any choice of the t_i , the finite-dimensional joint probabilities of the process. They determine many, but not all, properties of a stochastic process. For example, the finite dimensional distributions of a Brownian motion do not determine whether or not the process paths are continuous.

In general, the various concepts of equality between stochastic processes can be described as follows:

- *Property 1.* Two stochastic processes are weakly equivalent if they have the same finite-dimensional distributions. This is the weakest form of equality.
- *Property 2.* The process $X = X(t, \omega)$ is said to be equivalent or to be a modification of the process $Y = Y(t, \omega)$ if, for all t ,

$$P(X_t = Y_t) = 1$$

- *Property 3.* The process $X = X(t, \omega)$ is said to be strongly equivalent to or indistinguishable from the process $Y = Y(t, \omega)$ if

$$P(X_t = Y_t, \text{ for all } t) = 1$$

Property 3 implies Property 2, which in turn implies Property 1. Implications do not hold in the opposite direction. Two processes having the same finite distributions might have completely different paths. However it is possible to demonstrate that if one assumes that paths are continuous functions of time, Properties 2 and 3 become equivalent.

Information Structures

Let's now turn our attention to the question of time. We must introduce an appropriate representation of time as well as rules that describe the evolution of information, that is, information propagation, over time. The concepts of information and information propagation are fundamental in economics and finance theory.

The concept of information in finance is different from both the intuitive notion of information and that of information theory in which information is a quantitative measure related to the *a priori* probability of messages. In our context, information means the (progressive) revelation of the set of events to which the current state of the economy belongs. Though somewhat technical, this concept of information sheds light on the probabilistic structure of finance theory. The point is the following. Assets are represented by stochastic processes, that is, time-dependent random variables. But the probabilistic states on which these random variables are defined represent entire histories of the economy. To embed time into the probabilistic structure of states in a coherent way calls for information structures and filtrations (a concept we explain next).

It is assumed that the economy is in one of many possible states and that there is uncertainty on the state that has been realized.

Consider a time period of the economy. At the beginning of the period, there is complete uncertainty on the state of the economy (i.e., there is complete uncertainty on what path the economy will take). Different events have different probabilities, but there is no certainty. As time passes, uncertainty is reduced as the number of states to which the economy can belong is progressively reduced. Intuitively, revelation of information means the progressive reduction of the number of possible states; at the end of the period, the realized state is fully revealed. In continuous time and continuous states, the number of events is infinite at each instant. Thus its cardinality remains the same. We cannot properly say that the number of events shrinks. A more formal definition is required.

The progressive reduction of the set of possible states is formally expressed in the concepts of information structure and filtration. Let's start with *information structures*. Information structures apply only to discrete probabilities defined over a discrete set of states. At the initial instant T_0 , there is complete uncertainty on the state of the economy; the actual state is known only to belong to the largest possible event (that is, the entire space Ω). At the following instant T_1 , assuming that instants are discrete, the states are separated into a partition, a partition being a denumerable class of disjoint sets whose union is the space itself. The actual state belongs to one of the sets of the partitions. The revelation of information consists in ruling out all sets but one. For all the states of each partition, and only for these, random variables assume the same values.

Suppose, to exemplify, that only two assets exist in the economy and that each can assume only two possible prices and pay only two possible cash flows. At every moment there are 16 possible price-cash flow combinations. We can thus see that at the moment T_1 all the states are partitioned into 16 sets, each containing only one state. Each partition includes all the states that have a given set of prices and cash distributions at the moment T_1 . The same reasoning

can be applied to each instant. The evolution of information can thus be represented by a tree structure in which every path represents a state and every point a partition. Obviously the tree structure does not have to develop as symmetrically as in the above example; the tree might have a very generic structure of branches.

Filtration

The concept of information structure based on partitions provides a rather intuitive representation of the propagation of information through a tree of progressively finer partitions. However, this structure is not sufficient to describe the propagation of information in a general probabilistic context. In fact, the set of possible events is much richer than the set of partitions. It is therefore necessary to identify not only partitions but also a structure of events. The structure of events used to define the propagation of information is called a *filtration*. In the discrete case, however, the two concepts—information structure and filtration—are equivalent.

The concept of filtration is based on identifying all events that are known at any given instant. It is assumed that it is possible to associate to each trading moment t a σ -algebra of events $\mathfrak{F}_t \subset \mathfrak{F}$ formed by all events that are known prior to or at time t . It is assumed that events are never “forgotten,” that is, that $\mathfrak{F}_t \subset \mathfrak{F}_s$, if $t < s$. An ordering of time is thus created. This ordering is formed by an increasing sequence of σ -algebras, each associated to the time at which all its events are known. This sequence is a filtration. Indicated as $\{\mathfrak{F}_t\}$, a filtration is therefore an increasing sequence of all σ -algebras \mathfrak{F}_t , each associated to an instant t .

In the finite case, it is possible to create a mutual correspondence between filtrations and information structures. In fact, given an information structure, it is possible to associate to each partition the algebra generated by the same partition. Observe that a tree information structure is formed by partitions that create

increasing refinement: By going from one instant to the next, every set of the partition is decomposed. One can then conclude that the algebras generated by an information structure form a filtration.

On the other hand, given a filtration $\{\mathfrak{J}_t\}$, it is possible to associate a partition to each \mathfrak{J}_t . In fact, given any element that belongs to Ω , consider any other element that belongs to Ω such that, for each set of \mathfrak{J}_t , both either belong to or are outside this set. It is easy to see that classes of equivalence are thus formed, that these create a partition, and that the algebra generated by each such partition is precisely the \mathfrak{J}_t that has generated the partition.

A stochastic process is said to be adapted to the filtration $\{\mathfrak{J}_t\}$ if the variable X_t is measurable with respect to the σ -algebra \mathfrak{J}_t . It is assumed that the price and cash distribution processes $S_t(\omega)$ and $d_t(\omega)$ of every asset are adapted to $\{\mathfrak{J}_t\}$. This means that, for each t , no measurement of any price or cash distribution variable can identify events not included in the respective algebra or σ -algebra. Every random variable is a partial image of the set of states seen from a given point of view and at a given moment.

The concepts of filtration and of processes adapted to a filtration are fundamental. They ensure that information is revealed without anticipation. Consider the economy and associate at every instant a partition and an algebra generated by the partition. Every random variable defined at that moment assumes a value constant on each set of the partition. The knowledge of the realized values of the random variables does not allow identifying sets of events finer than partitions.

One might well ask: Why introduce the complex structure of σ -algebras as opposed to simply defining random variables? The point is that, from a logical point of view, the primitive concept is that of states and events. The evolution of time has to be defined on the primitive structure—it cannot simply be imposed on random variables. In practice, filtrations be-

come an important concept when dealing with conditional probabilities in a continuous environment. As the probability that a continuous random variable assumes a specific value is zero, the definition of conditional probabilities requires the machinery of filtration.

Conditional Probability and Conditional Expectation

Conditional probabilities and conditional averages are fundamental in the stochastic description of financial markets. For instance, one is generally interested in the probability distribution of the price of an asset at some date given its price at an earlier date. The widely used regression models are an example of conditional expectation models.

The *conditional probability* of event A given event B was defined earlier as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

This simple definition cannot be used in the context of continuous random variables because the conditioning event (i.e., one variable assuming a given value) has probability zero. To avoid this problem, we condition on σ -algebras and not on single zero-probability events. In general, as each instant is characterized by a σ -algebra \mathfrak{J}_t , the conditioning elements are the \mathfrak{J}_t .

The general definition of *conditional expectation* is the following. Consider a probability space $(\Omega, \mathfrak{J}, P)$ and a σ -algebra \mathfrak{G} contained in \mathfrak{J} and suppose that X is an integrable random variable on $(\Omega, \mathfrak{J}, P)$. We define the conditional expectation of X with respect to \mathfrak{G} , written as $E[X|\mathfrak{G}]$, as a random variable measurable with respect to \mathfrak{G} such that

$$\int_G E[X|\mathfrak{G}] dP$$

for every set $G \in \mathfrak{G}$. In other words, the *conditional expectation* is a random variable whose average on every event that belongs to \mathfrak{G} is equal to the average of X over those same events, but

it is \mathcal{G} -measurable while X is not. It is possible to demonstrate that such variables exist and are unique up to a set of measure zero.

Econometric models usually condition a random variable given another variable. In the previous framework, conditioning one random variable X with respect to another random variable Y means conditioning X given $\sigma\{Y\}$ (i.e., given the σ -algebra generated by Y). Thus $E[X|Y]$ means $E[X|\sigma\{Y\}]$.

This notion might seem to be abstract and to miss a key aspect of conditioning: intuitively, conditional expectation is a function of the conditioning variable. For example, given a stochastic price process, X_t , one would like to visualize conditional expectation $E[X_t | X_s]$, $s < t$ as a function of X_s that yields the expected price at a future date given the present price. This intuition is not wrong insofar as the conditional expectation $E[X|Y]$ of X given Y is a random variable function of Y .

However, we need to specify how conditional expectations are formed, given that the usual conditional probabilities cannot be applied as the conditioning event has probability zero. Here is where the above definition comes into play. The conditional expectation of a variable X given a variable Y is defined in full generality as a variable that is measurable with respect to the σ -algebra $\sigma(Y)$ generated by the conditioning variable Y and has the same expected value of Y on each set of $\sigma(Y)$. Later in this section we will see how conditional expectations can be expressed in terms of the joint p.d.f. of the conditioning and conditioned variables.

One can define conditional probabilities starting from the concept of conditional expectations. Consider a probability space $(\Omega, \mathfrak{J}, P)$, a sub- σ -algebra \mathcal{G} of \mathfrak{J} , and two events $A \in \mathfrak{J}$, $B \in \mathfrak{J}$. If I_A, I_B are the indicator functions of the sets A, B (the indicator function of a set assumes value 1 on the set, 0 elsewhere), we can define conditional probabilities of the event A , respectively, given \mathcal{G} or given the event B as

$$P(A|\mathcal{G}) = E[I_A|\mathcal{G}] \quad P(A|B) = E[I_A|I_B]$$

Using these definitions, it is possible to demonstrate that given two random variables X and Y with joint density $f(x, y)$, the conditional density of X given Y is

$$f(x|y) = \frac{f(x, y)}{f_Y(y)}$$

where the marginal density, defined as

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

is assumed to be strictly positive.

In the discrete case, the conditional expectation is a random variable that takes a constant value over the sets of the finite partition associated to \mathfrak{J}_t . Its value for each element of Ω is defined by the classical concept of conditional probability. Conditional expectation is simply the average over a partition assuming the classical conditional probabilities.

An important econometric concept related to conditional expectations is that of a martingale. Given a probability space $(\Omega, \mathfrak{J}, P)$ and a filtration $\{\mathfrak{J}_t\}$, a sequence of \mathfrak{J}_t -measurable random variables X_t is called a martingale if the following condition holds:

$$E[X_{t+1}|\mathfrak{J}_t] = X_t$$

A martingale translates the idea of a “fair game” as the expected value of the variable at the next period is the present value of the same value.

KEY POINTS

- Stochastic processes are time-dependent random variables.
- An information structure is a collection of partitions of events associated to each instant of time that become progressively finer with the evolution of time. A filtration is an increasing collection of σ -algebras associated to each instant of time.
- The states of the economy, intended as full histories of the economy, are represented as a

probability space. The revelation of information with time is represented by information structures or filtrations. Prices and other financial quantities are represented by adapted stochastic processes.

- By conditioning is meant the change in probabilities due to the acquisition of some information. It is possible to condition with respect to an event if the event has nonzero probability. In general terms, conditioning means conditioning with respect to a filtration or an information structure.
- A martingale is a stochastic process such that the conditional expected value is always equal to its present value. It embodies the idea of a fair game where today's wealth is the best forecast of future wealth.
- A time series is a discrete-time stochastic process, that is, a denumerable collection of random variables indexed by integer numbers.
- Any stationary time series admits an infinite moving average representation, that is to say, it can be represented as an infinite sum of white noise terms with appropriate coefficients.
- A time series is said to be invertible if it can also be represented as an infinite autoregression, that is, an infinite sum of all past terms with appropriate coefficients.

- ARMA models are parsimonious representations that involve only a finite number of moving average and autoregressive terms.
- An ARMA model is stationary if all the roots of the inverse characteristic equation of the AR or the MA part have roots with modulus strictly greater than one.
- A process is said to be integrated of order p if it becomes stationary after differencing p times.

NOTE

1. See Enders (2009), Gouriéroux and Monfort (1997), Hamilton (1994), and Tsay (2001) for a comprehensive discussion of modern time series econometrics.

REFERENCES

- Enders, W. (2009). *Applied Econometric Time Series: 3rd Ed.* Hoboken, NJ: John Wiley & Sons.
- Gouriéroux, C., and Monfort, A. (1997). *Time Series and Dynamic Models.* Cambridge: Cambridge University Press.
- Hamilton, J. D. (1994). *Time Series Analysis.* Princeton, NJ: Princeton University Press.
- Tsay, R.S. (2001). *Analysis of Financial Time Series.* Hoboken, NJ: John Wiley & Sons.

Extracting Risk-Neutral Density Information from Options Market Prices

ARTURO LECCADITO, PhD

Business Administration Department, Università della Calabria

RADU TUNARU, PhD

Professor of Quantitative Finance, Business School, University of Kent

Abstract: Information on risk-neutral density is valuable in financial markets for a wide range of participants. This density can be used to mark-to-market exotic options that are not very liquid on the market, for anticipation of effects determined by new policy or possible extreme events such as crashes, and even for designing new trading strategies. There are many models that have been proposed in the past for estimating the risk-neutral density, each with their pros and cons.

The concept of *risk-neutral density* (RND) plays an important theoretical role in asset pricing as outlined in Cox and Ross (1976), published very shortly after the publication of the Black-Scholes model. Since then, the estimation of RND has become an essential tool for central banks in monitoring the stability of the financial system and for measuring the impact of new policies. Investment banks also rely on the RND calibrated from liquid European vanilla options to determine the price of more exotic positions on their balance sheet that are not very liquid. Moreover, the first moments of the RND, such as implied volatility and skewness, can be used to design trading strategies.

One may argue that the information contained in option prices is redundant to the information provided by historical prices of

the underlying asset. However, based on the 1987 stock market crash, Jackwerth and Rubinstein (1996) demonstrated that this is not the case. Prior to the crash, the RND estimated at one-month horizon had been close to lognormal but subsequently the shape of the RND changed considerably. At the same time, they also revealed that the historical distribution had been lognormal and it remained like that after the crash. In other words, the option prices in the equity market contain different information from the historical equity prices.

In this entry we highlight the main steps for estimating the RND associated with an equity index. We exemplify the estimation procedure by applying a model based on both the *generalized inverse Gaussian distribution* that has been advocated in the literature for financial

modeling and the well-known lognormal mixture model that has been widely used by investment houses and central banks. These two models are straightforward and easy to apply since the option pricing formulas can be derived in closed form.

AN APPROPRIATE PARAMETRIC MODEL

The RND is recovered from a bundle of European vanilla call and put option prices on the same underlying asset X and with the same maturity T . The options differ in the exercise price K . Denoting with $f(\cdot)$ the probability density function of the underlying asset X under the risk-neutral probability measure \mathbb{Q} , the European vanilla call price for strike K is

$$C(K) = e^{-rT} \int_K^{\infty} (X_T - K) f(X_T) dX_T \quad (1)$$

where r is the continuous compounding risk-free rate.

The partial derivative of 1 with respect to the strike price K

$$\begin{aligned} \frac{\partial C}{\partial K} &= e^{-rT} \frac{\partial}{\partial K} \left[\int_K^{\infty} (X_T - K) f(X_T) dX_T \right] \\ &= -e^{-rT} \int_K^{\infty} f(X_T) dX_T = -e^{-rT} [1 - F(K)] \end{aligned}$$

where $F(\cdot)$ is the cumulative distribution function under the risk-neutral measure. Thus

$$F(K) = e^{rT} \frac{\partial C}{\partial K} + 1 \quad (2)$$

The RND probability function f can be obtained by derivation of the cumulative function F

$$f(K) = e^{rT} \frac{\partial^2 C}{\partial K^2} \quad (3)$$

One could then try to reconstruct either F or f from a grid of option prices using finite difference schemes. However, such numerical methods are notoriously unreliable and very sensitive to the sample of option prices available.

Over the years, two main classes of methods have emerged. First, parametric methods are

underpinned by univariate distributions such as the Weibull distribution (see Savickas, 2002, 2005), the generalized beta distribution (see McDonald and Xu, 1995; Anagnou et al., 2005), the generalized lambda distribution (see Corrado, 2001), the generalized gamma distribution (see Albota et al., 2009), the g-and-h distribution as proposed by Dutta and Babbel (2005); and a mixture of univariate distributions such as that proposed by Gemmill and Saflekos (2000) for two lognormals, and Melick and Thomas (1997) for three lognormals.

The second class is defined by semiparametric and nonparametric methods such as (1) expansion methods as used by Jarrow and Rudd (1982) and Corrado and Su (1997), (2) direct fitting of the implied volatility curve with splines or other interpolation methods as described by Shimko (1993), Anagnou et al. (2003), and Brunner and Hafner (2003), (3) kernel methods developed in Ait-Sahalia and Lo (1998) and Ait-Sahalia and Duarte (2003), and (4) maximum entropy methods as applied by Buchen and Kelly (1996) and Avellaneda (1998).

The nonparametric approach usually requires a large sample of data in order to achieve a good fit. In financial markets, for many asset classes, large samples may simply not be available. In this entry, we focus on the fully parametric approach.

The strategy for parametric models represented by a vector of parameters θ is to minimize some type of discrepancy measure between the theoretical options prices and the observed market prices.

Given the availability of N European call options $\{C(K_{i_j})\}_{j=1,\dots,N}$ and M put options $\{P(K_{s_j})\}_{j=1,\dots,M}$, all with the same maturity T , the problem that must be solved is the minimization of the function

$$\begin{aligned} H_1(\theta) &= \sum_{j=1}^N [C(K_{i_j}) - C^{mkt}(K_{i_j})]^2 \\ &+ \sum_{j=1}^M [P(K_{s_j}) - P^{mkt}(K_{s_j})]^2 \quad (4) \end{aligned}$$

subject to the forward constraint $\mathbb{E}^{\mathbb{Q}}[X_T] = F_0$, where F_0 is the forward price on the same underlying asset X and the last term of the sum accounts for the forward martingale condition that must be satisfied for any parametric model. The notation C^{mkt} , and P^{mkt} , relates, respectively, to the actual option prices from the market. The function H is a discrepancy measure between the theoretical prices obtained under the chosen parametric RND $f(\cdot; \theta)$ and the market prices.

While the H in (4) is widely used in practice, it is sometimes useful to consider other potential discrepancy measures such as

$$\begin{aligned} H_2(\theta) &= \sum_{j=1}^N \frac{[C(K_{i_j}) - C^{mkt}(K_{i_j})]^2}{C^{mkt}(K_{i_j})} \\ &\quad + \sum_{j=1}^M \frac{[P(K_{s_j}) - P^{mkt}(K_{s_j})]^2}{P^{mkt}(K_{s_j})} \\ H_3(\theta) &= \sum_{j=1}^N \frac{[C(K_{i_j}) - C^{mkt}(K_{i_j})]^2}{C(K_{i_j})} \\ &\quad + \sum_{j=1}^M \frac{[P(K_{s_j}) - P^{mkt}(K_{s_j})]^2}{P(K_{s_j})} \\ H_4(\theta) &= \sum_{j=1}^N |[C(K_{i_j}) - C^{mkt}(K_{i_j})]| \\ &\quad + \sum_{j=1}^M |[P(K_{s_j}) - P^{mkt}(K_{s_j})]| \end{aligned}$$

Since the market option prices that do not satisfy put-call parity are filtered out of the data used for calibration, it is possible to work with call prices only or with put prices only, if that is more convenient numerically.

TWO PARAMETRIC MODELS FOR RND ESTIMATION

In order to be able to reverse engineer the RND from options prices, a pricing formula for European vanilla options under the chosen distribution is needed. There is a great advantage in having the pricing formulas in closed form, other-

wise numerical integral approximation methods must be employed and this means that there is a risk of introducing errors in the estimation procedure.

Here we illustrate the RND estimation procedure for two special cases, the general inverse Gaussian (GIG) distribution and the lognormal mixture (LnMix) distribution. For both models, closed-form solutions for pricing European options are available.

Pricing Options with the GIG Distribution

The GIG distribution has been advocated for applications in financial modeling due to its flexibility to fit heavy tails (see Bibby and Sorensen, 2003). The probability density function of the GIG distribution is¹

$$f_{\text{GIG}}(x; \lambda, \chi, \psi) = \frac{x^{\lambda-1} \exp[-\frac{1}{2}(\chi x^{-1} + \psi x)]}{k_{\lambda}(\chi, \psi)} \times I_{(0, \infty)}(x) \quad (5)$$

where

$$k_{\lambda}(\chi, \psi) = \int_0^{\infty} x^{\lambda-1} \exp\left[-\frac{1}{2}(\chi x^{-1} + \psi x)\right] dx$$

is a normalizing constant that is related to the modified Bessel function of the third kind,

$$K_{\nu}(z) = \frac{1}{2} \int_0^{\infty} t^{\nu-1} \exp\left[-\frac{z}{2}(t^{-1} + t)\right] dt \quad (6)$$

via

$$k_{\lambda}(\chi, \psi) = 2 \left(\frac{\chi}{\psi}\right)^{\lambda/2} K_{\lambda}(\sqrt{\chi\psi}) \quad (7)$$

Further technical details on this distribution can be found in Paoletta (2007).

The GIG distribution is well defined, or "proper," for the parameter domain

$$\{(\lambda, \chi, \psi) \in \mathbb{R} \times (0, \infty) \times (0, \infty)\}$$

There are also two boundary cases possible: (1) $\lambda > 0$, $\chi = 0$ and $\psi > 0$ and (2) $\lambda < 0$, $\chi > 0$

and $\psi = 0$. Applying some standard algebraic routine leads to

$$\begin{aligned}
 P(K) &= Ke^{-rT} F_{GIG}(K; \lambda, \chi, \psi) \\
 &\quad - e^{-rT} \int_0^K x F_{GIG}(x; \lambda, \chi, \psi) dx \\
 &= Ke^{-rT} F_{GIG}(K; \lambda, \chi, \psi) \\
 &\quad - e^{-rT} \frac{k_{\lambda+1}(\chi, \psi)}{k_{\lambda}(\chi, \psi)} \\
 &\quad \times \int_0^K f_{GIG}(x; \lambda + 1, \chi, \psi) dx \\
 &= Ke^{-rT} F_{GIG}(K; \lambda, \chi, \psi) \\
 &\quad - e^{-rT} \sqrt{\frac{\chi}{\psi}} \frac{K_{\lambda+1}(\sqrt{\chi\psi})}{K_{\lambda}(\sqrt{\chi\psi})} \\
 &\quad \times F_{GIG}(K; \lambda + 1, \chi, \psi)
 \end{aligned}$$

This formula can be rewritten in terms of the forward price $F_0 = \mathbb{E}^Q(X_T)$ as

$$\begin{aligned}
 P(K) &= Ke^{-rT} F_{GIG}(K; \lambda, \chi, \psi) \\
 &\quad - F_0 e^{-rT} F_{GIG}(K; \lambda + 1, \chi, \psi) \quad (8)
 \end{aligned}$$

RND Estimation with the LnMix Distribution

The importance of fat tails and non-normal distributions in modeling equity stock and vanilla options has become prominent in the aftermath of the Black Monday 1987 crisis. The LnMix model is a convex combination of several lognormal individual models. Bahra (1997) was the first to propose using the LnMix model for RND estimation. An exact solution for options pricing of vanilla European call and put options

can be derived as a weighted sum of standard Black-Scholes prices. In practice, the preferred mixture model is the one based on two individual lognormal models.

If $LN(x; \alpha, \beta)$ is the lognormal distribution with parameters α and β , then the LnMix distribution is given by the following probability density function

$$\begin{aligned}
 f_{LN}(x; \alpha_1, \beta_1, \alpha_2, \beta_2, \eta) &= \eta LN(x; \alpha_1, \beta_1) \\
 &\quad + (1 - \eta) LN(x; \alpha_2, \beta_2) \quad (9)
 \end{aligned}$$

Bahra (1997) described the formulas for pricing European vanilla call and put options

$$\begin{aligned}
 C(K) &= e^{-rT} \{ \eta [e^{(\alpha_1 + 0.5\beta_1^2)} N(d_1) - KN(d_2)] \\
 &\quad + (1 - \eta) [e^{(\alpha_2 + 0.5\beta_2^2)} N(d_3) - KN(d_4)] \} \\
 P(K) &= e^{-rT} \{ \eta [e^{(\alpha_1 + 0.5\beta_1^2)} N(d_1) - KN(d_2)] \\
 &\quad + (1 - \eta) [e^{(\alpha_2 + 0.5\beta_2^2)} N(d_3) - KN(d_4)] \}
 \end{aligned}$$

where

$$\begin{aligned}
 d_1 &= \frac{\alpha_1 + \beta_1^2 - \log(K)}{\beta_1}, & d_2 &= d_1 - \beta_1 \\
 d_3 &= \frac{\alpha_2 + \beta_2^2 - \log(K)}{\beta_2}, & d_4 &= d_3 - \beta_2
 \end{aligned}$$

and N is the standard normal cumulative distribution function.

This model has five parameters $\alpha_1, \beta_1, \alpha_2, \beta_2,$ and η and one should expect a better fit of data with this model compared to the GIG model that has only three parameters. If the *calibration* goodness-of-fit results are very similar between the two models, then the model with fewer parameters should be preferred based on the principle of parsimony.

Table 1 Call Option Prices on May 29, 1998, on the FTSE100 Index

T	F_0	DF	5700	5750	5800	5850	5900	5950	6000	6050
Sep-98	5915.50	0.98	418.81	385.79	354.00	323.51	294.41	266.76	240.62	216.06
Dec-98	6000.11	0.96	586.56	553.83	521.93	490.86	460.62	431.28	402.89	375.54
Mar-99	6079.46	0.95	727.12	694.31	662.15	630.57	599.56	569.19	539.56	510.77
Jun-99	6128.55	0.93	837.32	804.79	772.79	741.21	710.02	679.35	649.34	620.15
Sep-99	6195.66	0.91	950.32	917.58	885.27	853.33	821.71	790.54	759.96	730.09
Dec-99	6269.07	0.90	1061.20	1028.20	995.70	963.43	931.46	899.9	868.85	838.46
Mar-00	6341.98	0.88	1167.80	1134.70	1001.90	1069.40	1037.2	1005.40	974.12	943.36
Jun-00	6383.58	0.87	1250.20	1217.30	1184.70	1152.30	1120.30	1088.60	1057.40	1026.70

Note: Initial value $X_0 = 5843.32$. In the second column the forward prices are reported. The third column reports the discount factors. Strike prices range from 5700 to 6050.

Table 2 Discrepancy Measures across Maturities for the Data in Table 1

	Sep-98	Dec-98	Mar-99	Jun-99	Sep-99	Dec-99	Mar-00	Jun-00
H_1 GIG	2.66E-05	3.21E-05	4.97E-05	5.32E-05	6.98E-05	8.03E-05	2.87E-04	1.41E-04
H_1 LnMix	2.64E-05	3.96E-05	5.12E-05	6.75E-05	8.85E-05	1.03E-04	2.90E-04	1.36E-04
H_2 GIG	5.26E-04	4.04E-04	4.82E-04	4.33E-04	4.91E-04	4.99E-04	1.58E-03	7.30E-04
H_2 LnMix	5.23E-04	4.98E-04	4.94E-04	5.50E-04	6.23E-04	6.37E-04	1.61E-03	7.00E-04
H_3 GIG	5.12E-04	4.93E-04	4.90E-04	5.46E-04	6.19E-04	6.34E-04	1.57E-03	6.98E-04
H_3 LnMix	5.15E-04	4.00E-04	4.78E-04	4.31E-04	4.89E-04	4.97E-04	1.55E-03	7.27E-04
H_4 GIG	0.0128	0.0140	0.0176	0.0181	0.0207	0.0222	0.0374	0.0294
H_4 LnMix	0.0127	0.0156	0.0177	0.0203	0.0233	0.0251	0.0376	0.0288

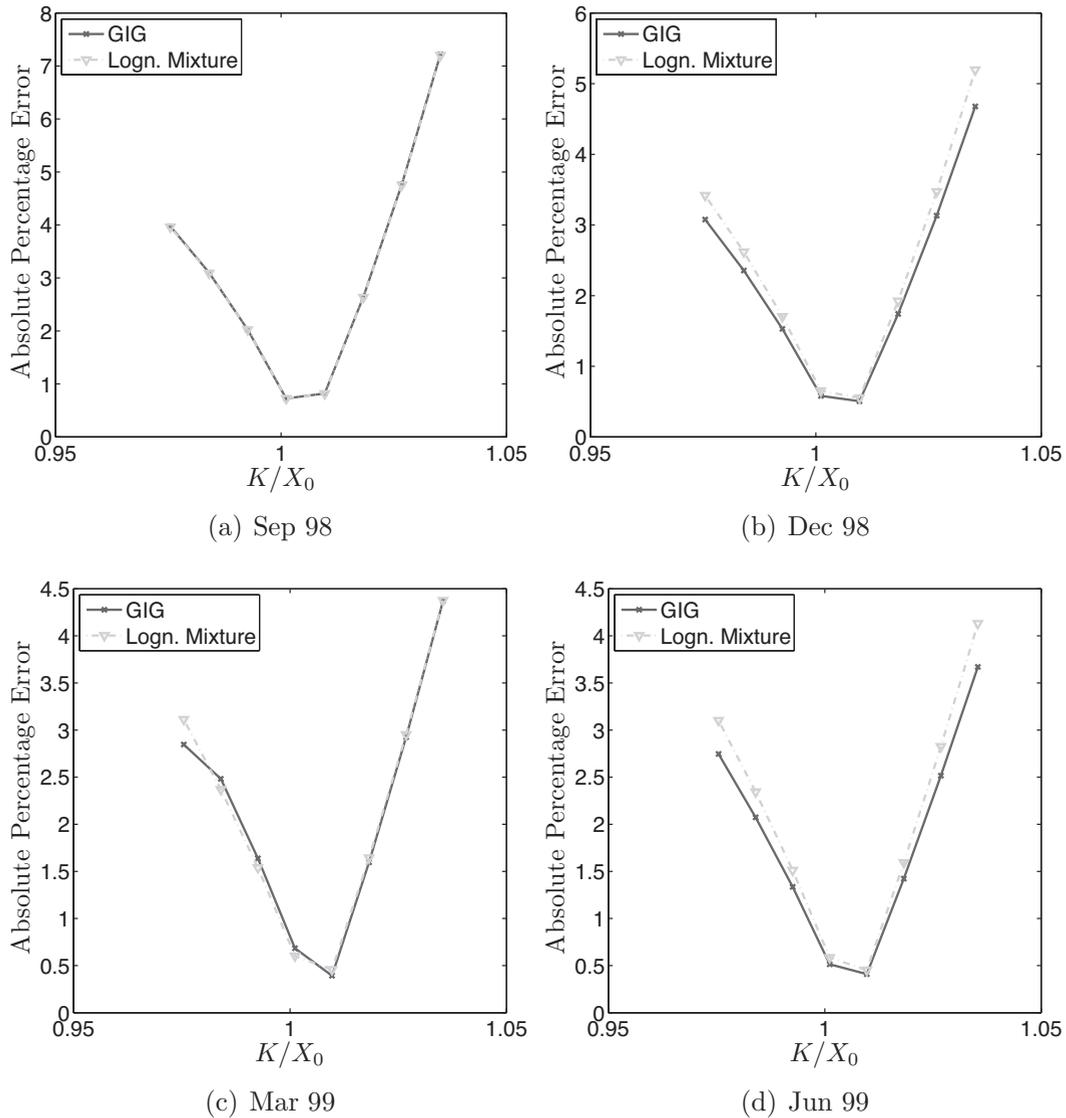


Figure 1 Absolute Percentage Errors for the First Four Maturities

FITTING THE MODELS TO DATA

For the RND estimation with a parametric model, the main elements are (1) formulas for pricing either European call or European put options, together with a formula for the forward price, (2) a minimization procedure for a non-linear function such as H_1 in the function given by (4), and (3) a set of market option prices.

Here we illustrate the calibration of the GIG and LnMix models using a dataset reported in Table 1, which is described in Rebonato (2004, pp. 290–291), and it is a typical example for the UK equity market.

The goodness of fit of the two models can be assessed to some extent from the results in Table 2, which reports the values obtained for the sum of squared residual $H_1(\hat{\theta}; \mathbf{m})$, where $\hat{\theta} = \arg \min_{\theta} H_1(\theta; \mathbf{m})$ and \mathbf{m} is the vector with

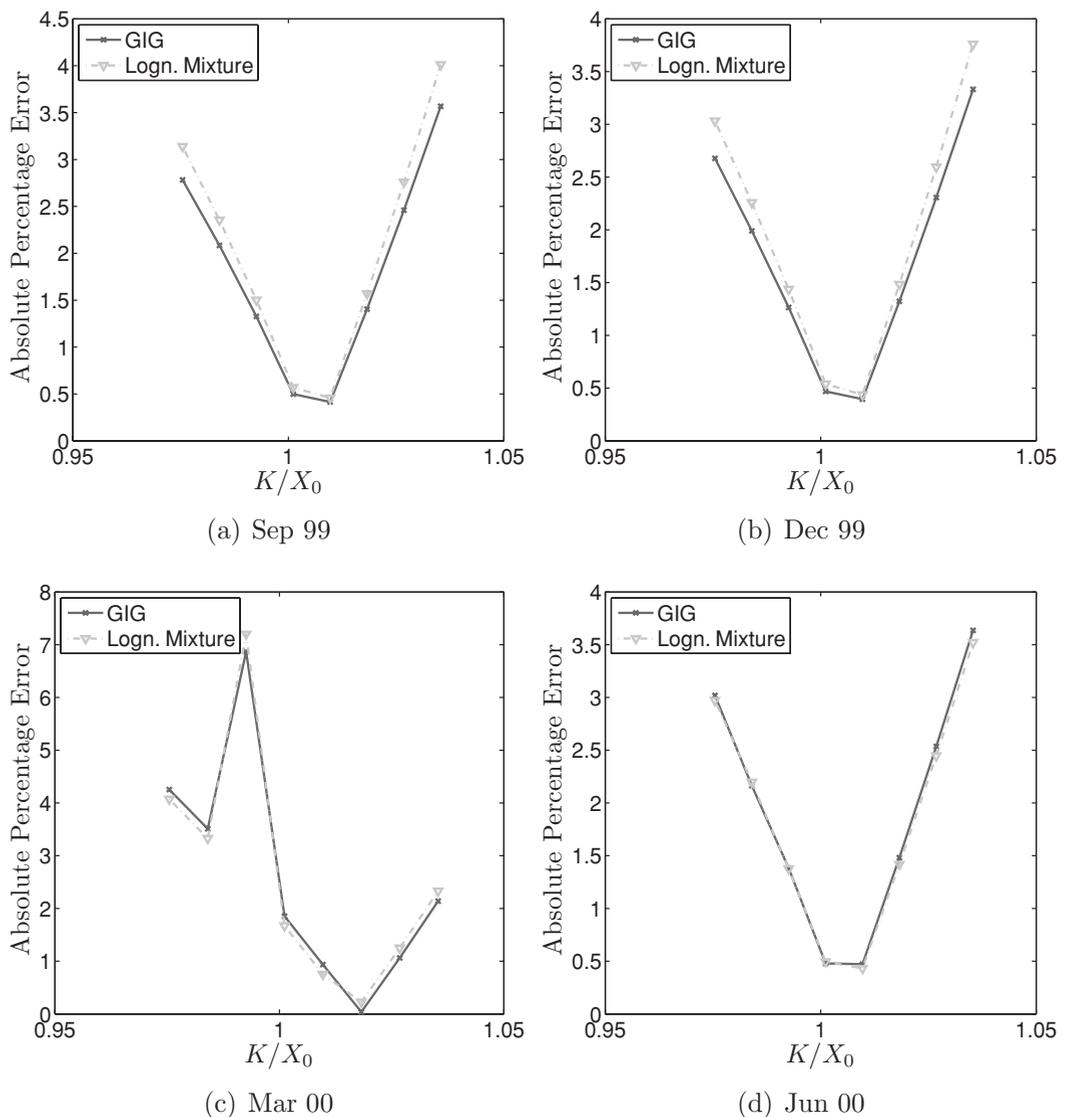


Figure 2 Absolute Percentage Errors for the Last Four Maturities

components K_j/X_0 reflecting moneyness. The smaller the value, the better is the fit. It is interesting that the GIG distribution, having three parameters, seems to calibrate across maturities very closely and is even a superior fit than the lognormal mixture (LnMix) model that uses five parameters.

A more informative comparison can be done by looking at the error structure versus moneyness. The fitting error for the two models and for each maturity are plotted in Figures 1 and 2 as the absolute percentage errors, defined for the European call option prices as $100 \times |C(\hat{\theta}; \mathbf{m}) - C^{mkt}(\mathbf{m})|/C^{mkt}(\mathbf{m})$, where $C(\hat{\theta}; \mathbf{m})$ is the same as the theoretical prices established in equation (1), which is calculated for the estimated parameter vector $\hat{\theta}$ following the minimization procedure focused on the function given in (4). In the neighborhood of at-the-money prices, the absolute percentage error is less than 1%, while out-of-the-money or in-the-money, it may go even higher.

Which parametric model to use depends on the task at hand. It is possible that some parametric models perform better for some asset classes (such as foreign exchange), while other models perform better for different asset classes (such as equity). Some models may have a superior fit in the tails.

KEY POINTS

- The information contained in the risk-neutral density is useful to many participants in financial markets. Central banks use this information in monitoring the stability of the financial system and for assessing the impact of new policies, and banks use it for marking positions in exotic derivatives that they hold.
- To recover the RND, a bundle of market prices for European vanilla call and put options on the same underlying asset and with the same maturity is used.
- Parametric and nonparametric models have been proposed for estimating the RND. For

several reasons, in practice, parametric models are better to employ. The main elements of a parametric model to estimate RND are an option pricing formula combined with a forward price formula, a minimization procedure, and a database of observed option prices.

- RND estimation can be done easily with parametric models for which pricing formulas are available for European vanilla options. The generalized inverse Gaussian model and the lognormal mixture model are examples of such models.
- The calibration is done by minimizing a discrepancy measure between the theoretical model prices and the observed option market prices.

NOTE

1. $I_A(x)$ is the indicator function being equal to 1 when $x \in A$ and zero otherwise.

REFERENCES

- Ait-Sahalia, Y., and J. Duarte. (2003). Nonparametric option pricing under shape restrictions. *Journal of Econometrics* 116 : 9–47.
- Ait-Sahalia, Y., and Lo, A. W. (1998). Nonparametric estimation of state-price densities implicit in financial asset prices. *Journal of Finance* 53: 499–547.
- Albota, G., Fabozzi, F., and Tunaru, R. (2009). Estimating risk-neutral density with parametric models in interest rate markets. *Quantitative Finance* 9: 55–70.
- Anagnou, I., Bedendo, M., Hodges, S., and Tompkins, R. (2005). The relation between implied and realized probability density functions. *Review of Futures Markets* 11: 41–66.
- Avellaneda, M. (1998). Minimum-relative-entropy calibration of asset-pricing models. *International Journal of Theoretical & Applied Finance* 1: 447–472.
- Bahra, B. (2007). Implied risk-neutral probability density functions from option prices: Theory and application. Report ISSN 1368-5562, Bank of England, London, EC2R 8AH.
- Bibby, B. M., and Sorensen, M. (2003). Hyperbolic processes in finance. In *Handbook of Heavy-Tailed*

- Distributions in Finance*, ed. S. T. Rachev. Chichester: Elsevier-North Holland, pp. 211–244.
- Brunner, B., and Hafner, R. (2003). Arbitrage-free estimation of the risk-neutral density from the implied volatility smile. *Journal of Computational Finance* 7: 75–106.
- Buchen, P. W., and Kelly, M. (1996). The maximum entropy distribution of an asset inferred from option prices. *Journal of Financial & Quantitative Analysis* 31: 143–159.
- Corrado, C. J. (2001). Option pricing based on the generalized lambda distribution. *Journal of Futures Markets* 21: 213–236.
- Corrado, C. J., and Su, T. (1997). Implied volatility skews and stock index skewness and kurtosis implied by S&P 500 index option prices. *Journal of Derivatives* 4: 8–19.
- Cox, J., and Ross, S. (1976). The valuation of options for alternative stochastic processes. *Journal of Financial Economics* 3: 145–166.
- Dutta, K. K., and Babbel, D. F. (2005). Extracting probabilistic information from the prices of interest rate options: Tests of distributional assumptions. *Journal of Business* 78: 841–870.
- Gemmill, G., and Saflekos, G. (2000). How useful are implied distributions? Evidence from stock-index options. *Journal of Derivatives* 7: 83–98.
- Jackwerth, J. C., and Rubinstein, M. (1996). Recovering probability distributions from options prices. *Journal of Finance* 51: 1611–1631.
- Melick, W. R., and Thomas, C. P. (1997). Recovering an asset's implied pdf from option prices: An application to crude oil during the Gulf crisis. *Journal of Financial & Quantitative Analysis* 32: 91–115.
- Paolella, M. (2007). *Intermediate Probability: A Computational Approach*. Chichester: John Wiley & Sons.
- Rebonato, R. (2004). *Volatility and Correlation*, 2nd ed., Chichester: John Wiley & Sons.
- Savickas, R. (2005). Evidence on delta hedging and implied volatilities for the Black-Scholes, Gamma and Weibull option-pricing models. *Journal of Financial Research* 28: 299–317.
- Shimko, D. (1993). Bounds of probability. *RISK* 6: 33–37.

Financial Statement Analysis

Financial Statements

PAMELA P. DRAKE, PhD, CFA

J. Gray Ferguson Professor of Finance, College of Business, James Madison University

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Much of the financial data that is used in financial modeling is drawn from the company's financial statements. The four basic financial statements are the balance sheet, the income statement, the statement of cash flows, and the statement of shareholders' equity. It is important to understand these data so that the information conveyed by them is interpreted properly in financial modeling. The financial statements are created using several assumptions that affect how to use and interpret the financial data.

Financial statements are summaries of the operating, financing, and investment activities of a business. Financial statements should provide information useful to both investors and creditors in making credit, investment, and other business decisions. And this usefulness means that investors and creditors can use these statements to predict, compare, and evaluate the amount, timing, and uncertainty of future cash flows.¹ In other words, *financial statements* provide the information needed to assess a company's future earnings and, therefore, the cash flows expected to result from those earnings.

Information from financial statements is typically used in financial modeling for forecasting and valuation purposes. In this entry, we discuss the general principles that guide the preparation of financial statements (generally

accepted accounting principles), the four basic financial statements (balance sheet, income statement, statement of cash flows, and statement of shareholders' equity), and the assumptions underlying the preparation of financial statements.

ACCOUNTING PRINCIPLES

The accounting data in financial statements are prepared by the firm's management according to a set of standards, referred to as *generally accepted accounting principles* (GAAP). Generally accepted accounting principles consist of the FASB Accounting Standards Codification, and, for publicly-traded companies, the rules and releases of the Securities and Exchange Commission.²

The financial statements of a company whose stock is publicly traded must, by law, be audited at least annually by independent public accountants (i.e., accountants who are not employees of the firm). In such an audit, the accountants examine the financial statements and the data from which these statements are prepared and attest—through the published auditor’s opinion—that these statements have been prepared according to GAAP. The auditor’s opinion focuses on whether the statements conform to GAAP and that there is adequate disclosure of any material change in accounting principles.

The financial statements are created using several assumptions that affect how we use and interpret the financial data:

- *Transactions are recorded at historical cost.* Therefore, the values shown in the statements are not market or replacement values, but rather reflect the original cost (adjusted for depreciation in the case of depreciable assets).
- *The appropriate unit of measurement is the dollar.* While this seems logical, the effects of inflation, combined with the practice of recording values at historical cost, may cause problems in using and interpreting these values.
- *The statements are recorded for predefined periods of time.* Generally, statements are produced to cover a chosen fiscal year or quarter, with the income statement and the statement of cash flows spanning a period’s time and the balance sheet and statement of shareholders’ equity as of the end of the specified period. But because the end of the fiscal year is generally chosen to coincide with the low point of activity in the operating cycle, the annual balance sheet and statement of shareholders’ equity may not be representative of values for the year.
- *Statements are prepared using accrual accounting and the matching principle.* Most businesses use accrual accounting, where income and revenues are matched in timing such that income is recorded in the period in which it is earned and expenses are reported in the period in which they are incurred in an attempt to generate revenues. The result of the use of accrual accounting is that reported income does not necessarily coincide with cash flows. Because the financial analyst is concerned ultimately with cash flows, he or she often must understand how reported income relates to a company’s cash flows.
- *It is assumed that the business will continue as a going concern.* The assumption that the business enterprise will continue indefinitely justifies the appropriateness of using historical costs instead of current market values because these assets are expected to be used up over time instead of sold.
- *Full disclosure requires providing information beyond the financial statements.* The requirement that there be full disclosure means that, in addition to the accounting numbers for such accounting items as revenues, expenses, and assets, narrative and additional numerical disclosures are provided in notes accompanying the financial statements. An analysis of financial statements is, therefore, not complete without this additional information.
- *Statements are prepared assuming conservatism.* In cases in which more than one interpretation of an event is possible, statements are prepared using the most conservative interpretation.

The financial statements and the auditors’ findings are published in the firm’s annual and quarterly reports sent to shareholders and the 10-K and 10-Q filings with the Securities and Exchange Commission (SEC). Also included in the reports, among other items, is a discussion by management, providing an overview of company events. The annual reports are much more detailed and disclose more financial information than the quarterly reports.

INFORMATION CONVEYED BY THE BASIC FINANCIAL STATEMENTS

In this section we will discuss the four basic financial statements and the information that they convey.

The Balance Sheet

The *balance sheet* is a report of the assets, liabilities, and equity of a firm at a point in time, generally at the end of a fiscal quarter or fiscal year. *Assets* are resources of the business enterprise, which are comprised of current or long-lived assets. How did the company finance these resources? It did so with liabilities and equity. *Liabilities* are obligations of the business enterprise that must be repaid at a future point in time, whereas equity is the ownership interest of the business enterprise. The relation between assets, liabilities and equity is simple, as reflected in the balance of what is owned and how it is financed, referred to as the accounting identity:

$$\text{Assets} = \text{Liabilities} + \text{Equity}$$

Assets

Assets are anything that the company owns that has a value. These assets may have a physical existence or not. Examples of physical assets include inventory items held for sale, office furniture, and production equipment. If an asset does not have a physical existence, we refer to it as an intangible asset, such as a trademark or a patent. You cannot see or touch an intangible asset, but it still contributes value to the company.

Assets may also be current or long-term, depending on how fast the company would be able to convert them into cash. Assets are generally reported in the balance sheet in order of liquidity, with the most liquid asset listed first and the least liquid listed last.

The most liquid assets of the company are the current assets. Current assets are assets that can be turned into cash in one operating cycle or one year, whichever is longer. This contrasts with the noncurrent assets, which cannot be liquidated quickly.

There are different types of current assets. The typical set of current assets is the following:

- Cash, bills, and currency are assets that are equivalent to cash (e.g., bank account).
- Marketable securities, which are securities that can be readily sold.
- Accounts receivable, which are amounts due from customers arising from trade credit.
- Inventories, which are investments in raw materials, work-in-process, and finished goods for sale.

A company's need for current assets is dictated, in part, by its operating cycle. The operating cycle is the length of time it takes to turn the investment of cash into goods and services for sale back into cash in the form of collections from customers. The longer the operating cycle, the greater a company's need for liquidity. Most firms' operating cycle is less than or equal to one year.

Noncurrent assets comprise both physical and nonphysical assets. Plant assets are physical assets, such as buildings and equipment, and are reflected in the balance sheet as gross plant and equipment and net plant and equipment. Gross plant and equipment, or gross property, plant, and equipment, is the total cost of investment in physical assets; that is, what the company originally paid for the property, plant, and equipment that it currently owns. Net plant and equipment, or net property, plant, and equipment, is the difference between gross plant and equipment and accumulated depreciation, and represents the book value of the plant and equipment assets. Accumulated depreciation is the sum of depreciation taken for

physical assets in the firm's possession. Therefore,

$$\begin{aligned} & \text{Gross plant and equipment} \\ & - \text{Accumulated depreciation} \\ \hline & = \text{Net plant and equipment} \end{aligned}$$

Companies may present just the net plant and equipment figure on the balance sheet, placing the detail with respect to accumulated depreciation in a footnote. Interpreting financial statements requires knowing a bit about how assets are depreciated for financial reporting purposes. Depreciation is the allocation of the cost of an asset over its useful life (or economic life). In the case of the fictitious Sample Company, whose balance sheet is shown in Table 1, the

original cost of the fixed assets (i.e., plant, property, and equipment)—less any write-downs for impairment—for year 2 is \$900 million. The accumulated depreciation for Sample in Year 1 is \$250 million; this means that the total depreciation taken on existing fixed assets over time is \$270 million. The net property, plant, and equipment account balance is \$630 million. This is also referred to as the book value or carrying value of these assets.

Intangible assets are assets that are not financial instruments, yet have no physical existence, such as patents, trademarks, copyrights, franchises, and formulas. Intangible assets may be amortized over some period, which is akin to depreciation. Keep in mind that a company may own a number of intangible assets that are not reported on the balance sheet. A company may only include an intangible asset's value on its balance sheet if (1) there are likely future benefits attributable specifically to the asset, and (2) the cost of the intangible asset can be measured.

Suppose a company has an active, ongoing investment in research and development to develop new products. It must expense what is spent on research and development each year because a given investment in R&D does not likely meet the two criteria because it is not until much later, after the R&D expense is made, that the economic viability of the investment is determined. If, on the other hand, a company buys a patent from another company, this cost may be capitalized and then amortized over the remaining life of the patent. So when you look at a company's assets on its balance sheet, you may not be getting the complete picture of what it owns.

Liabilities

We generally use the terms "liability" and "debt" as synonymous terms, though "liability" is actually a broader term, encompassing not only the explicit contracts that a company has, in terms of short-term and long-term debt obligations, but also obligations that are not specified in a contract, such as environmental

Table 1 The Sample Company Balance Sheet for Years 1 and 2 (in millions)

	Year 2	Year 1
Cash	\$40	\$30
Accounts receivable	100	90
Inventory	180	200
Other current assets	10	10
TOTAL CURRENT ASSETS	\$350	\$330
Property, plant, and equipment	\$900	\$800
Less accumulated depreciation	270	200
Net property, plant, and equipment	630	600
Intangible assets	20	20
TOTAL ASSETS	\$1,000	\$950
Accounts payable	\$150	\$140
Current maturities of long-term debt	60	40
TOTAL CURRENT LIABILITIES	\$180	\$165
Long-term debt	300	250
TOTAL LIABILITIES	\$380	\$325
Minority interest	30	15
Common stock	50	50
Additional paid-in capital	100	100
Retained earnings	500	400
TOTAL SHAREHOLDERS' EQUITY	650	550
TOTAL LIABILITIES AND SHAREHOLDERS' EQUITY	\$1,000	\$950

obligations or asset retirement obligations. Liabilities may be interest-bearing, such as a bond issue, or noninterest-bearing, such as amounts due to suppliers.

In the balance sheet, liabilities are presented in order of their due date and are often presented in two categories, current liabilities and long-term liabilities. Current liabilities are obligations due within one year or one operating cycle (whichever is longer). Current liabilities consist of:

- Accounts payable are amounts due to suppliers for purchases on credit.
- Wages and salaries payable are amounts due employees.
- Current portion of long-term indebtedness.
- Short-term bank loans.

Long-term liabilities are obligations that are due beyond one year. There are different types of long-term liabilities, including:

- Notes payables and bonds, which are indebtedness (loans) in the form of securities
- Capital leases, which are rental obligations that are long-term, fixed commitments
- Asset retirement liability, which is the contractual or statutory obligation to retire or decommission the asset and restore the site to required standards at the end of the asset's life
- Deferred taxes, which are taxes that may have to be paid in the future that are currently not due, though they are expensed for financial reporting purposes. Deferred taxes arise from differences between accounting and tax methods (e.g., depreciation methods).

Note that although deferred income taxes are often referred to as liabilities, some analysts will classify them as equity if the deferral is perceived to be perpetual. For example, a company that buys new depreciable assets each year will always have some level of deferred taxes; in

that case, an analyst will classify deferred taxes as equity.

Equity

The equity of a company is the ownership interest. The book value of equity, which for a corporation is often referred to as shareholders' equity or stockholders' equity, is basically the amount that investors paid the company for their ownership interest, plus any earnings (or less any losses), and minus any distributions to owners. For a corporation, equity is the amount that investors paid the corporation for the stock when it was initially sold, plus or minus any earnings or losses, less any dividends paid. Keep in mind that for any company, the reported amount of equity is an accumulation over time since the company's inception (or incorporation, in the case of a corporation).

Shareholders' equity is the carrying or book value of the ownership of a company. Shareholders' equity comprises:

+ Par value	A nominal amount per share of stock (sometimes prescribed by law), or the stated value, which is a nominal amount per share of stock assigned for accounting purposes if the stock has no par value.
+ Additional paid-in-capital	Also referred to as capital surplus, the amount paid for shares of stock by investors in excess of par or stated value.
– Treasury stock	The accounting value of shares of the firm's own stock bought by the firm.
+ Retained earnings	The accumulation of prior and current periods' earnings and losses, less any prior or current periods' dividends.
± Accumulated comprehensive income or loss	The total amount of income or loss that arises from transactions that result in income or losses, yet are not reported through the income statement. Items giving rise to this income include foreign currency translation adjustments and unrealized gains or losses on available-for-sale investments.

= Shareholders' equity

A Note on Minority Interest

On many companies' consolidated financial statements, you will notice a balance sheet account titled *Minority Interest*. When a company owns a substantial portion of another company, the accounting principles require that the company consolidate that company's financial statements into its own. Basically what happens in consolidating the financial statements is that the parent company will add the accounts of the subsidiary to its accounts (i.e., subsidiary inventory + parent inventory = consolidated inventory).³ If the parent does not own 100% of the subsidiary's ownership interest, an account is created, referred to as minority interest, which reflects the amount of the subsidiary's assets not owned by the parent. This account will be presented between liabilities and equity on the consolidated balance sheet. Is it a liability or an equity account? It is neither.

A similar adjustment takes place on the income statement. The minority interest account on the income statement reflects the income (or loss) in proportion to the equity in the subsidiary not owned by the parent.

Structure of the Balance Sheet

Consider a simple balance sheet for the Sample Company shown in Table 1 for fiscal years Year 1 and Year 2. The most recent fiscal year's data is presented in the left-most column of data. Notice that the accounting identity holds; that is, total assets are equal to the sum of the total liabilities and the total shareholders' equity.

The Income Statement

The *income statement* is a summary of operating performance over a period of time (e.g., a fiscal quarter or a fiscal year). We start with the revenue of the company over a period of time and then subtract the costs and expenses related to that revenue. The bottom line of the income statement consists of the owners' earnings for the period. To arrive at this "bottom line," we need to compare revenues and expenses. The

basic structure of the income statement includes the following:

Sales or revenues	⇐ Represent the amount of goods or services sold, in terms of price paid by customers
Less: Cost of goods sold (or cost of sales)	⇐ The amount of goods or services sold, in terms of cost to the firm
Gross profit	⇐ The difference between sales and cost of goods sold
Less: Selling and general expenditures	⇐ Salaries, administrative, marketing expenditures, etc.
Operating profit	⇐ Income from operations (ignores effects of financing decisions and taxes); earnings before interest and taxes (EBIT), operating income, and operating earnings
Less: Interest expense	⇐ Interest paid on debt
Net income before taxes	⇐ Earnings before taxes
Less: Taxes	⇐ Taxes expense for the current period
Net income	⇐ Operating profit less financing expenses (e.g., interest) and taxes
Less: Preferred stock dividends	⇐ Dividends paid to preferred shareholders
Earnings available to common shareholders	⇐ Net income less preferred stock dividends; residual income

Though the structure of the income statement varies by company, the basic idea is to present the operating results first, followed by non-operating results. The cost of sales, also referred to as the cost of goods sold, is deducted from revenues, producing a gross profit; that is, a profit without considering all other, general operating costs. These general operating expenses are those expenses related to the support of the general operations of the company, which includes salaries, marketing costs, and research and development. Depreciation, which is the amortized cost of physical assets, is also deducted from gross profit. The amount of the depreciation expense represents the cost of the wear and tear on the property, plant, and equipment of the company.

Table 2 The Sample Company Income Statement for Year 2 (in millions)

Sales	\$1,000
Cost of goods sold	600
Gross profit	\$400
Depreciation	50
Selling, general, and administrative expenses	160
Operating profit	\$190
Interest expense	23
Income before taxes	\$167
Taxes	67
Net income	\$100

Once we have the operating income, we have summarized the company's performance with respect to the operations of the business. But there is generally more to a company's performance. From operating income, we deduct interest expense and add any interest income. Further, adjustments are made for any other income or cost that is not a part of the company's core business.

There are a number of other items that may appear as adjustments to arrive at net income. One of these is extraordinary items, which are defined as unusual and infrequent gains or losses. Another adjustment would be for the expense related to the write-down of an asset's value.

In the case of the Sample Company, whose income statement is presented in Table 2, the income from operations—its core business—is \$190 million, whereas the net income (i.e., the "bottom line") is \$100 million.

Earnings Per Share

Companies provide information on *earnings per share* (EPS) in their annual and quarterly financial statement information, as well as in their periodic press releases. Generally, EPS is calculated as net income, divided by the number of shares outstanding. Companies must report both basic and diluted earnings per share.

Basic earnings per share is net income (minus preferred dividends), divided by the average number of shares outstanding. Diluted earnings

per share is net income (minus preferred dividends), divided by the number of shares outstanding considering all dilutive securities (e.g., convertible debt, options). Diluted earnings per share, therefore, gives the shareholder information about the potential dilution of earnings. For companies with a large number of dilutive securities (e.g., stock options, convertible preferred stock or convertible bonds), there can be a significant difference between basic and diluted EPS. You can see the effect of dilution by comparing the basic and diluted EPS.

More on Depreciation

There are different methods that can be used to allocate an asset's cost over its life. Generally, if the asset is expected to have value at the end of its economic life, the expected value, referred to as a salvage value (or residual value), is not depreciated; rather, the asset is depreciated down to its salvage value. There are different methods of depreciation that we classify as either straight-line or accelerated. Straight-line depreciation allocates the cost (less salvage value) in a uniform manner (equal amount per period) throughout the asset's life. Accelerated depreciation allocates the asset's cost (less salvage value) such that more depreciation is taken in the earlier years of the asset's life. There are alternative accelerated methods available, including:

- Declining balance method, in which a constant rate is applied to a declining amount (the undepreciated cost)
- Sum-of-the-years' digits method, in which a declining rate is applied to the asset's depreciable basis

Another method is the units-of-activity method, in which the useful life is defined in terms of a measure of units of production or some other metric or use (e.g., hours, miles). The depreciation expense in any period is determined as the usage in that period.

A common declining balance method is the double-declining balance method (DDB), which applies the rate that is twice that of the straight-line rate. In this case, the straight-line rate is 10% per year; therefore, the declining balance rate is 20% per year. We apply this rate of 20% against the original cost of \$1,000,000, resulting in a depreciation expense in the first year of \$200,000. In the second year, we apply this 20% against the undepreciated balance of $\$1,000,000 - 200,000 = \$800,000$, resulting in a depreciation of \$160,000.

Because the declining balance methods result in more depreciation sooner, relative to straight-line, and lower depreciation in the later years, companies may switch to straight-line in these later years. The same amount is depreciated over the life of the asset, but the pattern—and depreciation's impact on earnings—is modified slightly. In the case of the declining balance method, salvage value is not considered in the calculation of depreciation until the undepreciated balance reaches the salvage value.

For this same asset, the sum-of-the-years' digits (SYD) depreciation for the first year is the rate of $10/55$, or 18.18%, applied against the depreciable basis of $\$1,000,000 - 100,000 = \$900,000$:

$$\text{SYD first year} = \$900,000(10/55) = \$163,636$$

We calculate the denominator as the "sum of the years": $10 + 9 + 8 + 7 + 6 + 5 + 4 + 3 + 2 + 1 = 55$. In the second year, the rate is $9/55$ applied against the \$900,000, and so on.

Accelerated methods result in higher depreciation expenses in earlier years, relative to straight-line, as can be seen in Figure 1. As a result, accelerated methods result in lower reported earnings in earlier years, relative to straight-line. When comparing companies, it is important to understand whether the companies use different methods of depreciation because the choice of depreciation method affects both the balance sheet (through the carrying value of the asset) and the income statement (through the depreciation expense).

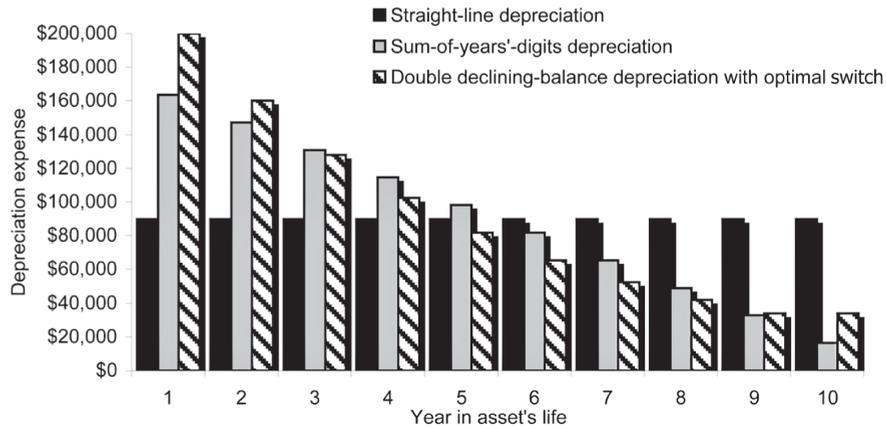
A major source of deferred income taxes and deferred tax assets is the accounting methods used for financial reporting purposes and tax purposes. In the case of financial accounting purposes, the company chooses the method that best reflects how its assets lose value over time, though most companies use the straight-line method. However, for tax purposes the company has no choice but to use the prescribed rates of depreciation, using the Modified Accelerated Cost Recovery System (MACRS). For tax purposes, a company does not have discretion over the asset's depreciable life or the rate of depreciation—they must use the MACRS system.

The MACRS system does not incorporate salvage value and is based on a declining balance system. The depreciable life for tax purposes may be longer than or shorter than that used for financial reporting purposes. For example, the MACRS rate for 3- and 5-year assets are as follows:

Year	3-year	5-year
1	33.33%	20.00%
2	44.45%	32.00%
3	14.81%	19.20%
4	7.41%	11.52%
5		11.52%
6		5.76%

You'll notice the fact that a 3-year asset is depreciated over four years and a 5-year asset is depreciated over six years. That is the result of using what is referred to as a half-year convention—using only half a year's worth of depreciation in the first year of an asset's life. This system results in a leftover amount that must still be depreciated in the last year (i.e., the fourth year in the case of a 3-year asset and the sixth year in the case of a 5-year asset). We provide a comparison of straight-line and MACRS depreciation in Figure 2. You can see that the methods produce different depreciation expenses, which result in the different income amounts for tax and financial reporting purposes.

Panel A: Depreciation Expense



Panel B: Book Value of the Asset

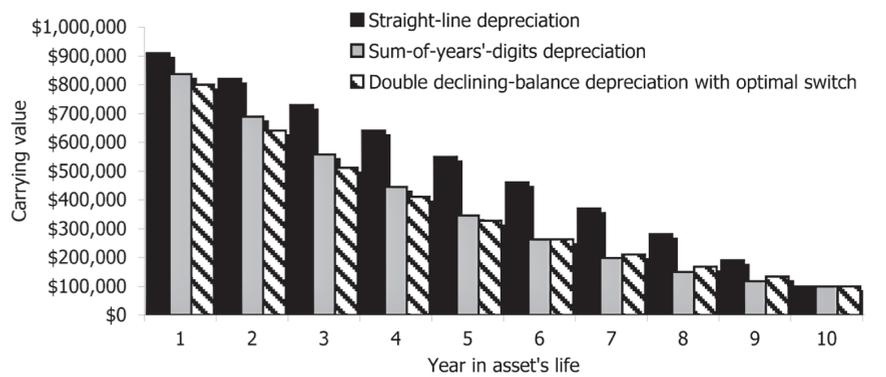


Figure 1 Comparison of Depreciation Expense and Book Value
 Depreciation expense each year for an asset with an original cost of \$1,000,000, a salvage value of \$10,000, and a 10-year useful life

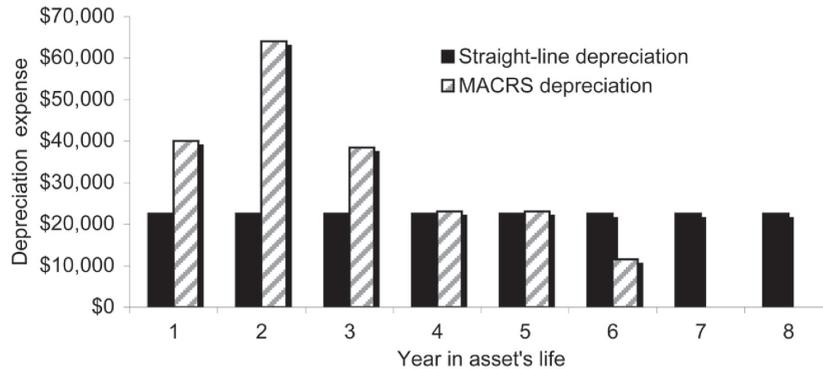
The Statement of Cash Flows

The *statement of cash flows* is the summary of a firm’s cash flows, summarized by operations, investment activities, and financing activities. A simplified cash flow statement is provided in Table 3 for the fictitious Sample Company. Cash flow from operations is cash flow from day-to-day operations. Cash flow from operating activities is basically net income adjusted for (1) noncash expenditures, and (2) changes in working capital accounts. The adjustment for

changes in working capital accounts is necessary to adjust net income that is determined using the accrual method to a cash flow amount. Increases in current assets and decreases in current liabilities are positive adjustments to arrive at the cash flow; decreases in current assets and increases in current liabilities are negative adjustments to arrive at the cash flow.

Cash flow for/from investing is the cash flows related to the acquisition (purchase) of plant, equipment, and other assets, as well as the

Panel A: Depreciation Expense



Panel B: Carrying Value

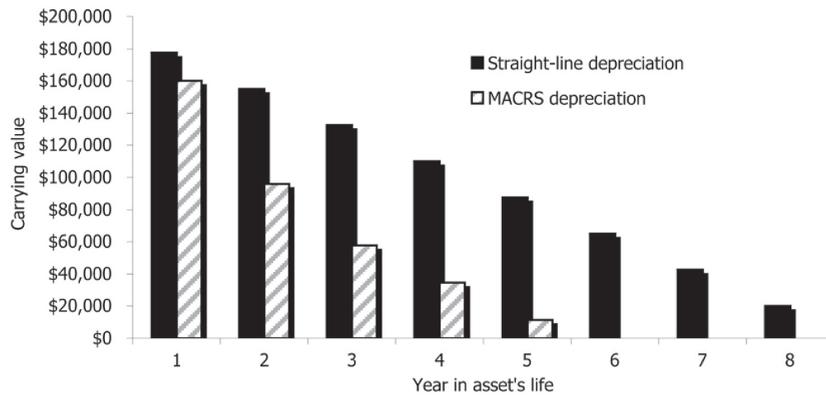


Figure 2 Depreciation for Financial Accounting Purposes versus Tax Purposes

Consider an asset that costs \$200,000 and has a salvage value of \$20,000. If the asset has a useful life of 8 years, but is classified as a 5-year asset for tax purposes, the depreciation and book value of the asset will be different between the financial accounting records and the tax records

proceeds from the sale of assets. Cash flow for/ from financing activities is the cash flow from activities related to the sources of capital funds (e.g., buy back common stock, pay dividends, issue bonds).

Not all of the classifications required by accounting principles are consistent with the true flow for the three types of activities. For example, interest expense is a financing cash flow, yet it affects the cash flow from operating activities because it is a deduction to arrive at net income. This inconsistency is also the case for interest

income and dividend income, both of which result from investing activities, but show up in the cash flow from operating activities through their contribution to net income.

The sources of a company's cash flows can reveal a great deal about the company and its prospects. For example, a financially healthy company tends to consistently generate cash flows from operations (that is, positive operating cash flows) and invests cash flows (that is, negative investing cash flows). To remain viable, a company must be able to generate funds

Table 3 The Sample Company Statement of Cash Flows, for the period ending December 31, 2006 (in millions)

Net income	\$100	
Add depreciation	50	
Subtract increase in accounts receivable	-10	
Add decrease in inventory	20	
Add increase in accounts payable	50	
Cash flow from operations		\$210
Retire debt	-\$100	
Cash flow for financing		-100
Purchase of equipment	-\$100	
Cash flow for investment		-100
Change in cash flow		\$10

from its operations; to grow, a company must continually make capital investments.

The change in cash flow—also called net cash flow—is the bottom line in the statement of cash flows and is equal to the change in the cash account as reported on the balance sheet. For the Sample Company, shown in Table 3, the net change in cash flow is a positive \$10 million; this is equal to the change in the cash account from \$50 million in Year 1 to \$60 million in Year 2.

By studying the cash flows of a company over time, we can gauge a company's financial health. For example, if a company relies on external financing to support its operations (that is, reliant on cash flows from financing and not from operations) for an extended period of time, this is a warning sign of financial trouble up ahead.

The Statement of Stockholders' Equity

The *statement of stockholders' equity* (also referred to as the *statement of shareholders' equity*) is a summary of the changes in the equity accounts, including information on stock options exercised, repurchases of shares, and Treasury shares. The basic structure is to include a reconciliation of the balance in each component of equity from the beginning of the fiscal year with the end of the fiscal year, detailing changes

attributed to net income, dividends, purchases or sales of Treasury stock. The components are common stock, additional paid-in capital, retained earnings, and Treasury stock. For each of these components, the statement begins with the balance of each at the end of the previous fiscal period and then adjustments are shown to produce the balance at the end of the current fiscal period.

In addition, there is a reconciliation of any gains or losses that affect stockholders' equity but which do not flow through the income statement, such as foreign-currency translation adjustments and unrealized gains on investments. These items are of interest because they are part of comprehensive income, and hence income to owners, but they are not represented on the company's income statement.

Why Bother About the Footnotes?

Footnotes to the financial statements contain additional information, supplementing or explaining financial statement data. These notes are presented in both the annual report and the 10-K filing (with the SEC), though the latter usually provides a greater depth of information.

The footnotes to the financial statements provide information pertaining to:

- *The significant accounting policies and practices that the company uses.* This helps the analyst with the interpretation of the results, comparability of the results to other companies and to other years for the same company, and in assessing the quality of the reported information.
- *Income taxes.* The footnotes tell us about the company's current and deferred income taxes, breakdowns by the type of tax (e.g., federal versus state), and the effective tax rate that the company is paying.
- *Pension plans.* The detail about pension plans, including the pension assets and the pension liability, is important in determining whether a company's pension plan is overfunded or underfunded.

- *Leases*. You can learn about both the capital leases, which are the long-term lease obligations that are reported on the balance sheet, and about the future commitments under operating leases, which are not reflected on the balance sheet.
- *Long-term debt*. You can find detailed information about the maturity dates and interest rates on the company's debt obligations.

The phrase “the devil is in the details” applies aptly to the footnotes of a company's financial statement. Through the footnotes, a company is providing information that is crucial in analyzing a company's financial health and performance. If footnotes are vague or confusing, as they were in the case of Enron prior to the break in the scandal, the analyst must ask questions to help understand this information.

ACCOUNTING FLEXIBILITY

The generally accepted accounting principles provide some choices in the manner in which some transactions and assets are accounted. For example, a company may choose to account for inventory, and hence costs of sales, using Last-in, First-out (LIFO) or First-in, First-out (FIFO). This is intentional because these principles are applied to a broad set of companies and no single set of methods offers the best representation of a company's condition or performance for all companies. Ideally, a company's management, in consultation with the accountants, chooses those accounting methods and presentations that are most appropriate for the company.

A company's management has always had the ability to manage earnings through the judicious choice of accounting methods within the GAAP framework. The company's “watchdogs” (i.e., the accountants) should keep the company's management in check. However, recent scandals have revealed that the watchdog function of the accounting firms was not working well. Additionally, some companies'

management used manipulation of financial results and outright fraud to distort the financial picture.

The Sarbanes-Oxley Act of 2002 offers some comfort in terms of creating the oversight board for the auditing accounting firms. In addition, the Securities and Exchange Commission, the Financial Accounting Standards Board, and the International Accounting Standards Board are tightening some of the flexibility that companies had in the past.

Pro Forma Financial Data

Pro forma financial information is really a misnomer—the information is neither pro forma (that is, forward looking), nor reliable financial data. What is it? Creative accounting. It started during the Internet-tech boom in the 1990s and persists today: Companies release financial information that is prepared according to its own liking, using accounting methods that they create.

Why did companies start doing this? What is wrong with generally accepted accounting principles (GAAP)? During the Internet-tech stock boom, many startup companies quickly went public and then felt the pressures to generate profits. However, profits in that industry were hard to come by during that period of time. What some companies did is generate financial data that they included in company releases that reported earnings not calculated using GAAP—but rather by methods of their own. In some cases, these alternative methods hid a lot of the ills of these companies.

The use of pro forma financial data may be helpful, but also may be misleading to investors. Analysts routinely adjust published financial statement data to remove unusual, nonrecurring items. This can give the analyst a better predictor of the continued performance of the company. So what is wrong with the company itself doing this? Nothing, unless it becomes misleading, such as a company including its nonrecurring gains, but not including its

nonrecurring losses. In concern for the possibility of misleading information being given to investors, the Securities and Exchange Commission now requires that if companies release pro forma financial data, they must also reconcile this data with GAAP.⁴

KEY POINTS

- There are four basic financial statements: the balance sheet, the income statement, the statement of cash flows, and the statement of stockholders' equity.
 - The balance sheet and the statement of shareholders' equity are statements with values of accounts at a point in time. In the case of the balance sheet, the company presents data as of the end of the most recent two years; in the case of the statement of shareholders' equity, from the latest fiscal year to the end. The income statement and the statement of cash flows provide data on earnings and cash flows over the period, whether that period is a fiscal quarter or year.
 - The information conveyed in the footnotes is essential to the understanding of financial statements. There is detail in these footnotes that gives us a better idea of the financial health of the company. The financial statements and the accompanying footnotes provide the accounting principles that guide companies in the preparation of financial statements.
- Not only must the accounting methods that a company uses be understood, but the choices that a company has made among the available accounting methods should be understood.

NOTES

1. The purpose, focus, and objectives of financial statements are detailed in Financial Accounting Standards Board (1978, 1980).
2. Effective July 1, 2009, Financial Accounting Standards Board (FASB) Accounting Standards Codification.
3. There are other adjustments made for intercorporate transactions, but we will not go into these in this entry.
4. Securities and Exchange Commission RIN3235-A169, "Conditions for Use of Non-GAAP Financial Measures," effective March 28, 2003.

REFERENCES

- Financial Accounting Standards Board (1978). Objectives of financial reporting by business enterprises. *Statement of Financial Accounting Concepts No. 1*. Stamford, CT: FASB.
- Financial Accounting Standards Board (1980). Qualitative characteristics of accounting information. *Statement of Financial Accounting Concepts No. 2*. Stamford, CT: FASB.
- Financial Accounting Standards Board (2009). *FASB Accounting Standards Codification*. Stamford, CT: FASB.

Financial Ratio Analysis

PAMELA P. DRAKE, PhD, CFA

J. Gray Ferguson Professor of Finance, College of Business, James Madison University

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Financial analysis involves the selection, evaluation, and interpretation of financial data and other pertinent information to assist in evaluating the operating performance and financial condition of a company. The operating performance of a company is a measure of how well a company has used its resources—its assets, both tangible and intangible—to produce a return on its investment. The financial condition of a company is a measure of its ability to satisfy its obligations, such as the payment of interest on its debt in a timely manner. The analyst has many tools available in the analysis of financial information. These tools include financial ratio analysis and quantitative analysis. The analyst must understand how to use these tools, along with economics and accounting information, in the most effective manner.

Financial analysis is one of the many tools useful in valuation because it helps analysts and investors gauge returns and risks. In this entry, we explain and illustrate financial ratios—one of the tools of financial analysis. In financial ratio analysis we select the relevant information—primarily the financial statement data—and evaluate it. We show how to incorporate market data and economic data in the analysis of financial ratios. Finally, we show how to interpret financial ratio analysis, identifying the pitfalls that occur when it's not done properly.

RATIOS AND THEIR CLASSIFICATION

A ratio is a mathematical relation between two quantities. Suppose you have 200 apples and

100 oranges. The ratio of apples to oranges is $200/100$, which we can conveniently express as 2:1 or 2. A financial ratio is a comparison between one bit of financial information and another. Consider the ratio of current assets to current liabilities, which we refer to as the current ratio. This ratio is a comparison between assets that can be readily turned into cash—current assets—and the obligations that are due in the near future—current liabilities. A current ratio of 2 or 2:1 means that we have twice as much in current assets as we need to satisfy obligations due in the near future.

Ratios can be classified according to the way they are constructed and the financial characteristic they are describing. For example, we will see that the current ratio is constructed as a coverage ratio (the ratio of current

assets—available funds—to current liabilities—the obligation) that we use to describe a firm's liquidity (its ability to meet its immediate needs).

There are as many different *financial ratios* as there are possible combinations of items appearing on the income statement, balance sheet, and statement of cash flows. We can classify ratios according to how they are constructed or according to the financial characteristic that they capture.

Ratios can be constructed in the following four ways:

1. As a *coverage ratio*. A coverage ratio is a measure of a firm's ability to "cover," or meet, a particular financial obligation. The denominator may be any obligation, such as interest or rent, and the numerator is the amount of the funds available to satisfy that obligation.
2. As a *return ratio*. A return ratio indicates a net benefit received from a particular investment of resources. The net benefit is what is left over after expenses, such as operating earnings or net income, and the resources may be total assets, fixed assets, inventory, or any other investment.
3. As a *turnover ratio*. A turnover ratio is a measure of how much a firm gets out of its assets. This ratio compares the gross benefit from an activity or investment with the resources employed in it.
4. As a *component percentage*. A component percentage is the ratio of one amount in a financial statement, such as sales, to the total of amounts in that financial statement, such as net profit.

In addition, we can also express financial data in terms of time—say, how many days' worth of inventory we have on hand—or on a per-share basis—say, how much a firm has earned for each share of common stock. Both are measures we can use to evaluate operating performance or financial condition.

When we assess a firm's operating performance, a concern is whether the company is

applying its assets in an efficient and profitable manner. When an analyst assesses a firm's financial condition, a concern is whether the company is able to meet its financial obligations. The analyst can use financial ratios to evaluate five aspects of operating performance and financial condition:

1. Return on investment
2. Liquidity
3. Profitability
4. Activity
5. Financial leverage

There are several ratios reflecting each of the five aspects of a firm's operating performance and financial condition. We apply these ratios to the Fictitious Corporation, whose balance sheets, income statements, and statement of cash flows for two years are shown in Tables 1, 2, and 3, respectively. We refer to the most recent fiscal year for which financial statements

Table 1 Fictitious Corporation Balance Sheets for Years Ending December 31, in Thousands

	Current Year	Prior Year
ASSETS		
Cash	\$400	\$200
Marketable securities	200	0
Accounts receivable	600	800
Inventories	<u>1,800</u>	<u>1,000</u>
Total current assets	<u>\$3,000</u>	<u>\$2,000</u>
Gross plant and equipment	\$11,000	\$10,000
Accumulated depreciation	<u>(4,000)</u>	<u>(3,000)</u>
Net plant and equipment	7,000	7,000
Intangible assets	<u>1,000</u>	<u>1,000</u>
Total assets	<u>\$11,000</u>	<u>\$10,000</u>
LIABILITIES AND SHAREHOLDERS' EQUITY		
Accounts payable	\$500	\$400
Other current liabilities	500	200
Long-term debt	<u>4,000</u>	<u>5,000</u>
Total liabilities	<u>\$5,000</u>	<u>\$5,600</u>
Common stock, \$1 par value; Authorized 2,000,000 shares Issued 1,500,000 and 1,200,000 shares	1,500	1,200
Additional paid-in capital	1,500	800
Retained earnings	<u>3,000</u>	<u>2,400</u>
Total shareholders' equity	<u>6,000</u>	<u>4,400</u>
Total liabilities and shareholders' equity	<u>\$11,000</u>	<u>\$10,000</u>

Table 2 Fictitious Corporation Income Statements for Years Ending December 31, in Thousands

	Current Year	Prior Year
Sales	\$10,000	\$9,000
Cost of goods sold	(6,500)	(6,000)
Gross profit	\$3,500	\$3,000
Lease expense	(1,000)	(500)
Administrative expense	(500)	(500)
Earnings before interest and taxes (EBIT)	\$2,000	\$2,000
Interest	(400)	(500)
Earnings before taxes	\$1,600	\$1,500
Taxes	(400)	(500)
Net income	\$1,200	\$1,000
Preferred dividends	(100)	(100)
Earnings available to common shareholders	\$1,100	\$900
Common dividends	(500)	(400)
Retained earnings	\$600	\$500

are available as the “current year.” The “prior year” is the fiscal year prior to the current year.

The ratios we introduce here are by no means the only ones that can be formed using financial data, though they are some of the more commonly used. After becoming comfortable with the tools of financial analysis, an analyst will be able to create ratios that serve a particular evaluation objective.

RETURN-ON-INVESTMENT RATIOS

Return-on-investment ratios compare measures of benefits, such as earnings or net income, with measures of investment. For example, if an analyst wants to evaluate how well the firm uses its assets in its operations, he could calculate the *return on assets*—sometimes called the *basic earning power ratio*—as the ratio of earnings before interest and taxes (EBIT) (also known as *operating earnings*) to total assets:

$$\text{Basic earning power} = \frac{\text{Earnings before interest and taxes}}{\text{Total assets}}$$

Table 3 Fictitious Company Statement of Cash Flows, Years Ended December 31, in Thousands

	Current Year	Prior Year
Cash flow from (used for) operating activities		
Net income	\$1,200	\$1,000
Add or deduct adjustments to cash basis:		
Change in accounts receivables	\$200	\$(200)
Change in accounts payable	100	400
Change in marketable securities	(200)	200
Change in inventories	(800)	(600)
Change in other current liabilities	300	0
Depreciation	1,000	1,000
	<u>600</u>	<u>800</u>
Cash flow from operations	\$1,800	\$1,800
Cash flow from (used for) investing activities		
Purchase of plant and equipment	\$(1,000)	\$0
Cash flow from (used for) investing activities	\$(1,000)	\$0
Cash flow from (used for) financing activities		
Sale of common stock	\$1,000	\$0
Repayment of long-term debt	(1,000)	(1,500)
Payment of preferred dividends	(100)	(100)
Payment of common dividends	(500)	(400)
Cash flow from (used for) financing activities	(600)	(1,900)
Increase (decrease) in cash flow	\$200	\$(100)
Cash at the beginning of the year	200	300
Cash at the end of the year	\$400	\$200

For Fictitious Corporation, for the current year:

$$\begin{aligned} \text{Basic earning power} &= \frac{\$2,000,000}{\$11,000,000} \\ &= 0.1818 \text{ or } 18.18\% \end{aligned}$$

For every dollar invested in assets, Fictitious earned about 18 cents in the current year. This measure deals with earnings from operations; it does not consider how these operations are financed.

Another return-on-assets ratio uses net income—operating earnings less interest and taxes—instead of earnings before interest and taxes:

$$\text{Return on assets} = \frac{\text{Net income}}{\text{Total assets}}$$

(In actual application the same term, return on assets, is often used to describe both ratios. It is only in the actual context or through an examination of the numbers themselves that we know which return ratio is presented. We use two different terms to describe these two return-on-asset ratios in this entry simply to avoid any confusion.)

For Fictitious in the current year:

$$\begin{aligned}\text{Return on assets} &= \frac{\$1,200,000}{\$11,000,000} \\ &= 0.1091 \text{ or } 10.91\%\end{aligned}$$

Thus, without taking into consideration how assets are financed, the return on assets for Fictitious is 18%. Taking into consideration how assets are financed, the return on assets is 11%. The difference is due to Fictitious financing part of its total assets with debt, incurring interest of \$400,000 in the current year; hence, the return-on-assets ratio excludes taxes of \$400,000 in the current year from earnings in the numerator.

If we look at Fictitious's liabilities and equities, we see that the assets are financed in part by liabilities (\$1 million short term, \$4 million long term) and in part by equity (\$800,000 preferred stock, \$5.2 million common stock). Investors may not be interested in the return the firm gets from its total investment (debt plus equity), but rather shareholders are interested in the return the firm can generate on their investment. The *return on equity* is the ratio of the net income shareholders receive to their equity in the stock:

$$\begin{aligned}\text{Return on equity} & \\ &= \frac{\text{Net income}}{\text{Book value of shareholders' equity}}\end{aligned}$$

For Fictitious Corporation, there is only one type of shareholder: common. For the current year:

$$\begin{aligned}\text{Return on equity} &= \frac{\$1,200,000}{\$6,000,000} \\ &= 0.2000 \text{ or } 20.00\%\end{aligned}$$

Recap: Return-on-Investment Ratios

The return-on-investment ratios for Fictitious Corporation for the current year are:

$$\begin{aligned}\text{Basic earning power} &= 18.18\% \\ \text{Return on assets} &= 10.91\% \\ \text{Return on equity} &= 20.00\%\end{aligned}$$

These return-on-investment ratios indicate:

- Fictitious earns over 18% from operations, or about 11% overall, from its assets.
- Shareholders earn 20% from their investment (measured in book value terms).

These ratios do not provide information on:

- Whether this return is due to the profit margins (that is, due to costs and revenues) or to how efficiently Fictitious uses its assets.
- The return shareholders earn on their actual investment in the firm, that is, what shareholders earn relative to their actual investment, not the book value of their investment. For example, \$100 may be invested in the stock, but its value according to the balance sheet may be greater than or, more likely, less than \$100.

DuPont System

The returns on investment ratios provides a "bottom line" on the performance of a company, but do not tell us anything about the "why" behind this performance. For an understanding of the "why," an analyst must dig a bit deeper into the financial statements. A method that is useful in examining the source of performance is the DuPont system. The DuPont system is a method of breaking down return ratios into their components to determine which areas are responsible for a firm's performance. To see how it is used, let us take a closer look at the first definition of the return on assets:

$$\begin{aligned}\text{Basic earning power} & \\ &= \frac{\text{Earnings before interest and taxes}}{\text{Total assets}}\end{aligned}$$

Suppose the return on assets changes from 20% in one period to 10% the next period. We do not know whether this decreased return is due to a less efficient use of the firm's assets—that is, lower activity—or to less effective management of expenses (that is, lower profit margins). A lower return on assets could be due to lower activity, lower margins, or both. Because an analyst is interested in evaluating past operating performance to evaluate different aspects of the management of the firm and to predict future performance, knowing the source of these returns is valuable.

Let us take a closer look at the return on assets and break it down into its components: measures of activity and profit margin. We do this by relating both the numerator and the denominator to sales activity. Divide both the numerator and the denominator of the basic earning power by revenues:

$$\text{Basic earning power} = \frac{\text{Earnings before interest and taxes/Revenues}}{\text{Revenues total assets/Revenues}}$$

which is equivalent to:

$$\text{Basic earning power} = \left(\frac{\text{Earnings before interest and taxes}}{\text{Revenues}} \right) \left(\frac{\text{Revenues}}{\text{Revenues total assets}} \right)$$

This says that the earning power of the company is related to profitability (in this case, operating profit) and a measure of activity (total asset turnover).

$$\begin{aligned} \text{Basic earning power} &= (\text{Operating profit margin}) \\ &(\text{Total asset turnover}) \end{aligned}$$

When analyzing a change in the company's basic earning power, an analyst could look at this breakdown to see the change in its components: operating profit margin and total asset turnover.

This method of analyzing return ratios in terms of profit margin and turnover ratios, re-

ferred to as the DuPont System, is credited to the E.I. DuPont Corporation, whose management developed a system of breaking down return ratios into their components.

Let's look at the return on assets of Fictitious for the two years. Its returns on assets were 20% in the prior year and 18.18% in the current year. We can decompose the firm's returns on assets for the two years to obtain:

Year	Basic Earning Power	Operating Profit Margin	Total Asset Turnover
Prior	20.00%	22.22%	0.9000 times
Current	18.18	20.00	0.9091 times

We see that operating profit margin declined over the two years, yet asset turnover improved slightly, from 0.9000 to 0.9091. Therefore, the return-on-assets decline is attributable to lower profit margins.

The return on assets can be broken down into its components in a similar manner:

$$\text{Return on assets} = \left(\frac{\text{Net income}}{\text{Revenues}} \right) \left(\frac{\text{Revenues}}{\text{Revenues total assets}} \right)$$

or

$$\begin{aligned} \text{Return on assets} &= (\text{Net profit margin})(\text{Total asset turnover}) \end{aligned}$$

The basic earning power ratio relates to the return on assets. Recognizing that:

$$\text{Net income} = \text{Earnings before tax}(1 - \text{Tax rate})$$

then

$$\begin{aligned} \text{Net income} &= \text{Earnings before interest and taxes} \\ &\times \left(\frac{\text{Earnings before taxes}}{\text{Earnings before interest and taxes}} \right) (1 - \text{Tax rate}) \\ &\quad \uparrow \quad \quad \quad \uparrow \\ &\quad \text{equity's share of earnings} \quad \text{tax retention \%} \end{aligned}$$

The ratio of earnings before taxes to earnings before interest and taxes reflects the interest burden of the company, whereas the term $(1 - \text{tax rate})$ reflects the company's tax burden.

Therefore,

$$\begin{aligned} \text{Return on assets} &= \left(\frac{\text{Earnings before interest and taxes}}{\text{Revenues}} \right) \\ &\times \left(\frac{\text{Revenues}}{\text{Revenues total assets}} \right) \\ &\times \left(\frac{\text{Earnings before taxes}}{\text{Earnings before interest and taxes}} \right) \\ &\quad (1 - \text{Tax rate}) \end{aligned}$$

or

$$\begin{aligned} \text{Return on assets} &= (\text{Operating profit margin})(\text{Total asset turnover}) \\ &\times (\text{Equity's share of earnings})(\text{Tax retention \%}) \end{aligned}$$

The breakdown of a return-on-equity ratio requires a bit more decomposition because instead of total assets as the denominator, the denominator in the return is shareholders' equity. Because activity ratios reflect the use of all of the assets, not just the proportion financed by equity, we need to adjust the activity ratio by the proportion that assets are financed by equity (that is, the ratio of the book value of shareholders' equity to total assets):

$$\begin{aligned} \text{Return on equity} &= (\text{Return on assets}) \left(\frac{\text{Total assets}}{\text{Shareholder's equity}} \right) \\ &= \left(\frac{\text{Net income}}{\text{Revenues}} \right) \left(\frac{\text{Revenues}}{\text{Total assets}} \right) \left(\frac{\text{Total assets}}{\text{Shareholder's equity}} \right) \\ &\quad \uparrow \\ &\quad \text{Equity multiplier} \end{aligned}$$

The ratio of total assets to shareholders' equity is referred to as the *equity multiplier*. The equity multiplier, therefore, captures the effects of how a company finances its assets, referred to as its financial leverage. Multiplying the total asset turnover ratio by the equity multiplier allows us to break down the return-on-equity ratios into three components: profit margin, asset turnover, and financial leverage. For example, the return on equity can be broken down

into three parts:

$$\begin{aligned} \text{Return on equity} &= (\text{Net profit margin})(\text{Total asset turnover}) \\ &\quad (\text{Equity multiplier}) \end{aligned}$$

Applying this breakdown to Fictitious for the two years:

Year	Return on Equity	Net Profit Margin	Total Asset Turnover	Total Debt to Assets	Equity Multiplier
Prior	22.73%	11.11%	0.9000 times	56.00%	2.2727
Current	20.00	12.00	0.9091	45.45%	1.8332

The return on equity decreased over the two years because of a lower operating profit margin and less use of financial leverage.

The analyst can decompose the return on equity further by breaking out the equity's share of before-tax earnings (represented by the ratio of earnings before and after interest) and tax retention percentage. Consider the example in Figure 1, in which we provide a DuPont breakdown of the return on equity for Microsoft Corporation for the fiscal year ending June 30, 2006, in Panel A. The return on equity of 31.486% can be broken down into three and then five components, as shown in this figure. We can also use this breakdown to compare the return on equity for the 2005 and 2006 fiscal years, as shown in Panel B. As you can see, the return on equity improved from 2005 to 2006 and, using this breakdown, we can see that this was due primarily to the improvement in the asset turnover and the increased financial leverage.

This decomposition allows the analyst to take a closer look at the factors that are controllable by a company's management (e.g., asset turnover) and those that are not controllable (e.g., tax retention). The breakdowns lead the analyst to information on both the balance sheet and the income statement. And this is not the only breakdown of the return ratios—further decomposition is possible.

For the fiscal year ending June 30, 2006,

$$\text{Return on equity} = \frac{\text{Net income}}{\text{Total assets}} = \frac{\$12.599}{\$40.014} = 0.31486 \text{ or } 31.486\%$$

Breaking return on equity into three components:

$$\begin{aligned} \text{Return on equity} &= \frac{\text{Net income}}{\text{Revenues}} \times \frac{\text{Revenues}}{\text{Total assets}} \times \frac{\text{Total assets}}{\text{Shareholders' equity}} \\ &= \frac{\$12.599}{\$44.282} \times \frac{\$44.282}{\$69.597} \times \frac{\$69.597}{\$40.014} = 0.31486 \text{ or } 31.486\% \end{aligned}$$

Breaking the return on equity into five components:

$$\begin{aligned} \text{Return on equity} &= \left(\frac{\text{Earnings before interest and taxes}}{\text{Revenues}} \right) \times \left(\frac{\text{Earnings before taxes}}{\text{Earnings before interest and taxes}} \right) \times (1 - \text{Tax rate}) \times \left(\frac{\text{Revenues}}{\text{Total assets}} \right) \times \left(\frac{\text{Total assets}}{\text{Shareholders' equity}} \right) \\ \text{Return on equity} &= \left(\frac{\$18.262}{\$44.282} \right) \times \left(\frac{\$18.262}{\$18.262} \right) \times (1 - 0.31010) \times \left(\frac{\$44.282}{\$69.597} \right) \times \left(\frac{\$69.597}{\$40.014} \right) \\ &= 0.41240 \times 1.0 \times 0.68990 \times 0.63626 \times 1.73932 \\ &= 0.31486 \text{ or } 31.486\% \end{aligned}$$

Comparing the components between the June 30, 2006 fiscal year and the June 30, 2005 fiscal year,

$$\text{Return on equity} = \left(\frac{\text{Earnings before interest and taxes}}{\text{Revenues}} \right) \times \left(\frac{\text{Earnings before taxes}}{\text{Earnings before interest and taxes}} \right) \times (1 - \text{Tax rate}) \times \left(\frac{\text{Revenues}}{\text{Total assets}} \right) \times \left(\frac{\text{Total assets}}{\text{Shareholders' equity}} \right)$$

$$\text{Return on equity June 30, 2006} = 0.41240 \times 1.0 \times 0.68990 \times 0.63626 \times 1.73932 = 31.486\%$$

$$\text{Return on equity June 30, 2005} = 0.41791 \times 1.0 \times 0.73695 \times 0.56186 \times 1.47179 = 25.468\%$$

Figure 1 The DuPont System Applied to Microsoft Corporation

LIQUIDITY

Liquidity reflects the ability of a firm to meet its short-term obligations using those assets that are most readily converted into cash. Assets that may be converted into cash in a short period of time are referred to as *liquid assets*; they are listed in financial statements as current assets. Current assets are often referred to as *working capital*, since they represent the resources needed for the day-to-day operations of the firm's long-term capital investments. Current assets are used to satisfy short-term obligations, or current liabilities. The amount by which current assets exceed current liabilities is referred to as the *net working capital*.

Operating Cycle

How much liquidity a firm needs depends on its *operating cycle*. The operating cycle is the duration from the time cash is invested in goods

and services to the time that investment produces cash. For example, a firm that produces and sells goods has an operating cycle comprising four phases:

1. Purchase raw materials and produce goods, investing in inventory.
2. Sell goods, generating sales, which may or may not be for cash.
3. Extend credit, creating accounts receivable.
4. Collect accounts receivable, generating cash.

The four phases make up the cycle of cash use and generation. The operating cycle would be somewhat different for companies that produce services rather than goods, but the idea is the same—the operating cycle is the length of time it takes to generate cash through the investment of cash.

What does the operating cycle have to do with liquidity? The longer the operating cycle, the more current assets are needed

(relative to current liabilities) since it takes longer to convert inventories and receivables into cash. In other words, the longer the operating cycle, the greater the amount of net working capital required.

To measure the length of an operating cycle we need to know:

- The time it takes to convert the investment in inventory into sales (that is, cash → inventory → sales → accounts receivable).
- The time it takes to collect sales on credit (that is, accounts receivable → cash).

We can estimate the operating cycle for Fictitious Corporation for the current year, using the balance sheet and income statement data. The number of days Fictitious ties up funds in inventory is determined by the total amount of money represented in inventory and the average day's cost of goods sold. The current investment in inventory—that is, the money “tied up” in inventory—is the ending balance of inventory on the balance sheet. The *average day's cost of goods sold* is the cost of goods sold on an average day in the year, which can be estimated by dividing the cost of goods sold (which is found on the income statement) by the number of days in the year. The average day's cost of goods sold for the current year is:

$$\begin{aligned} & \text{Average day's cost of goods sold} \\ &= \frac{\text{Cost of goods sold}}{365 \text{ days}} \\ &= \frac{\$6,500,000}{365 \text{ days}} \\ &= \$17,808 \text{ per day} \end{aligned}$$

In other words, Fictitious incurs, on average, a cost of producing goods sold of \$17,808 per day.

Fictitious has \$1.8 million of inventory on hand at the end of the year. How many days' worth of goods sold is this? One way to look at this is to imagine that Fictitious stopped buying more raw materials and just finished producing whatever was on hand in inventory, using available raw materials and work-in-process.

How long would it take Fictitious to run out of inventory?

We compute the days sales in inventory (DSI), also known as the *number of days of inventory*, by calculating the ratio of the amount of inventory on hand (in dollars) to the average day's cost of goods sold (in dollars per day):

$$\begin{aligned} & \text{Days sales in inventory} \\ &= \frac{\text{Amount of inventory on hand}}{\text{Average day's cost of goods sold}} \\ &= \frac{\$1,800,000}{\$17,808 \text{ per day}} = 101 \text{ days} \end{aligned}$$

In other words, Fictitious has approximately 101 days of goods on hand at the end of the current year. If sales continued at the same price, it would take Fictitious 101 days to run out of inventory.

If the ending inventory is representative of the inventory throughout the year, then it takes about 101 days to convert the investment in inventory into sold goods. Why worry about whether the year-end inventory is representative of inventory at any day throughout the year? Well, if inventory at the end of the fiscal year-end is lower than on any other day of the year, we have understated the DSI. Indeed, in practice most companies try to choose fiscal year-ends that coincide with the slow period of their business. That means the ending balance of inventory would be lower than the typical daily inventory of the year. To get a better picture of the firm, we could, for example, look at quarterly financial statements and take averages of quarterly inventory balances. However, here for simplicity we make a note of the problem of representatives and deal with it later in the discussion of financial ratios.

It should be noted that as an attempt to make the inventory figure more representative, some suggest taking the average of the beginning and ending inventory amounts. This does nothing to remedy the representativeness problem because the beginning inventory is simply the ending inventory from the previous year and, like the ending value from the current year, is

measured at the low point of the operating cycle. A preferred method, if data are available, is to calculate the average inventory for the four quarters of the fiscal year.

We can extend the same logic for calculating the number of days between a sale—when an account receivable is created—and the time it is collected in cash. If we assume that Fictitious sells all goods on credit, we can first calculate the *average credit sales per day* and then figure out how many days' worth of credit sales are represented by the ending balance of receivables.

The average credit sales per day are:

$$\begin{aligned}\text{Credit sales per day} &= \frac{\text{Credit sales}}{365 \text{ days}} \\ &= \frac{\$10,000,000}{365 \text{ days}} \\ &= \$27,397 \text{ per day}\end{aligned}$$

Therefore, Fictitious generates \$27,397 of credit sales per day. With an ending balance of accounts receivable of \$600,000, the *days sales outstanding* (DSO), also known as the *number of days of credit*, in this ending balance is calculated by taking the ratio of the balance in the accounts receivable account to the credit sales per day:

$$\begin{aligned}\text{Days sales outstanding} &= \frac{\text{Accounts receivable}}{\text{Credit sales per day}} \\ &= \frac{\$600,000}{\$27,397 \text{ per day}} \\ &= 22 \text{ days}\end{aligned}$$

If the ending balance of receivables at the end of the year is representative of the receivables on any day throughout the year, then it takes, on average, approximately 22 days to collect the accounts receivable. In other words, it takes 22 days for a sale to become cash.

Using what we have determined for the inventory cycle and cash cycle, we see that for Fictitious:

$$\begin{aligned}\text{Operating cycle} &= \text{DSI} + \text{DSO} \\ &= 101 \text{ days} + 22 \text{ days} \\ &= 123 \text{ days}\end{aligned}$$

We also need to look at the liabilities on the balance sheet to see how long it takes a firm to pay its short-term obligations. We can apply the same logic to accounts payable as we did to accounts receivable and inventories. How long does it take a firm, on average, to go from creating a payable (buying on credit) to paying for it in cash?

First, we need to determine the amount of an *average day's purchases on credit*. If we assume all the Fictitious purchases are made on credit, then the total purchases for the year would be the cost of goods sold less any amounts included in cost of goods sold that are not purchases. For example, depreciation is included in the cost of goods sold yet is not a purchase. Since we do not have a breakdown on the company's cost of goods sold showing how much was paid for in cash and how much was on credit, let us assume for simplicity that purchases are equal to cost of goods sold less depreciation. The average day's purchases then become:

$$\begin{aligned}\text{Average day's purchases} &= \frac{\text{Cost of goods sold} - \text{Depreciation}}{365 \text{ days}} \\ &= \frac{\$6,500,000 - \$1,000,000}{365 \text{ days}} \\ &= \$15,068 \text{ per day}\end{aligned}$$

The *days payables outstanding* (DPO), also known as the number of days of purchases, represented in the ending balance in accounts payable, is calculated as the ratio of the balance in the accounts payable account to the average day's purchases:

$$\begin{aligned}\text{Days payables outstanding} &= \frac{\text{Accounts payable}}{\text{Average day's purchases}}\end{aligned}$$

For Fictitious in the current year:

$$\begin{aligned}\text{Days payables outstanding} &= \frac{\$500,000}{\$15,065 \text{ per day}} \\ &= 33 \text{ days}\end{aligned}$$

This means that on average Fictitious takes 33 days to pay out cash for a purchase.

The operating cycle tells us how long it takes to convert an investment in cash back into cash (by way of inventory and accounts receivable). The number of days of payables tells us how long it takes to pay on purchases made to create the inventory. If we put these two pieces of information together, we can see how long, on net, we tie up cash. The difference between the operating cycle and the number of days of purchases is the *cash conversion cycle* (CCC), also known as the *net operating cycle*:

$$\text{Cash conversion cycle} = \text{Operating cycle} \\ - \text{DPO}$$

Or, substituting for the operating cycle,

$$\text{CCC} = \text{DSI} + \text{DSO} - \text{DPO}$$

The cash conversion cycle for Fictitious in the current year is:

$$\text{CCC} = 101 + 22 - 33 = 90 \text{ days}$$

The CCC is how long it takes for the firm to get cash back from its investments in inventory and accounts receivable, considering that purchases may be made on credit. By not paying for purchases immediately (that is, using trade credit), the firm reduces its liquidity needs. Therefore, the longer the net operating cycle, the greater the required liquidity.

Measures of Liquidity

The analyst can describe a firm's ability to meet its current obligations in several ways. The *current ratio* indicates the firm's ability to meet or cover its current liabilities using its current assets:

$$\text{Current ratio} = \frac{\text{Current assets}}{\text{Current liabilities}}$$

For the Fictitious Corporation, the current ratio for the current year is the ratio of current assets, \$3 million, to current liabilities, the sum of accounts payable and other current liabilities,

or \$1 million.

$$\text{Current ratio} = \frac{\$3,000,000}{\$1,000,000} = 3.0 \text{ times}$$

The current ratio of 3.0 indicates that Fictitious has three times as much as it needs to cover its current obligations during the year. However, the current ratio groups all current asset accounts together, assuming they are all as easily converted to cash. Even though, by definition, current assets can be transformed into cash within a year, not all current assets can be transformed into cash in a short period of time.

An alternative to the current ratio is the *quick ratio*, also called the *acid-test ratio*, which uses a slightly different set of current accounts to cover the same current liabilities as in the current ratio. In the quick ratio, the least liquid of the current asset accounts, inventory, is excluded. Hence:

$$\text{Quick ratio} = \frac{\text{Current assets} - \text{Inventory}}{\text{Current liabilities}}$$

We typically leave out inventories in the quick ratio because inventories are generally perceived as the least liquid of the current assets. By leaving out the least liquid asset, the quick ratio provides a more conservative view of liquidity.

For Fictitious in the current year:

$$\begin{aligned} \text{Quick ratio} &= \frac{\$3,000,000 - \$1,800,000}{\$1,000,000} \\ &= \frac{\$1,200,000}{\$1,000,000} = 1.2 \text{ times} \end{aligned}$$

Still another way to measure the firm's ability to satisfy short-term obligations is the *net working capital-to-sales ratio*, which compares net working capital (current assets less current liabilities) with sales:

$$\begin{aligned} \text{Net working capital-to-sales ratio} \\ &= \frac{\text{Net working capital}}{\text{Sales}} \end{aligned}$$

This ratio tells us the "cushion" available to meet short-term obligations relative to sales. Consider two firms with identical working capital of \$100,000, but one has sales of \$500,000 and

the other sales of \$1 million. If they have identical operating cycles, this means that the firm with the greater sales has more funds flowing in and out of its current asset investments (inventories and receivables). The company with more funds flowing in and out needs a larger cushion to protect itself in case of a disruption in the cycle, such as a labor strike or unexpected delays in customer payments. The longer the operating cycle, the more of a cushion (net working capital) a firm needs for a given level of sales.

For Fictitious Corporation:

$$\begin{aligned} \text{Net working capital-to-sales-ratio} \\ = \frac{\$3,000,000 - 1,000,000}{\$10,000,000} = 0.2000 \text{ or } 20\% \end{aligned}$$

The ratio of 0.20 tells us that for every dollar of sales, Fictitious has 20 cents of net working capital to support it.

Recap: Liquidity Ratios

Operating cycle and liquidity ratio information for Fictitious using data for the current year, in summary, is:

Days sales in inventory	= 101 days
Days sales outstanding	= 22 days
Operating cycle	= 123 days
Days payables outstanding	= 33 days
Cash conversion cycle	= 90 days
Current ratio	= 3.0
Quick ratio	= 1.2
Net working capital-to-sales ratio	= 20%

Given the measures of time related to the current accounts—the operating cycle and the cash conversion cycle—and the three measures of liquidity—current ratio, quick ratio, and net working capital-to-sales ratio—we know the following about Fictitious Corporation’s ability to meet its short-term obligations:

- Inventory is less liquid than accounts receivable (comparing days of inventory with days of credit).

- Current assets are greater than needed to satisfy current liabilities in a year (from the current ratio).
- The quick ratio tells us that Fictitious can meet its short-term obligations even without resorting to selling inventory.
- The net working capital “cushion” is 20 cents for every dollar of sales (from the net working capital-to-sales ratio.)

What don’t ratios tell us about liquidity? They don’t provide us with answers to the following questions:

- How liquid are the accounts receivable? How much of the accounts receivable will be collectible? Whereas we know it takes, on average, 22 days to collect, we do not know how much will never be collected.
- What is the nature of the current liabilities? How much of current liabilities consists of items that recur (such as accounts payable and wages payable) each period and how much consists of occasional items (such as income taxes payable)?
- Are there any unrecorded liabilities (such as operating leases) that are not included in current liabilities?

PROFITABILITY RATIOS

Liquidity ratios indicate a firm’s ability to meet its immediate obligations. Now we extend the analysis by adding *profitability ratios*, which help the analyst gauge how well a firm is managing its expenses. *Profit margin ratios* compare components of income with sales. They give the analyst an idea of which factors make up a firm’s income and are usually expressed as a portion of each dollar of sales. For example, the profit margin ratios we discuss here differ only in the numerator. It is in the numerator that we can evaluate performance for different aspects of the business.

For example, suppose the analyst wants to evaluate how well production facilities are managed. The analyst would focus on gross

profit (sales less cost of goods sold), a measure of income that is the direct result of production management. Comparing gross profit with sales produces the *gross profit margin*:

$$\begin{aligned} \text{Gross profit margin} \\ = \frac{\text{Revenues} - \text{Cost of goods sold}}{\text{Revenues}} \end{aligned}$$

This ratio tells us the portion of each dollar of sales that remains after deducting production expenses. For Fictitious Corporation for the current year:

$$\begin{aligned} \text{Gross profit margin} &= \frac{\$10,000,000 - \$6,500,000}{\$10,000,000} \\ &= \frac{\$3,500,000}{\$10,000,000} \\ &= 0.3500 \text{ or } 35\% \end{aligned}$$

For each dollar of revenues, the firm's gross profit is 35 cents. Looking at sales and cost of goods sold, we can see that the gross profit margin is affected by:

- Changes in sales volume, which affect cost of goods sold and sales.
- Changes in sales price, which affect revenues.
- Changes in the cost of production, which affect cost of goods sold.

Any change in gross profit margin from one period to the next is caused by one or more of those three factors. Similarly, differences in gross margin ratios among firms are the result of differences in those factors.

To evaluate operating performance, we need to consider operating expenses in addition to the cost of goods sold. To do this, remove operating expenses (e.g., selling and general administrative expenses) from gross profit, leaving operating profit, also referred to as earnings before interest and taxes (EBIT). The *operating profit margin* is therefore:

$$\begin{aligned} \text{Operating profit margin} \\ = \frac{\text{Revenues} - \text{Cost of goods sold} - \text{Operating expenses}}{\text{Revenues}} \\ = \frac{\text{Revenues earnings before interest and taxes}}{\text{Revenues}} \end{aligned}$$

For Fictitious in the current year:

$$\begin{aligned} \text{Operating profit margin} &= \frac{\$2,000,000}{\$10,000,000} \\ &= 0.20 \text{ or } 20\% \end{aligned}$$

Therefore, for each dollar of revenues, Fictitious has 20 cents of operating income. The operating profit margin is affected by the same factors as gross profit margin, plus operating expenses such as:

- Office rent and lease expenses
- Miscellaneous income (e.g., income from investments)
- Advertising expenditures
- Bad debt expense

Most of these expenses are related in some way to revenues, though they are not included directly in the cost of goods sold. Therefore, the difference between the gross profit margin and the operating profit margin is due to these indirect items that are included in computing the operating profit margin.

Both the gross profit margin and the operating profit margin reflect a company's operating performance. But they do not consider how these operations have been financed. To evaluate both operating and financing decisions, the analyst must compare net income (that is, earnings after deducting interest and taxes) with revenues. The result is the *net profit margin*:

$$\text{Net profit margin} = \frac{\text{Net income}}{\text{Revenues}}$$

The net profit margin tells the analyst the net income generated from each dollar of revenues; it considers financing costs that the operating profit margin does not consider. For Fictitious for the current year:

$$\text{Net profit margin} = \frac{\$1,200,000}{\$10,000,000} = 0.12 \text{ or } 12\%$$

For every dollar of revenues, Fictitious generates 12 cents in profits.

Recap: Profitability Ratios

The profitability ratios for Fictitious in the current year are:

Gross profit margin	= 35%
Operating profit margin	= 20%
Net profit margin	= 12%

They indicate the following about the operating performance of Fictitious:

- Each dollar of revenues contributes 35 cents to gross profit and 20 cents to operating profit.
- Every dollar of revenues contributes 12 cents to owners' earnings.
- By comparing the 20-cent operating profit margin with the 12-cent net profit margin, we see that Fictitious has 8 cents of financing costs for every dollar of revenues.

What these ratios do not indicate about profitability is the sensitivity of gross, operating, and net profit margins to:

- Changes in the sales price
- Changes in the volume of sales

Looking at the profitability ratios for one firm for one period gives the analyst very little information that can be used to make judgments regarding future profitability. Nor do these ratios provide the analyst any information about why current profitability is what it is. We need more information to make these kinds of judgments, particularly regarding the future profitability of the firm. For that, turn to activity ratios, which are measures of how well assets are being used.

ACTIVITY RATIOS

Activity ratios—for the most part, turnover ratios—can be used to evaluate the benefits produced by specific assets, such as inventory or accounts receivable, or to evaluate the benefits produced by the totality of the firm's assets.

Inventory Management

The *inventory turnover ratio* indicates how quickly a firm has used inventory to generate

the goods and services that are sold. The inventory turnover is the ratio of the cost of goods sold to inventory:

$$\text{Inventory turnover ratio} = \frac{\text{Cost of goods sold}}{\text{Inventory}}$$

For Fictitious for the current year:

$$\begin{aligned} \text{Inventory turnover ratio} &= \frac{\$6,500,000}{\$1,800,000} \\ &= 3.61 \text{ times} \end{aligned}$$

This ratio indicates that Fictitious turns over its inventory 3.61 times per year. On average, cash is invested in inventory, goods and services are produced, and these goods and services are sold 3.6 times a year. Looking back to the number of days of inventory, we see that this turnover measure is consistent with the results of that calculation: There are 101 calendar days of inventory on hand at the end of the year; dividing 365 days by 101 days, or 365/101 days, we find that inventory cycles through (from cash to sales) 3.61 times a year.

Accounts Receivable Management

In much the same way inventory turnover can be evaluated, an analyst can evaluate a firm's management of its accounts receivable and its credit policy. The *accounts receivable turnover ratio* is a measure of how effectively a firm is using credit extended to customers. The reason for extending credit is to increase sales. The downside to extending credit is the possibility of default—customers not paying when promised. The benefit obtained from extending credit is referred to as *net credit sales*—sales on credit less returns and refunds.

$$\begin{aligned} \text{Accounts receivable turnover} \\ &= \frac{\text{Net credit sales}}{\text{Accounts receivable}} \end{aligned}$$

Looking at the Fictitious Corporation income statement, we see an entry for sales, but we do not know how much of the amount stated is on credit. In the case of evaluating a firm, an analyst would have an estimate of the amount

of credit sales. Let us assume that the entire sales amount represents net credit sales. For Fictitious for the current year:

$$\begin{aligned}\text{Accounts receivable turnover} &= \frac{\$10,000,000}{\$600,000} \\ &= 16.67 \text{ times}\end{aligned}$$

Therefore, almost 17 times in the year there is, on average, a cycle that begins with a sale on credit and finishes with the receipt of cash for that sale. In other words, there are 17 cycles of sales to credit to cash during the year.

The number of times accounts receivable cycle through the year is consistent with the days sales outstanding (22) that we calculated earlier—accounts receivable turn over 17 times during the year, and the average number of days of sales in the accounts receivable balance is $365 \text{ days} / 16.67 \text{ times} = 22 \text{ days}$.

Overall Asset Management

The inventory and accounts receivable turnover ratios reflect the benefits obtained from the use of specific assets (inventory and accounts receivable). For a more general picture of the productivity of the firm, an analyst can compare the sales during a period with the total assets that generated these revenues.

One way is with the *total asset turnover ratio*, which indicates how many times during the year the value of a firm's total assets is generated in revenues:

$$\text{Total assets turnover} = \frac{\text{Revenues}}{\text{Total assets}}$$

For Fictitious in the current year:

$$\begin{aligned}\text{Total assets turnover} &= \frac{\$10,000,000}{\$11,000,000} \\ &= 0.91 \text{ times}\end{aligned}$$

The turnover ratio of 0.91 indicated that in the current year, every dollar invested in total assets generates 91 cents of sales. Or, stated differently, the total assets of Fictitious turn over almost once during the year. Because total as-

sets include both tangible and intangible assets, this turnover indicates how efficiently all assets were used.

An alternative is to focus only on fixed assets, the long-term, tangible assets of the firm. The *fixed-asset turnover* is the ratio of revenues to fixed assets:

$$\text{Fixed asset turnover ratio} = \frac{\text{Revenues}}{\text{Fixed assets}}$$

For Fictitious in the current year:

$$\begin{aligned}\text{Fixed asset turnover ratio} &= \frac{\$10,000,000}{\$7,000,000} \\ &= 1.43 \text{ times}\end{aligned}$$

Therefore, for every dollar of fixed assets, Fictitious is able to generate \$1.43 of revenues.

Recap: Activity Ratios

The activity ratios for Fictitious Corporation are:

Inventory turnover ratio	= 3.61 times
Accounts receivable turnover ratio	= 16.67 times
Total asset turnover ratio	= 0.91 times
Fixed-asset turnover ratio	= 1.43 times

From these ratios the analyst can determine that:

- Inventory flows in and out almost four times a year (from the inventory turnover ratio).
- Accounts receivable are collected in cash, on average, 22 days after a sale (from the number of days of credit). In other words, accounts receivable flow in and out almost 17 times during the year (from the accounts receivable turnover ratio).

Here is what these ratios do not indicate about the firm's use of its assets:

- The sales not made because credit policies are too stringent.
- How much of credit sales is not collectible.
- Which assets contribute most to the turnover.

FINANCIAL LEVERAGE RATIOS

A firm can finance its assets with equity or with debt. Financing with debt legally obligates the firm to pay interest and to repay the principal as promised. Equity financing does not obligate the firm to pay anything because dividends are paid at the discretion of the board of directors. There is always some risk, which we refer to as business risk, inherent in any business enterprise. But how a firm chooses to finance its operations—the particular mix of debt and equity—may add financial risk on top of business risk. *Financial risk* is risk associated with a firm's ability to satisfy its debt obligations, and is often measured using the extent to which debt financing is used relative to equity.

Financial leverage ratios are used to assess how much financial risk the firm has taken on. There are two types of financial leverage ratios: component percentages and coverage ratios. Component percentages compare a firm's debt with either its total capital (debt plus equity) or its equity capital. Coverage ratios reflect a firm's ability to satisfy fixed financing obligations, such as interest, principal repayment, or lease payments.

Component Percentage Ratios

A ratio that indicates the proportion of assets financed with debt is the *debt-to-assets ratio*, which compares total liabilities (short-term + long-term debt) with total assets:

$$\text{Total debt-to-assets ratio} = \frac{\text{Debt}}{\text{Total assets}}$$

For Fictitious in the current year:

$$\begin{aligned} \text{Total debt-to-assets ratio} &= \frac{\$5,000,000}{\$11,000,000} \\ &= 0.4546 \text{ or } 45.46\% \end{aligned}$$

This ratio indicates that 45% of the firm's assets are financed with debt (both short term and long term).

Another way to look at the financial risk is in terms of the use of debt relative to the use of eq-

uity. The debt-to-equity ratio indicates how the firm finances its operations with debt relative to the book value of its shareholders' equity:

$$\begin{aligned} \text{Debt-to-equity ratio} \\ &= \frac{\text{Debt}}{\text{Book value of shareholders' equity}} \end{aligned}$$

For Fictitious for the current year, using the book-value definition:

$$\begin{aligned} \text{Debt-to-equity ratio} &= \frac{\$5,000,000}{\$6,000,000} \\ &= 0.8333 \text{ or } 83.33\% \end{aligned}$$

For every \$1 of book value of shareholders' equity, Fictitious uses 83 cents of debt.

Both of these ratios can be stated in terms of total debt, as above, or in terms of long-term debt or even simply interest-bearing debt. And it is not always clear in which form—total, long-term debt, or interest-bearing—the ratio is calculated. Additionally, it is often the case that the current portion of long-term debt is excluded in the calculation of the long-term versions of these debt ratios.

Book Value versus Market Value

One problem with using a financial ratio based on the book value of equity to analyze financial risk is that there is seldom a strong relationship between the book value and market value of a stock. The distortion in values on the balance sheet is obvious by looking at the book value of equity and comparing it with the market value of equity. The book value of equity consists of:

- The proceeds to the firm of all the stock issues since it was first incorporated, less any stock repurchased by the firm.
- The accumulative earnings of the firm, less any dividends, since it was first incorporated.

Let's look at an example of the book value versus the market value of equity. IBM was incorporated in 1911, so the book value of its equity represents the sum of all its stock issued and all its earnings, less any dividends paid since 1911. As of the end of 2006, IBM's book value of

equity was approximately \$28.5 billion, yet its market value was \$142.8 billion.

Book value generally does not give a true picture of the investment of shareholders in the firm because:

- Earnings are recorded according to accounting principles, which may not reflect the true economics of transactions.
- Due to inflation, the earnings and proceeds from stock issued in the past do not reflect today's values.

Market value, on the other hand, is the value of equity as perceived by investors. It is what investors are willing to pay. So why bother with book value? For two reasons: First, it is easier to obtain the book value than the market value of a firm's securities, and second, many financial services report ratios using book value rather than market value.

However, any of the ratios presented in this entry that use the book value of equity can be restated using the market value of equity. For example, instead of using the book value of equity in the debt-to-equity ratio, the market value of equity to measure the firm's financial leverage can be used.

Coverage Ratios

The ratios that compare debt to equity or debt to assets indicate the amount of financial leverage, which enables an analyst to assess the financial condition of a firm. Another way of looking at the financial condition and the amount of financial leverage used by the firm is to see how well it can handle the financial burdens associated with its debt or other fixed commitments.

One measure of a firm's ability to handle financial burdens is the *interest coverage ratio*, also referred to as the *times interest-covered ratio*. This ratio tells us how well the firm can cover or meet the interest payments associated with debt. The ratio compares the funds available to pay interest (that is, earnings before interest and taxes)

with the interest expense:

$$\text{Interest coverage ratio} = \frac{\text{EBIT}}{\text{Interest expense}}$$

The greater the interest coverage ratio, the better able the firm is to pay its interest expense.

For Fictitious for the current year:

$$\text{Interest coverage ratio} = \frac{\$2,000,000}{\$400,000} = 5 \text{ times}$$

An interest coverage ratio of 5 means that the firm's earnings before interest and taxes are five times greater than its interest payments.

The interest coverage ratio provides information about a firm's ability to cover the interest related to its debt financing. However, there are other costs that do not arise from debt but that nevertheless must be considered in the same way we consider the cost of debt in a firm's financial obligations. For example, lease payments are fixed costs incurred in financing operations. Like interest payments, they represent legal obligations.

What funds are available to pay debt and debt-like expenses? Start with EBIT and add back expenses that were deducted to arrive at EBIT. The ability of a firm to satisfy its fixed financial costs—its fixed charges—is referred to as the *fixed-charge coverage ratio*. One definition of the fixed-charge coverage considers only the lease payments:

$$\begin{aligned} \text{Fixed-charge coverage ratio} \\ = \frac{\text{EBIT} + \text{Lease expense}}{\text{Interest} + \text{Lease expense}} \end{aligned}$$

For Fictitious for the current year:

$$\begin{aligned} \text{Fixed-charge coverage ratio} \\ = \frac{\$2,000,000 + \$1,000,000}{\$400,000 + \$1,000,000} \\ = 2.14 \text{ times} \end{aligned}$$

This ratio tells us that Fictitious's earnings can cover its fixed charges (interest and lease payments) more than two times over.

What fixed charges to consider is not entirely clear-cut. For example, if the firm is required to set aside funds to eventually or periodically

retire debt—referred to as sinking funds—is the amount set aside a fixed charge? As another example, since preferred dividends represent a fixed financing charge, should they be included as a fixed charge? From the perspective of the common shareholder, the preferred dividends must be covered either to enable the payment of common dividends or to retain earnings for future growth. Because debt principal repayment and preferred stock dividends are paid on an after-tax basis—paid out of dollars remaining after taxes are paid—this fixed charge must be converted to before-tax dollars. The fixed charge coverage ratio can be expanded to accommodate the sinking funds and preferred stock dividends as fixed charges.

Up to now we considered earnings before interest and taxes as funds available to meet fixed financial charges. EBIT includes noncash items such as depreciation and amortization. If an analyst is trying to compare funds available to meet obligations, a better measure of available funds is cash flow from operations, as reported in the statement of cash flows. A ratio that considers cash flows from operations as funds available to cover interest payments is referred to as the *cash-flow interest coverage ratio*.

$$\text{Cash flow interest coverage ratio} = \frac{\text{Cash flow from operations} + \text{Interest} + \text{Taxes}}{\text{Interest}}$$

The amount of cash flow from operations that is in the statement of cash flows is net of interest and taxes. So we have to add back interest and taxes to cash flow from operations to arrive at the cash flow amount before interest and taxes in order to determine the cash flow available to cover interest payments.

For Fictitious for the current year:

$$\begin{aligned} \text{Cash flow interest coverage ratio} &= \frac{\$1,800,000 + \$400,000 + \$400,000}{\$400,000} \\ &= \frac{\$2,600,000}{\$400,000} = 6.5 \text{ times} \end{aligned}$$

This coverage ratio indicates that, in terms of cash flows, Fictitious has 6.5 times more cash than is needed to pay its interest. This is a better picture of interest coverage than the five times reflected by EBIT. Why the difference? Because cash flow considers not just the accounting income, but noncash items as well. In the case of Fictitious, depreciation is a non-cash charge that reduced EBIT but not cash flow from operations—it is added back to net income to arrive at cash flow from operations.

Recap: Financial Leverage Ratios

Summarizing, the financial leverage ratios for Fictitious Corporation for the current year are:

Debt-to-assets ratio	= 45.45%
Debt-to-equity ratio	= 83.33%
Interest coverage ratio	= 5.00 times
Fixed-charge coverage ratio	= 2.14 times
Cash-flow interest coverage ratio	= 6.50 times

These ratios indicate that Fictitious uses its financial leverage as follows:

- Assets are 45% financed with debt, measured using book values.
- Long-term debt is approximately two-thirds of equity. When equity is measured in market value terms, long-term debt is approximately one-sixth of equity.

These ratios do not indicate:

- What other fixed, legal commitments the firm has that are not included on the balance sheet (for example, operating leases).
- What the intentions of management are regarding taking on more debt as the existing debt matures.

COMMON-SIZE ANALYSIS

An analyst can evaluate a company's operating performance and financial condition through ratios that relate various items of information contained in the financial statements. Another

way to analyze a firm is to look at its financial data more comprehensively.

Common-size analysis is a method of analysis in which the components of a financial statement are compared with each other. The first step in common-size analysis is to break down a financial statement—either the balance sheet or the income statement—into its parts. The next step is to calculate the proportion that each item represents relative to some benchmark. This form of common-size analysis is sometimes referred to as *vertical common-size analysis*. Another form of common-size analysis is *horizontal common-size analysis*, which uses either an income statement or a balance sheet in a fiscal year and compares accounts to the corresponding items in another year. In common-size analysis of the balance sheet, the benchmark is total assets. For the income statement, the benchmark is sales.

Let us see how it works by doing some common-size financial analysis for the Fictitious Corporation. The company's balance sheet is restated in Table 4. This statement does not look precisely like the balance sheet we have seen before. Nevertheless, the data are the same but reorganized. Each item in the original balance sheet has been restated as a proportion

Table 4 Fictitious Corporation Common-Size Balance Sheets for Years Ending December 31

	Current Year	Prior Year
Asset Components		
Cash	3.6%	2.0%
Marketable securities	1.8%	0.0%
Accounts receivable	5.5%	8.0%
Inventory	<u>16.4%</u>	<u>10.0%</u>
Current assets	27.3%	20.0%
Net plant and equipment	63.5%	70.0%
Intangible assets	<u>9.2%</u>	<u>10.0%</u>
Total assets	100.0%	100.0%
Liability and shareholders' equity components		
Accounts payable	4.6%	4.0%
Other current liabilities	4.6%	2.0%
Long-term debt	<u>36.4%</u>	<u>50.0%</u>
Total liabilities	45.4%	56.0%
Shareholders' equity	<u>54.6%</u>	<u>44.0%</u>
Total liabilities and shareholders' equity	100.0%	100.0%

of total assets for the purpose of common size analysis. Hence, we refer to this as the *common-size balance sheet*.

In this balance sheet, we see, for example, that in the current year cash is 3.6% of total assets, or $\$400,000/\$11,000,000 = 0.036$. The largest investment is in plant and equipment, which comprises 63.6% of total assets. On the liabilities side, that current liabilities are a small portion (9.1%) of liabilities and equity.

The common-size balance sheet indicates in very general terms how Fictitious has raised capital and where this capital has been invested. As with financial ratios, however, the picture is not complete until trends are examined and compared with those of other firms in the same industry.

In the income statement, as with the balance sheet, the items may be restated as a proportion of sales; this statement is referred to as the *common-size income statement*. The common-size income statements for Fictitious for the two years are shown in Table 5. For the current year, the major costs are associated with goods sold (65%); lease expense, other expenses, interest, taxes, and dividends make up smaller portions of sales. Looking at gross profit, EBIT, and net income, these proportions are the profit margins we calculated earlier. The common-size income statement provides information on the profitability of different aspects of the firm's business. Again, the picture is not yet complete.

Table 5 Fictitious Corporation Common-Size Income Statement for Years Ending December 31

	Current Year	Prior Year
Sales	100.0%	100.0%
Cost of goods sold	<u>65.0%</u>	<u>66.7%</u>
Gross profit	35.0%	33.3%
Lease and administrative expenses	<u>15.0%</u>	<u>11.1%</u>
Earnings before interest and taxes	20.0%	22.2%
Interest expense	<u>4.0%</u>	<u>5.6%</u>
Earnings before taxes	16.0%	16.6%
Taxes	<u>4.0%</u>	<u>5.5%</u>
Net income	12.0%	11.1%
Common dividends	<u>6.0%</u>	<u>5.6%</u>
Retained earnings	6.0%	5.5%

For a more complete picture, the analyst must look at trends over time and make comparisons with other companies in the same industry.

USING FINANCIAL RATIO ANALYSIS

Financial analysis provides information concerning a firm's operating performance and financial condition. This information is useful for an analyst in evaluating the performance of the company as a whole, as well as of divisions, products, and subsidiaries. An analyst must also be aware that financial analysis is also used by analysts and investors to gauge the financial performance of the company.

But financial ratio analysis cannot tell the whole story and must be interpreted and used with care. Financial ratios are useful but, as noted in the discussion of each ratio, there is information that the ratios do not reveal. For example, in calculating inventory turnover, we need to assume that the inventory shown on the balance sheet is representative of inventory throughout the year. Another example is in the calculation of accounts receivable turnover. We assumed that all sales were on credit. If we are on the outside looking in—that is, evaluating a firm based on its financial statements only, such as the case of a financial analyst or investor—and therefore do not have data on credit sales, assumptions must be made that may or may not be correct.

In addition, there are other areas of concern that an analyst should be aware of in using financial ratios:

- Limitations in the accounting data used to construct the ratios.
- Selection of an appropriate benchmark firm or firms for comparison purposes.
- Interpretation of the ratios.
- Pitfalls in forecasting future operating performance and financial condition based on past trends.

KEY POINTS

- The basic data for financial analysis are the financial statement data. These data are used to analyze relationships between different elements of a firm's financial statements. Through this analysis, a picture of the operating performance and financial condition of a firm can be developed.
- Looking at the calculated financial ratios, in conjunction with industry and economic data, judgments about past and future financial performance and condition can be made.
- Financial ratios can be classified by type—coverage, return, turnover, or component percentage—or by the financial characteristic that we wish to measure—liquidity, profitability activity, financial leverage, or return.
- Liquidity ratios indicate firm's ability to satisfy short-term obligations. These ratios are closely related to a firm's operating cycle, which tells us how long it takes a firm to turn its investment in current assets back into cash.
- Profitability ratios indicate how well a firm manages its assets, typically in terms of the proportion of revenues that are left over after expenses.
- Activity ratios measure how efficiently a firm manages its assets, that is, how effectively a firm uses its assets to generate sales.
- Financial leverage ratios indicate (1) to what extent a firm uses debt to finance its operations and (2) its ability to satisfy debt and debt-like obligations.
- Return-on-investment ratios provide a gauge for how much of each dollar of an investment is generated in a period.
- The DuPont system breaks down return ratios into their profit margin and activity ratios, allowing us to analyze changes in return on investments.
- Common-size analysis expresses financial statement data relative to some benchmark item—usually total assets for the balance sheet and sales for the income statement. Representing financial data in this way allows

an analyst to spot trends in investments and profitability.

- Interpretation of financial ratios requires an analyst to put the trends and comparisons in perspective with the company's significant events. In addition to company-specific events, issues that can cause the analysis of financial ratios to become more challenging include the use of historical accounting values, changes in accounting principles, and accounts that are difficult to classify.
- Comparison of financial ratios across time and with competitors is useful in gauging performance. In comparing ratios over time, an analyst should consider changes in accounting and significant company events. In comparing ratios with a benchmark, an analyst

must take care in the selection of the companies that constitute the benchmark and the method of calculation.

REFERENCES

- Bernstein, L. A. (1999). *Analysis of Financial Statements*, 5th edition. New York: McGraw-Hill.
- Fabozzi, F. J., Drake, P. P., and Polimeni, R. S. (2007). *The Complete CFO Handbook: From Accounting to Accountability*. Hoboken, NJ: John Wiley & Sons.
- Fridson, M., and Alvarez, F. (2011). *Financial Statement Analysis: A Practitioner's Guide*, 4th edition. Hoboken, NJ: John Wiley & Sons.
- Peterson, P. P., and Fabozzi, F. J. (2012). *Analysis of Financial Statements*, 3rd edition. Hoboken, NJ: John Wiley & Sons.

Cash-Flow Analysis

PAMELA P. DRAKE, PhD, CFA

J. Gray Ferguson Professor of Finance, College of Business, James Madison University

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: An objective of financial analysis is to assess a company's operating performance and financial condition. The information that is available for analysis includes economic, market, and financial information. But some of the most important financial data are provided by the company in its annual and quarterly financial statements. These choices make it quite difficult to compare financial performance and condition across companies, and also provide an opportunity for the management of financial numbers through judicious choice of accounting methods. Cash flows provide a way of transforming net income based on an accrual system to a more comparable basis. Additionally, cash flows are essential ingredients in valuation: The value of a company today is the present value of its expected future cash flows. Therefore, understanding past and current cash flows may help in forecasting future cash flows and, hence, determine the value of the company. Moreover, understanding cash flow allows the assessment of the ability of a firm to maintain current dividends and its current capital expenditure policy without relying on external financing.

One of the key financial measures that an analyst should understand is the company's cash flow. This is because the cash flow aids the analyst in assessing the ability of the company to satisfy its contractual obligations and maintain current dividends and current capital expenditure policy without relying on external financing. Moreover, an analyst must understand why this measure is important for external parties, specifically stock analysts covering the company. The reason is that the basic valuation principle followed by stock analysts is that the value of a company today is the present

value of its expected future cash flows. In this entry, we discuss *cash-flow analysis*.

DIFFICULTIES WITH MEASURING CASH FLOW

The primary difficulty with measuring a cash flow is that it is a flow: Cash flows into the company (cash inflows) and cash flows out of the company (cash outflows). At any point in time there is a stock of cash on hand, but the stock of cash on hand varies among companies because

of the size of the company, the cash demands of the business, and a company's management of working capital. So what is cash flow? Is it the total amount of cash flowing into the company during a period? Is it the total amount of cash flowing out of the company during a period? Is it the net of the cash inflows and outflows for a period? Well, there is no specific definition of cash flow—and that's probably why there is so much confusion regarding the measurement of cash flow. Ideally, a measure of the company's operating performance that is comparable among companies is needed—something other than net income.

A simple, yet crude method of calculating cash flow requires simply adding noncash expenses (e.g., depreciation and amortization) to the reported net income amount to arrive at cash flow. For example, the estimated cash flow for Procter & Gamble (P&G) for 2002 is:

$$\begin{aligned} &\text{Estimated cash flow} \\ &= \text{Net income} \quad + \quad \text{Depreciation and} \\ &\quad \quad \quad \text{amortization} \\ &= \$4,352 \text{ million} \quad + \quad 1,693 \text{ million} \\ &= \$6,045 \text{ million} \end{aligned}$$

This amount is not really a cash flow, but simply earnings before depreciation and amortization. Is this a cash flow that stock analysts should use in valuing a company? Though not a cash flow, this estimated cash flow does allow a quick comparison of income across firms that may use different depreciation methods and depreciable lives. (As an example of the use of this estimate of cash flow, *The Value Line Investment Survey*, published by Value Line, Inc., reports a cash flow per share amount, calculated as reported earnings plus depreciation, minus any preferred dividends, stated per share of common stock.) [*Guide to Using the Value Line Investment Survey* (New York: Value Line, Inc.), p. 19.]

The problem with this measure is that it ignores the many other sources and uses of cash during the period. Consider the sale of goods for credit. This transaction generates sales for

the period. Sales and the accompanying cost of goods sold are reflected in the period's net income and the estimated cash flow amount. However, until the account receivable is collected, there is no cash from this transaction. If collection does not occur until the next period, there is a misalignment of the income and cash flow arising from this transaction. Therefore, the simple estimated cash flow ignores some cash flows that, for many companies, are significant.

Another estimate of cash flow that is simple to calculate is earnings before interest, taxes, depreciation, and amortization (EBITDA). However, this measure suffers from the same accrual-accounting bias as the previous measure, which may result in the omission of significant cash flows. Additionally, EBITDA does not consider interest and taxes, which may also be substantial cash outflows for some companies. (For a more detailed discussion of the EBITDA measure, see Eastman [1997].)

These two rough estimates of cash flows are used in practice not only for their simplicity, but because they experienced widespread use prior to the disclosure of more detailed information in the statement of cash flows. Currently, the measures of cash flow are wide ranging, including the simplistic cash flow measures, measures developed from the statement of cash flows, and measures that seek to capture the theoretical concept of *free cash flow*.

CASH FLOWS AND THE STATEMENT OF CASH FLOWS

Prior to the adoption of the statement of cash flows, the information regarding cash flows was quite limited. The first statement that addressed the issue of cash flows was the statement of financial position, which was required starting in 1971 (APB Opinion No. 19, "Reporting Changes in Financial Position"). This statement was quite limited, requiring an analysis of the

sources and uses of funds in a variety of formats. In its earlier years of adoption, most companies provided this information using what is referred to as the *working capital concept*—a presentation of working capital provided and applied during the period. Over time, many companies began presenting this information using the *cash concept*, which is a most detailed presentation of the cash flows provided by operations, investing, and financing activities.

Consistent with the cash concept format of the funds flow statement, the statement of cash flows is now a required financial statement. The requirement that companies provide a statement of cash flows applies to fiscal years after 1987 (Statement of Financial Accounting Standards No. 95, “Statement of Cash Flows”). This statement requires the company to classify cash flows into three categories, based on the activity: operating, investing, and financing. Cash flows are summarized by activity and within activity by type (e.g., asset dispositions are reported separately from asset acquisitions).

The reporting company may report the cash flows from operating activities on the statement of cash flows using either the *direct method*—reporting all cash inflows and outflows—or the *indirect method*—starting with net income and making adjustments for depreciation and other noncash expenses and for changes in working capital accounts. Though the direct method is recommended, it is also the most burdensome for the reporting company to prepare. Most companies report cash flows from operations using the indirect method. The indirect method has the advantage of providing the financial statement user with a reconciliation of the company’s net income with the change in cash. The indirect method produces a cash flow from operations that is similar to the estimated cash flow measure discussed previously, yet it encompasses the changes in working capital accounts that the simple measure does not. For example, Procter & Gamble’s cash flow from operating activities (taken from their 2002 statement of cash flows) is \$7,742 million,

which is over \$1 billion more than the cash flow that we estimated earlier. (Procter & Gamble’s fiscal year ends June 30, 2002.)

The classification of cash flows into the three types of activities provides useful information that can be used by an analyst to see, for example, whether the company is generating sufficient cash flows from operations to sustain its current rate of growth. However, the classification of particular items is not necessarily as useful as it could be. Consider some of the classifications:

- Cash flows related to interest expense are classified in operations, though they are clearly financing cash flows.
- Income taxes are classified as operating cash flows, though taxes are affected by financing (e.g., deduction for interest expense paid on debt) and investment activities (e.g., the reduction of taxes from tax credits on investment activities).
- Interest income and dividends received are classified as operating cash flows, though these flows are a result of investment activities.

Whether these items have a significant effect on the analysis depends on the particular company’s situation. Procter & Gamble, for example, has very little interest and dividend income, and its interest expense of \$603 million is not large relative to its earnings before interest and taxes (\$6,986 million). Table 1 shows that by adjusting P&G’s cash flows for the interest expense only (and related taxes) changes the complexion of its cash flows slightly to reflect greater cash-flow generation from operations and less cash flow reliance on financing activities.

The adjustment is for \$603 million of interest and other financing costs, less its tax shield (the amount that the tax bill is reduced by the interest deduction) of \$211 (estimated from the average tax rate of 35% of \$603): adjustment = $\$603(1 - 0.35) = \392 .

Table 1 Adjusted Cash Flow for P&G (2002)

(In Millions)	As Reported	As Adjusted
Cash flow from operations	\$7,741	\$8,134
Cash flow for investing activities	(6,835)	(6,835)
Cash flow from (for) financing activities	197	(195)

Source: Procter & Gamble 2002 Annual Report.

For other companies, however, this adjustment may provide a less flattering view of cash flows. Consider Amazon.com's fiscal year results. Interest expense to financing, along with their respective estimated tax effects, results in more reliance on cash flow from financing as can be seen in Table 2.

Looking at the relation among the three cash flows in the statement provides a sense of the activities of the company. A young, fast-growing company may have negative cash flows from operations, yet positive cash flows from financing activities (that is, operations may be financed in large part with external financing). As a company grows, it may rely to a lesser extent on external financing. The typical, mature company generates cash from operations and reinvests part or all of it back into the company. Therefore, cash flow related to operations is positive (that is a source of cash) and cash flow related to investing activities is negative (that is, a use of cash). As a company matures, it may seek less financing externally and may even use cash to reduce its reliance on external financing (e.g., repay debts). We can classify companies on the basis of the pattern of their sources of cash flows, as shown in Table 3.

Table 2 Adjusted Cash Flow, Amazon.com (2001)

(In Millions)	As Reported	As Adjusted
Cash flow from operations	\$(120)	\$(30)
Cash flow for investing activities	(253)	(253)
Cash flow from financing activities	(107)	17

The adjustment is based on interest expense of \$139 million, and a tax rate of 35%.

Source: Amazon.com 2001 10-K.

Though additional information is required to assess a company's financial performance and condition, examination of the sources of cash flows, especially over time, gives us a general idea of the company's operations. P&G's cash flow pattern is consistent with that of a mature company, whereas Amazon.com's cash flows are consistent with those of a fast-growing company that is reliant on outside funds for growth.

Fridson (2002) suggests reformatting the statement of cash flows as shown in Table 4. From the basic cash flow, the nondiscretionary cash needs are subtracted resulting in a cash flow referred to as discretionary cash flow. By restructuring the statement of cash flows in this way, it can be seen how much flexibility the company has when it must make business decisions that may adversely impact the long-run financial health of the enterprise.

For example, consider a company with a basic cash flow of \$800 million and operating cash flow of \$500 million. Suppose that this company pays dividends of \$130 million and that its capital expenditure is \$300 million. Then the discretionary cash flow for this company is \$200 million found by subtracting the \$300 million capital expenditure from the operating cash flow of \$500 million. This means that even after maintaining a dividend payment of \$130 million, its cash flow is positive. Notice that asset sales and other investing activity are not needed to generate cash to meet the dividend payments because in Table 4 these items are subtracted after accounting for the dividend payments. In fact, if this company planned to increase its capital expenditures, the format in Table 4 can be used to assess how much that expansion can be before affecting dividends and/or increasing financing needs.

Though we can classify a company based on the sources and uses of cash flows, more data are needed to put this information in perspective. What is the trend in the sources and uses of cash flows? What market, industry, or company-specific events affect the company's cash flows? How does the company being

Table 3 Patterns of Sources of Cash Flows

Cash Flow	Financing Growth Externally and Internally	Financing Growth Internally	Mature	Temporary Financial Downturn	Financial Distress	Downsizing
Operations	+	+	+	-	-	+
Investing activities	-	-	-	+	-	+
Financing activities	+	-	+ or -	+	-	-

analyzed compare with other companies in the same industry in terms of the sources and uses of funds?

Let's take a closer look at the incremental information provided by cash flows. Consider Wal-Mart Stores, Inc., which had growing sales and net income from 1990 to 2005, as summarized in Figure 1. We see that net income grew each year, with the exception of 1995, and that sales grew each year.

We get additional information by looking at the cash flows and their sources, as graphed in Figure 2. We see that the growth in Wal-Mart was supported both by internally generated funds and, to a lesser extent, through external financing. Wal-Mart's pattern of cash flows suggests that Wal-Mart is a mature company that

has become less reliant on external financing, funding most of its growth in recent years (with the exception of 1999) with internally generated funds.

FREE CASH FLOW

Cash flows without any adjustment may be misleading because they do not reflect the cash outflows that are necessary for the future existence of a firm. An alternative measure, free cash flow, was developed by Jensen (1986) in his theoretical analysis of agency costs and corporate takeovers. In theory, free cash flow is the cash flow left over after the company funds all positive net present value projects. Positive net present value projects are those capital investment projects for which the present value of expected future cash flows exceeds the present value of project outlays, all discounted at the cost of capital. (The cost of capital is the cost to the company of funds from creditors and shareholders. The cost of capital is basically a hurdle: If a project returns more than its cost of capital, it is a profitable project.) In other words, free cash flow is the cash flow of the firm, less capital expenditures necessary to stay in business (that is, replacing facilities as necessary) and grow at the expected rate (which requires increases in working capital).

The theory of free cash flow was developed by Jensen to explain behaviors of companies that could not be explained by existing economic theories. Jensen observed that companies that generate free cash flow should disgorge that cash rather than invest the funds in less profitable investments. There are many ways in which companies can disgorge this

Table 4 Suggested Reformatting of Cash Flow Statement to Analyze a Company's Flexibility

	Basic cash flow
Less:	Increase in adjusted working capital
	Operating cash flow
Less:	Capital expenditures
	Discretionary cash flow
Less:	Dividends
Less:	Asset sales and other investing activities
	Cash flow before financing
Less:	Net (increase) in long-term debt
Less:	Net (increase) in notes payable
Less:	Net purchase of company's common stock
Less:	Miscellaneous
	Cash flow

Notes:

1. The basic cash flow includes net earnings, depreciation, and deferred income taxes, less items in net income not providing cash.
2. The increase in adjusted working capital excludes cash and payables.

Source: This format was suggested by Fridson (1995).

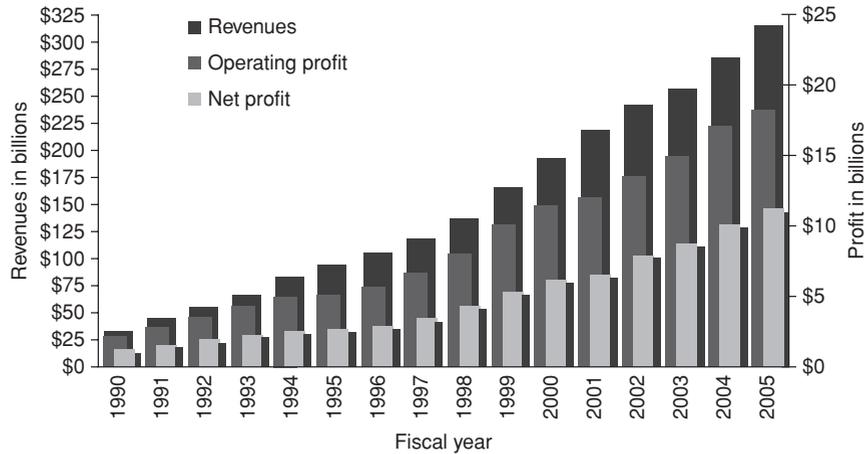


Figure 1 Wal-Mart Stores, Inc., Revenues, Operating Profit, and Net Income, 1990–2005
 Source: Wal-Mart Stores, Inc., Annual Report, various years.

excess cash flow, including the payment of cash dividends, the repurchase of stock, and debt issuance in exchange for stock. The debt-for-stock exchange, for example, increases the company’s leverage and future debt obligations, obligating the future use of excess cash flow. If a company does not disgorge this free cash flow, there is the possibility that another company—a company whose cash flows are less than its profitable investment opportunities or a company that is willing to purchase and lever-up the company—will attempt to acquire the free-cash-flow-laden company.

As a case in point, Jensen observed that the oil industry illustrates the case of wasting re-

sources: The free cash flows generated in the 1980s were spent on low-return exploration and development and on poor diversification attempts through acquisitions. He argues that these companies would have been better off paying these excess cash flows to shareholders through share repurchases or exchanges with debt.

By itself, the fact that a company generates free cash flow is neither good nor bad. What the company does with this free cash flow is what is important. And this is where it is important to measure the free cash flow as that cash flow in excess of profitable investment opportunities. Consider the simple numerical exercise

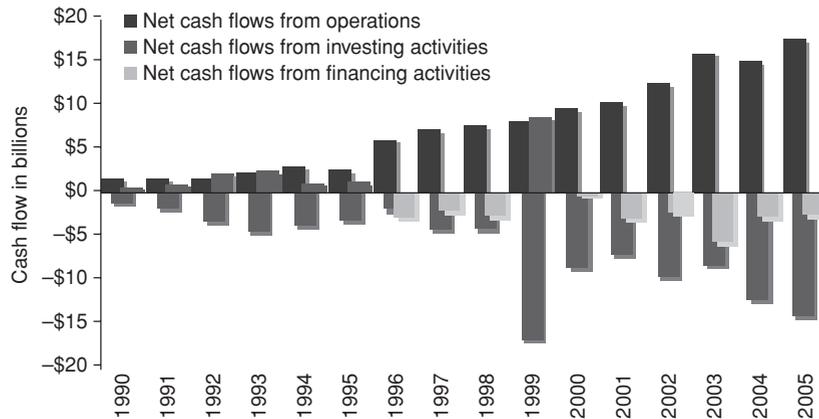


Figure 2 Wal-Mart Stores, Inc., Cash Flows, 1990–2005
 Source: Wal-Mart Stores, Inc., Annual Report, various years.

with the Winner Company and the Loser Company:

	Winner Company	Loser Company
Cash flow before capital expenditures	\$1,000	\$1,000
Capital expenditures, positive net present value projects	(750)	(250)
Capital expenditures, negative net present value projects	0	(500)
Cash flow	\$250	\$250
Free cash flow	\$250	\$750

These two companies have identical cash flows and the same total capital expenditures. However, the Winner Company spends only on profitable projects (in terms of positive net present value projects), whereas the Loser Company spends on both profitable projects and wasteful projects. The Winner Company has a lower free cash flow than the Loser Company, indicating that they are using the generated cash flows in a more profitable manner. The lesson is that the existence of a high level of free cash flow is not necessarily good—it may simply suggest that the company is either a very good takeover target or the company has the potential for investing in unprofitable investments.

Positive free cash flow may be good or bad news; likewise, negative free cash flow may be good or bad news:

	Good News	Bad News
Positive free cash flow	The company is generating substantial operating cash flows, beyond those necessary for profitable projects.	The company is generating more cash flows than it needs for profitable projects and may waste these cash flows on unprofitable projects.
Negative free cash flow	The company has more profitable projects than it has operating cash flows and must rely on external financing to fund these projects.	The company is unable to generate sufficient operating cash flows to satisfy its investment needs for future growth.

Therefore, once the free cash flow is calculated, other information (e.g., trends in profitability) must be considered to evaluate the operating performance and financial condition of the firm.

CALCULATING FREE CASH FLOW

There is some confusion when this theoretical concept is applied to actual companies. The primary difficulty is that the amount of capital expenditures necessary to maintain the business at its current rate of growth is generally not known; companies do not report this item and may not even be able to determine how much of a period's capital expenditures are attributed to maintenance and how much are attributed to expansion.

Consider Procter & Gamble's property, plant, and equipment for 2002, which comprise some, but not all, of P&G's capital investment:

Additions to property, plant, and equipment	\$1,679 million
Dispositions of property, plant, and equipment	(227)
Net change before depreciation	\$1,452 million

(In addition to the traditional capital expenditures (that is, changes in property, plant, and equipment), P&G also has cash flows related to investment securities and acquisitions. These investments are long-term and are hence part of P&G's investment activities cash outflow of \$6,835 million.)

How much of the \$1,679 million is for maintaining P&G's current rate of growth and how much is for expansion? Though there is a positive net change of \$1,452 million, does it mean that P&G is expanding? Not necessarily: The additions are at current costs, whereas the dispositions are at historical costs. The additions of \$1,679 are less than P&G's depreciation and amortization expense for 2001 of \$1,693 million, yet it is not disclosed in the financial reports how much of this latter amount reflects amortization. (P&G's depreciation and amortization

are reported together as \$1,693 million on the statement of cash flows.) The amount of necessary capital expenditures is therefore elusive.

Some estimate free cash flow by assuming that all capital expenditures are necessary for the maintenance of the current growth of the company. Though there is little justification in using all expenditures, this is a practical solution to an impractical calculation. This assumption allows us to estimate free cash flows using published financial statements.

Another issue in the calculation is defining what is truly “free” cash flow. Generally, we think of “free” cash flow as that being left over after all necessary financing expenditures are paid; this means that free cash flow is after interest on debt is paid. Some calculate free cash flow before such financing expenditures, others calculate free cash flow after interest, and still others calculate free cash flow after both interest and dividends (assuming that dividends are a commitment, though not a legal commitment).

There is no one correct method of calculating free cash flow and different analysts may arrive at different estimates of free cash flow for a company. The problem is that it is impossible to measure free cash flow as dictated by the theory, so many methods have arisen to calculate this cash flow. A simple method is to start with the cash flow from operations and then deduct capital expenditures. For P&G in 2002,

Cash flow from operations	\$7,742
Deduct capital expenditures	<u>(1,692)</u>
Free cash flow	\$6,050

Though this approach is rather simple, the cash flow from the operations amount includes a deduction for interest and other financing expenses. Making an adjustment for the after-tax interest and financing expenses, as we did earlier for Procter & Gamble,

Cash flow from operations (as reported)	\$7,742
Adjustment	<u>392</u>
Cash flow from operations (as adjusted)	\$8,134
Deduct capital expenditures	<u>(1,692)</u>
Free cash flow	\$6,442

We can relate free cash flow directly to a company’s income. Starting with net income, we can estimate free cash flow using four steps:

- Step 1: Determine earnings before interest and taxes (EBIT).
- Step 2: Calculate earnings before interest but after taxes.
- Step 3: Adjust for noncash expenses (e.g., depreciation).
- Step 4: Adjust for capital expenditures and changes in working capital.

Using these four steps, we can calculate the free cash flow for Procter & Gamble for 2002, as shown in Table 5.

NET FREE CASH FLOW

There are many variations in the calculation of cash flows that are used in analyses of companies’ financial condition and operating performance. As an example of these variations, consider the alternative to free cash flow developed by Fitch, a company that rates corporate debt instruments. This cash flow measure, referred to as *net free cash flow* (NFCF), is free cash flow less interest and other financing costs and taxes. In this approach, free cash flow is defined as earnings before depreciation, interest, and taxes, less capital expenditures. Capital expenditures encompass all capital spending, whether for maintenance or expansion, and no changes in working capital are considered.

The basic difference between NFCF and free cash flow is that the financing expenses—interest and, in some cases, dividends—are deducted. If preferred dividends are perceived as nondiscretionary—that is, investors come to expect the dividends—dividends may be included with the interest commitment to arrive at net free cash flow. Otherwise, dividends are deducted from net free cash flow to produce cash flow. Another difference is that NFCF does not consider changes in working capital in the analysis.

Table 5 Calculation of Procter & Gamble's Free Cash Flow for 2002, in Millions*

<i>Step 1:</i>		
Net income	\$4,352	
Add taxes	2,031	
Add interest	<u>603</u>	
Earnings before interest and taxes	\$6,986	
<i>Step 2:</i>		
Earnings before interest and taxes	\$6,986	
Deduct taxes (@35%)	<u>(2,445)</u>	
Earnings before interest	\$4,541	
<i>Step 3:</i>		
Earnings before interest	\$4,541	
Add depreciation and amortization	1,693	
Add increase in deferred taxes	<u>389</u>	
Earnings before noncash expenses	\$6,623	
<i>Step 4:</i>		
Earnings before noncash expenses		\$6,623
Deduct capital expenditures		<u>(1,679)</u>
Add decrease in receivables	\$96	
Add decrease in inventories	159	
Add cash flows from changes in accounts payable, accrued expenses, and other liabilities	684	
Deduct cash flow from changes in other operating assets and liabilities	<u>(98)</u>	
Cash flow from change in working capital accounts		<u>841</u>
Free cash flow		\$5,785

*Procter & Gamble's fiscal year ended June 30, 2002. Charges in operating accounts are taken from Procter & Gamble's Statement of Cash Flows.

Further, cash taxes are deducted to arrive at net free cash flow. Cash taxes are the income tax expense restated to reflect the actual cash flow related to this obligation, rather than the accrued expense for the period. Cash taxes are the income tax expense (from the income statement) adjusted for the change in deferred income taxes (from the balance sheets). For Procter & Gamble in 2002,

Income tax expense	\$2,031
Deduct increase in deferred income tax	<u>(389)</u>
Cash taxes	\$1,642

(Note that cash taxes require taking the tax expense and either increasing this to reflect any decrease in deferred taxes [that is, the payment this period of tax expense recorded in a prior period] or decreasing this amount to reflect any increase in deferred taxes [that is, the deferment of some of the tax expense].)

In the case of Procter & Gamble for 2002,

EBIT	\$6,986
Add depreciation and amortization	<u>1,693</u>
EBITDA	\$8,679
Deduct capital expenditures	<u>(1,679)</u>
Free cash flow	\$7,000
Deduct interest	(603)
Deduct cash taxes	<u>(1,642)</u>
Net free cash flow	\$4,755
Deduct cash common dividends	<u>(2,095)</u>
Net cash flow	\$2,660

The free cash flow amount per this calculation differs from the \$5,785 that we calculated earlier for two reasons: Changes in working capital and the deduction of taxes on operating earnings were not considered.

Net cash flow gives an idea of the unconstrained cash flow of the company. This cash flow measure may be useful from a creditor's perspective in terms of evaluating the company's ability to fund additional debt. From a

shareholder's perspective, net cash flow (that is, net free cash flow net of dividends) may be an appropriate measure because this represents the cash flow that is reinvested in the company.

USEFULNESS OF CASH FLOWS IN FINANCIAL ANALYSIS

The usefulness of cash flows for financial analysis depends on whether cash flows provide unique information or provide information in a manner that is more accessible or convenient for the analyst. The cash flow information provided in the statement of cash flows, for example, is not necessarily unique because most, if not all, of the information is available through analysis of the balance sheet and income statement. What the statement does provide is a classification scheme that presents information in a manner that is easier to use and, perhaps, more illustrative of the company's financial position.

An analysis of cash flows and the sources of cash flows can reveal the following information:

- **The sources of financing the company's capital spending.** Does the company generate in-

ternally (that is, from operations) a portion or all of the funds needed for its investment activities? If a company cannot generate cash flow from operations, this may indicate problems up ahead. Reliance on external financing (e.g., equity or debt issuance) may indicate a company's inability to sustain itself over time.

- **The company's dependence on borrowing.** Does the company rely heavily on borrowing that may result in difficulty in satisfying future debt service?
- **The quality of earnings.** Large and growing differences between income and cash flows suggest a low quality of earnings.

Consider the financial results of Krispy Kreme Doughnuts, Inc., a wholesaler and retailer of donuts. Krispy Kreme grew from having fewer than 200 stores before its initial public offering (IPO) in 2000 to over 400 stores at the end of its 2005 fiscal year. Accompanying this growth in stores is the growth in operating and net income, as we show in Figure 3. The growth in income continued after the IPO as the number of stores increased, but the tide in income turned in the 2004 fiscal year and losses continued into the 2005 fiscal year as well.

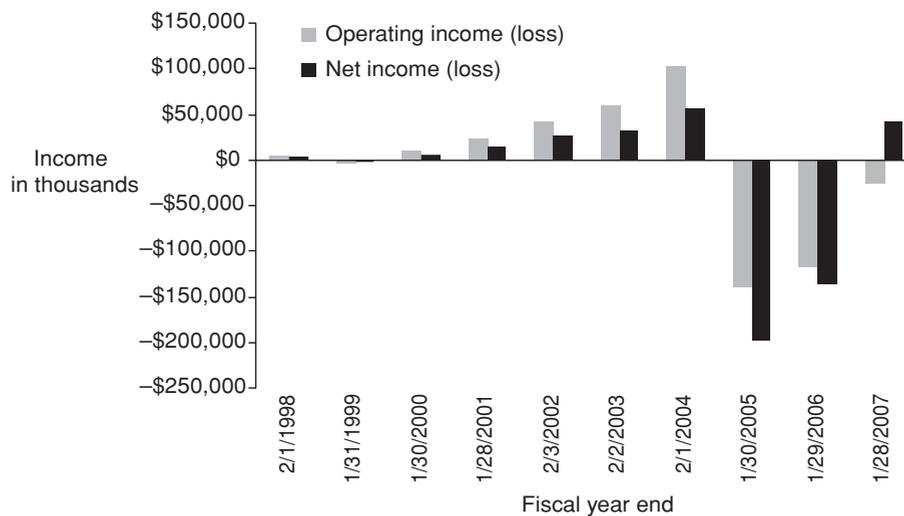


Figure 3 Krispy, Kreme Doughnuts, Inc. Income, 1997–2006
 Source: Krispy Kreme Doughnuts, Inc., 10-K filings, various years.

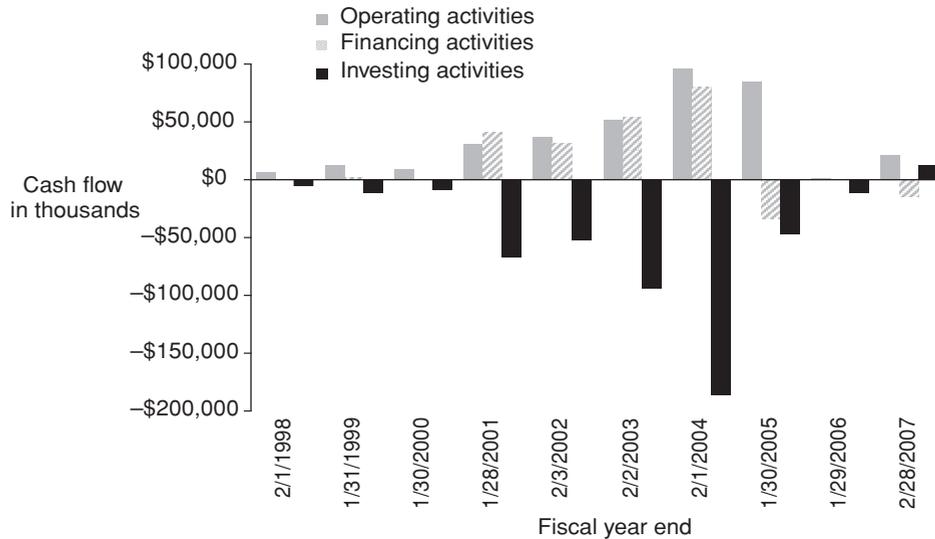


Figure 4 Krispy, Kreme Doughnuts, Inc., Cash Flows, 1997–2006
 Source: Krispy, Kreme Doughnuts, Inc., 10-K filings, various years.

Krispy Kreme's growth just after its IPO was financed by both operating activities and external financing, as we show in Figure 4. However, approximately half of the funds to support its rapid growth and to purchase some of its franchised stores in the 2000–2003 fiscal years came from long-term financing. This resulted in problems as the company's debt burden became almost three times its equity as revenue growth slowed by the 2005 fiscal year. Krispy Kreme demonstrated some ability to turn itself around in the 2006 fiscal year, partly by slowing its expansion through new stores.

Ratio Analysis

One use of cash-flow information is in ratio analysis, primarily with the balance sheet and income statement information. Once such ratio is the cash flow-based ratio, the *cash-flow interest coverage ratio*, which is a measure of financial risk. There are a number of other cash flow-based ratios that an analyst may find useful in evaluating the operating performance and financial condition of a company.

A useful ratio to help further assess a company's cash flow is the *cash flow-to-capital expenditures ratio*, or *capital expenditures coverage ratio*:

ditures ratio, or capital expenditures coverage ratio:

$$\begin{aligned} &\text{Cash flow-to-capital expenditures} \\ &= \frac{\text{Cash flow}}{\text{Capital expenditures}} \end{aligned}$$

The cash-flow measure in the numerator should be one that has not already removed capital expenditures; for example, including free cash flow in the numerator would be inappropriate.

This ratio provides information about the financial flexibility of the company and is particularly useful for capital-intensive firms and utilities (see Fridson, 2002, p. 173). The larger the ratio, the greater the financial flexibility. However, one must carefully examine the reasons why this ratio may be changing over time and why it might be out of line with comparable firms in the industry. For example, a declining ratio can be interpreted in two ways. First, the firm may eventually have difficulty adding to capacity via capital expenditures without the need to borrow funds. The second interpretation is that the firm may have gone through a period of major capital expansion and therefore it will take time for revenues to be generated

that will increase the cash flow from operations to bring the ratio to some normal long-run level.

Another useful cash flow ratio is the *cash flow-to-debt ratio*:

$$\text{Cash flow to debt} = \frac{\text{Cash flow}}{\text{Debt}}$$

where debt can be represented as total debt, long-term debt, or a debt measure that captures a specific range of maturity (e.g., debt maturing in five years). This ratio gives a measure of a company's ability to meet maturing debt obligations. A more specific formulation of this ratio is Fitch's CFAR ratio, which compares a company's three-year average net free cash flow to its maturing debt over the next five years (see McConville, 1996). By comparing the company's average net free cash flow to the expected obligations in the near term (that is, five years), this ratio provides information on the company's credit quality.

Using Cash-Flow Information

The analysis of cash flows provides information that can be used along with other financial data to help assess the financial condition of a company. Consider the cash flow-to-debt ratio calculated using three different measures of cash flow—EBITDA, free cash flow, and cash flow from operations (from the statement of cash flows)—each compared with long-term debt, as shown in Figure 5 for Weirton Steel.

This example illustrates the need to understand the differences among the cash flow measures. The effect of capital expenditures in the 1988–1991 period can be seen by the difference between the free-cash-flow measure and the other two measures of cash flow; both EBITDA and cash flow from operations ignore capital expenditures, which were substantial outflows for this company in the earlier period.

Cash-flow information may help a stock or bond analyst identify companies that may encounter financial difficulties. Consider the study by Largay and Stickney (1980) that an-

alyzed the financial statements of W. T. Grant during the 1966–1974 period preceding its bankruptcy in 1975 and ultimate liquidation. They noted that financial indicators such as profitability ratios, turnover ratios, and liquidity ratios showed some downtrends, but provided no definite clues to the company's impending bankruptcy. A study of cash flows from operations, however, revealed that company operations were causing an increasing drain on cash, rather than providing cash. (For the period investigated, a statement of changes of financial position [on a working capital basis]) was required to be reported prior to 1988.] This necessitated an increased use of external financing, the required interest payments on which exacerbated the cash-flow drain. Cash-flow analysis clearly was a valuable tool in this case since W. T. Grant had been running a negative cash flow from operations for years. Yet none of the traditional ratios discussed above take into account the cash flow from operations. Use of the cash flow-to-capital expenditures ratio and the cash flow-to-debt ratio would have highlighted the company's difficulties.

Dugan and Samson (1996) examined the use of operating cash flow as an early warning signal of a company's potential financial problems. The subject of the study was Allied Products Corporation because for a decade this company exhibited a significant divergence between cash flow from operations and net income. For parts of the period, net income was positive while cash flow from operations was a large negative value. In contrast to W. T. Grant, which went into bankruptcy, the auditor's report in the 1991 Annual Report of Allied Products Corporation did issue a going-concern warning. Moreover, the stock traded in the range of \$2 to \$3 per share. There was then a turnaround of the company by 1995. In its 1995 annual report, net income increased dramatically from prior periods (to \$34 million) and there was a positive cash flow from operations (\$29 million). The stock traded in the \$25 range by the spring of 1996. As with the W. T. Grant study, Dugan and

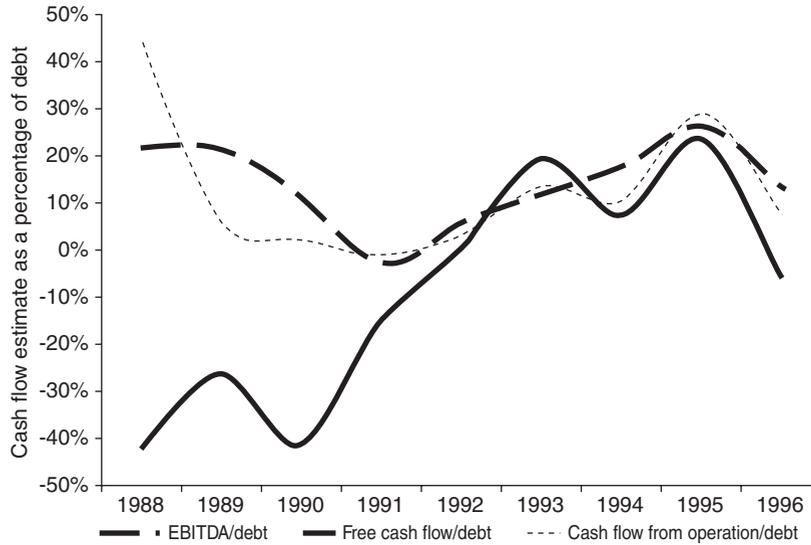


Figure 5 Cash Flow to Debt Using Alternative Estimates of Cash Flow for Weirton Steel, 1988–1996
Source: Weirton Steel's 10-K reports, various years.

Samson (1996) found that the economic realities of a firm are better reflected in its cash flow from operations.

The importance of cash-flow analysis in bankruptcy prediction is supported by the study by Foster and Ward (1997), who compared trends in the statement of cash flows components—cash flow from operations, cash flow for investment, and cash flow for financing—between healthy companies and companies that subsequently sought bankruptcy. They observe that healthy companies tend to have relatively stable relations among the cash flows for the three sources, correcting any given year's deviation from their norm within one year. They also observe that unhealthy companies exhibit declining cash flows from operations and financing and declining cash flows for investment one and two years prior to the bankruptcy. Further, unhealthy companies tend to expend more cash flows to financing sources than they bring in during the year prior to bankruptcy. These studies illustrate the importance of examining cash flow information in assessing the financial condition of a company.

KEY POINTS

- The term “cash flow” has many meanings and the challenge is to determine the cash-flow definition and calculation that is appropriate. The simplest calculation of cash flow is the sum of net income and noncash expenses. This measure, however, does not consider other sources and uses of cash during the period.
- The statement of cash flows provides a useful breakdown of the sources of cash flows: operating activities, investing activities, and financing activities. Though attention is generally focused on the cash flows from operations, what the company does with the cash flows (that is, investing or paying off financing obligations) and what the sources of invested funds are (that is, operations versus external financing) must be investigated. Minor adjustments can be made to the items classified in the statement of cash flows to improve the classification.
- Examination of the different patterns of cash flows is necessary to get a general idea of the activities of the company. For example, a company whose only source of cash flow is

from investing activities, suggesting the sale of property or equipment, may be experiencing financial distress.

- Free cash flow is a company's cash flow that remains after making capital investments that maintain the company's current rate of growth. It is not possible to calculate free cash flow precisely, resulting in many different variations in calculations of this measure. A company that generates free cash flow is not necessarily performing well or poorly; the existence of free cash flow must be taken in context with other financial data and information on the company.
- One of the variations in the calculation of a cash-flow measure is net free cash flow, which is, essentially, free cash flow less any financing obligations. This is a measure of the funds available to service additional obligations to suppliers of capital.

REFERENCES

- Bernstein, L. A. (1999). *Analysis of Financial Statements*. 5th edition. New York: McGraw Hill.
- Dugan, M. T., and Samson, W. D. (1996). Operating cash flow: Early indicators of financial difficulty and recovery. *Journal of Financial Statement Analysis*, Summer: 41–50.
- Eastman, K. (1997). EBITDA: An overrated tool for cash flow analysis. *Commercial Lending Review*, January-February: 64–69.
- Fabozzi, F. J., Drake, P. P., and Polimeni, R. S. (2007). *The Complete CFO Handbook: From Accounting to Accountability*. Hoboken, NJ: John Wiley & Sons.
- Fridson, M. (2002). *Financial Statement Analysis: A Practitioner's Guide*, 3rd edition. New York: John Wiley & Sons.
- Jensen, M. C. (1986). Agency costs of free cash flow, corporate finance, and takeovers. *American Economic Review* 76, 2: 323–329.
- Largay, J. A., and Stickney, C. P. (1980). Cash flows, ratio analysis and the W. T. Grant company bankruptcy. *Financial Analysts Journal* July/August: 51–54.
- McConville, D. J. (1996). Cash flow ratios gains respect as useful tool for credit Rating. *Corporate Cashflow*, January: 18.
- Peterson, P. P., and Fabozzi, F. J. (2012). *Analysis of Financial Statements*, 3rd edition. Hoboken, NJ: John Wiley & Sons.
- Stumpp, P. M. (2001). Critical failings of EBITDA as a cash flow measure. In F. J. Fabozzi (ed.), *Bond Credit Analysis: Framework and Case Studies* (pp. 139–170). Hoboken, NJ: John Wiley & Sons.

Finite Mathematics for Financial Modeling

Important Functions and Their Features

MARKUS HÖCHSTÖTTER, PhD

Assistant Professor, University of Karlsruhe

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Probability theory can be understood as a particular field in mathematics. Hence, it is only to be expected that it relies intensely on theory from analysis and algebra. For example, the fact that the cumulative probability over all values a random variable can assume has to be equal to one is not always feasible to check for without a profound knowledge of mathematics. Continuous probability distributions involve a good deal of analysis and the more sophisticated a distribution is, the more mathematics is necessary to handle it.

In this entry, we review the functions that are used in financial modeling: continuous functions, the indicator function, the derivative of a function, *monotonic functions*, and the integral. Moreover, as special functions, we get to know the factorial, the gamma, beta, and Bessel functions as well as the characteristic function of random variables. (For a more detailed discussion of these functions, see Khuri [2003], MacCluer [2009], and Richardson [2008].)

CONTINUOUS FUNCTION

In this section, we introduce general continuous functions.

General Idea

Let $f(x)$ be a *continuous function* for some real-valued variable x . The general idea behind continuity is that the graph of $f(x)$ does not exhibit gaps. In other words, $f(x)$ can be thought of as

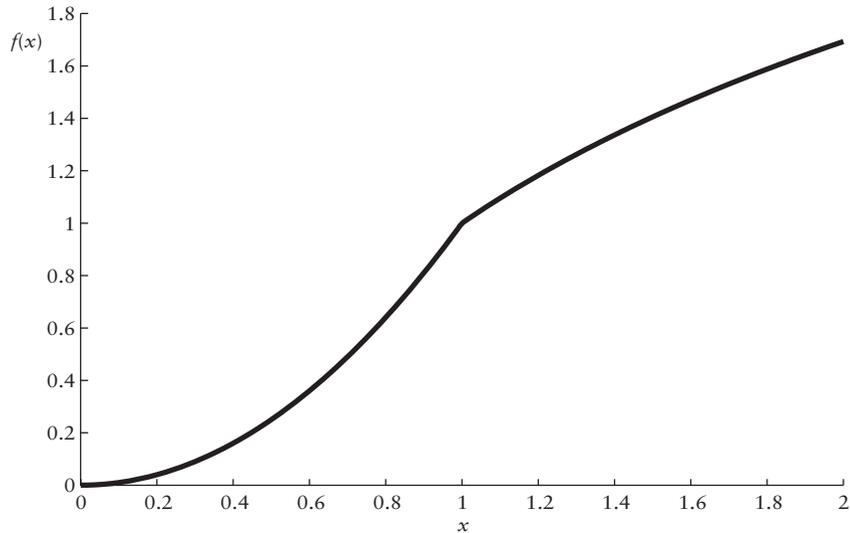


Figure 1 Continuous Function $f(x)$

Note: For $x \in [0,1)$, $f(x) = x^2$ and for $x \in [1,2)$, $f(x) = 1 + \ln(x)$.

being seamless. We illustrate this in Figure 1. For increasing x , from $x = 0$ to $x = 2$, we can move along the graph of $f(x)$ without ever having to jump. In the figure, the graph is generated by the two functions $f(x) = x^2$ for $x \in [0,1)$, and $f(x) = \ln(x) + 1$ for $x \in [1, 2)$.

Note that the function $f(x) = \ln(x)$ is the natural logarithm. It is the inverse function to the

exponential function $g(x) = e^x$ where $e = 2.7183$ is the Euler constant. The inverse has the effect that $f(g(x)) = \ln(e^x) = x$, that is, \ln and e cancel each other out.

A function $f(x)$ is discontinuous if we have to jump when we move along the graph of the function. For example, consider the graph in Figure 2. Approaching $x = 1$ from the left, we

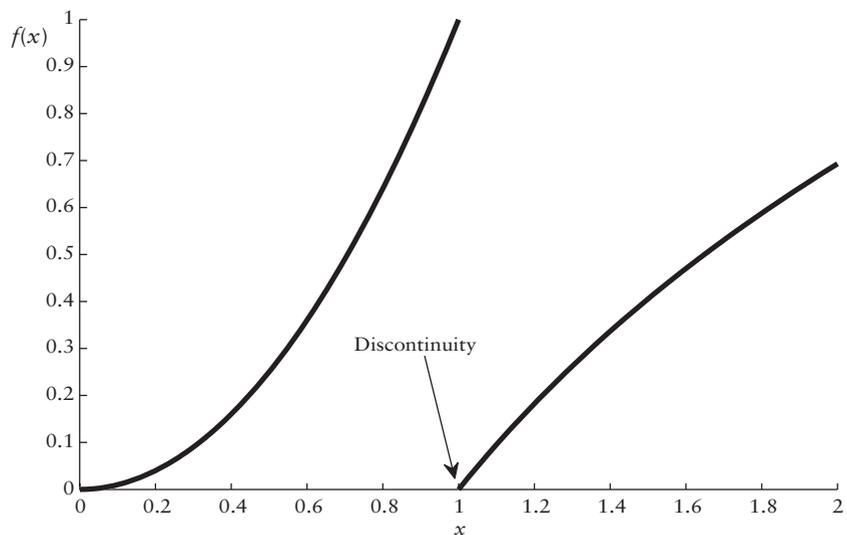


Figure 2 Discontinuous Function $f(x)$

Note: For $x \in [0,1)$, $f(x) = x^2$ and for $x \in [1,2)$, $f(x) = \ln(x)$.

have to jump from $f(x) = 1$ to $f(1) = 0$. Thus, the function f is discontinuous at $x = 1$. Here, f is given by $f(x) = x^2$ for $x \in [0,1)$, and $f(x) = \ln(x)$ for $x \in [1,2)$.

Formal Derivation

For a formal treatment of continuity, we first concentrate on the behavior of f at a particular value x^* .

We say that that a function $f(x)$ is *continuous* at x^* if, for any positive distance δ , we obtain a related distance $\varepsilon(\delta)$ such that

$$f(x^*) - \delta \leq f(x) \leq f(x^*) + \delta, \quad \text{for all } x \in (x^* - \varepsilon(\delta), x^* + \varepsilon(\delta))$$

What does that mean? We use Figure 3 to illustrate. (The function is $f(x) = \sin(x)$ with $x^* = 0.2$.) At x^* , we have the value $f(x^*)$. Now, we select a neighborhood around $f(x^*)$ of some arbitrary distance δ as indicated by the dashed horizontal lines through $f(x^*) - \delta$ and $f(x^*) + \delta$, respectively. From the intersections of these horizontal lines and the function graph (solid line),

we extend two vertical dash-dotted lines down to the x -axis so that we obtain the two values x^L and x^U , respectively. Now, we measure the distance between x^L and x^* and also the distance between x^U and x^* . The smaller of the two yields the distance $\varepsilon(\delta)$. With this distance $\varepsilon(\delta)$ on the x -axis, we obtain the environment $(x^* - \varepsilon(\delta), x^* + \varepsilon(\delta))$ about x^* . (Note that $x^L = x^* - \varepsilon(\delta)$, since the distance between x^L and x^* is the shorter one.) The environment is indicated by the dashed lines extending vertically above $x^* - \varepsilon(\delta)$ and $x^* + \varepsilon(\delta)$, respectively. We require that all x that lie in $(x^* - \varepsilon(\delta), x^* + \varepsilon(\delta))$ yield values $f(x)$ inside of the environment $[f(x^*) - \delta, f(x^*) + \delta]$. We can see by Figure 3 that this is satisfied.

Let us repeat this procedure for a smaller distance δ . We obtain new environments $[f(x^*) - \delta, f(x^*) + \delta]$ and $(x^* - \varepsilon(\delta), x^* + \varepsilon(\delta))$. If, for all x in $(x^* - \varepsilon(\delta), x^* + \varepsilon(\delta))$, the $f(x)$ are inside of $[f(x^*) - \delta, f(x^*) + \delta]$, again, then we can take an even smaller δ . We continue this for successively smaller values of δ just short of becoming 0 or until the condition on the $f(x)$ is no longer satisfied. As we can easily see in Figure 3, we could go on forever and the condition on the $f(x)$

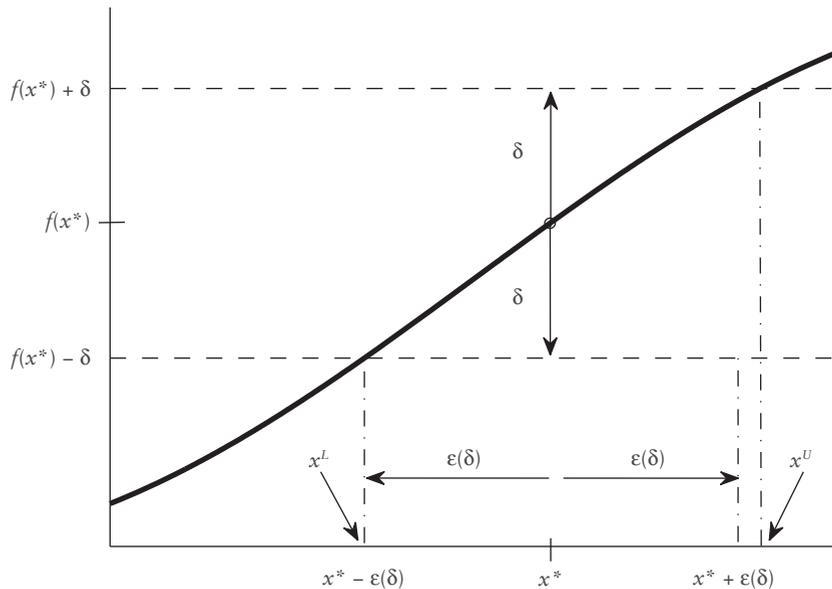


Figure 3 Continuity Criterion
 Note: Function $f = \sin(x)$, for $-1 \leq x \leq 1$.

would always be satisfied. Hence, the graph of f is seamless or continuous at x .

Finally, we say that the function f is *continuous* if it is continuous at all x for which f is defined, that is, in the domain of f . Note that only the domain of f is of interest. For example, the square root function $f(x) = \sqrt{x}$ is only defined for $x \geq 0$. Thus, we do not care about whether f is continuous for any x other than $x \geq 0$.

INDICATOR FUNCTION

The indicator function acts like a switch. Often, it is denoted by $1_A(X)$ where A is the event of interest and X is a random variable. So, $1_A(X)$ is 1 if the event A is true, that is, if X assumes a value in A . Otherwise, $1_A(X)$ is 0. Formally, this is expressed as

$$1_A(X) = \begin{cases} 1 & X \in A \\ 0 & \text{otherwise} \end{cases}$$

Usually, indicator functions are applied if we are interested in whether a certain event has occurred or not. For example, in a simple way,

the value V of a company may be described by a real numbered random variable X on $\Omega = \mathbf{R}$ with a particular probability distribution P . Now, the value V of the company may be equal to X as long as X is greater than 0. In the case where X assumes a negative value or 0, then V is automatically 0, that is, the company is bankrupt. So, the event of interest is $A = [0, \infty)$, that is, we want to know whether X is still positive. Using the indicator function this can be expressed as

$$1_{[0, \infty)}(X) = \begin{cases} 1 & X \in [0, \infty) \\ 0 & \text{otherwise} \end{cases}$$

Finally, the company value can be given as

$$V = 1_{[0, \infty)}(X) \cdot X = \begin{cases} X & X \in [0, \infty) \\ 0 & \text{otherwise} \end{cases}$$

The company value V as a function is depicted in Figure 4. We can clearly detect the kink at $x = 0$ where the indicator function becomes 1 and, hence, $V = X$.

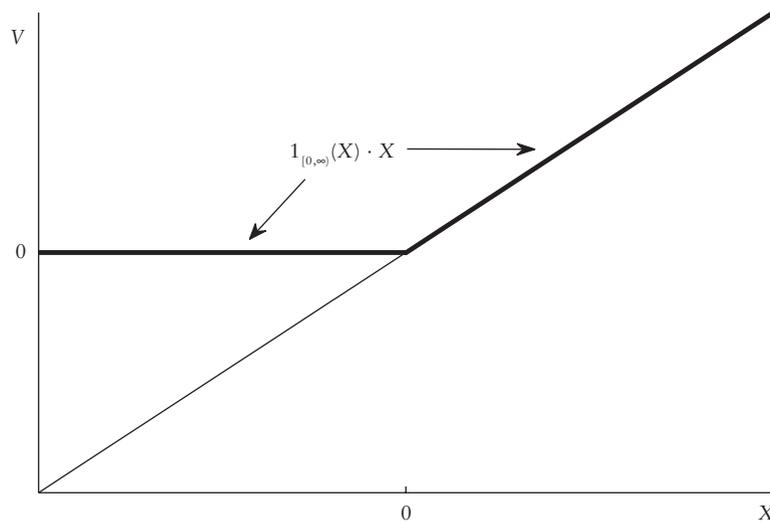


Figure 4 The Company Value V as a Function of the Random Variable X Using the Indicator Function $1_{[0, \infty)}(X) \cdot X$

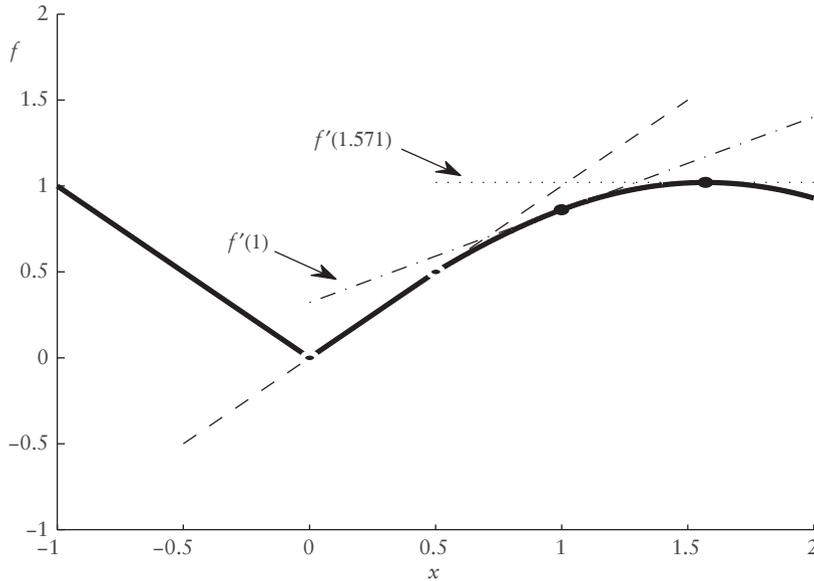


Figure 5 Function f (solid) with Derivatives $f'(x)$ at x , for $0 < x < 0.5$ (dashed), $x = 1$ (dash-dotted), and $x = 1.571$ (dotted)

DERIVATIVES

Suppose we have some continuous function f with the graph given by the solid line in Figure 5. We now might be interested in the growth rate of f at some position x . That is, we might want to know by how much f increases or decreases

when we move from some x by a step of a given size, say Δx , to the right. This difference in f we denote by Δf . This Δ symbol is called delta.

Let us next have a look at the graphs given by the solid lines in Figure 6. These represent the graphs of f and g . The important difference

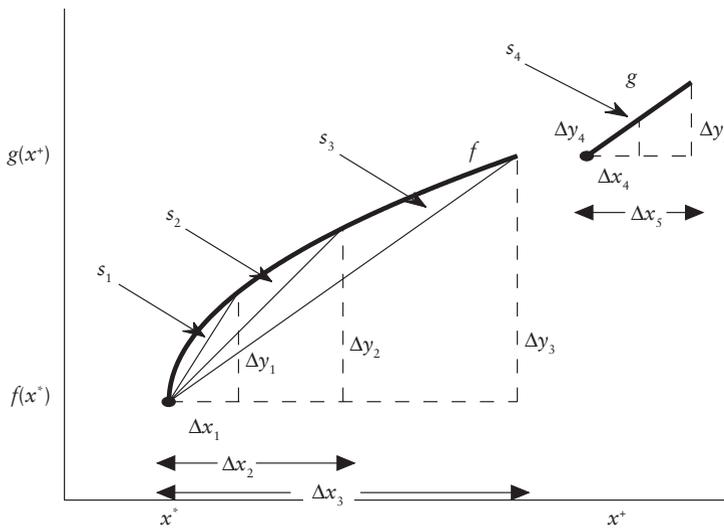


Figure 6 Functions f and g with Slopes Measured at the Points $(x^*, f(x^*))$ and $(x^+, g(x^+))$ Indicated by the \bullet Symbol

between f and g is that, while g is linear, f is not, as can be seen by f 's curvature.

We begin the analysis of the graphs' slopes with function g on the top right of the figure. Let us focus on the point $(x^+, g(x^+))$ given by the solid circle at the lower end of graph g . Now, when we move to the right by Δx_4 along the horizontal dashed line, the corresponding increase in g is given by Δy_4 , as indicated by the vertical dashed line. If, on the other hand, we moved to the right by the longer distance, Δx_5 , the according increment of g would be given by Δy_5 . (This vertical increment Δy_5 is also indicated by a vertical dashed line.) Since g is linear, it has constant slope everywhere and, hence, also at the point $(x^+, f(x^+))$. We denote that slope by s_4 . This implies that the ratios representing the relative increments (i.e., the slopes) have to be equal. That is,

$$s_4 = \frac{\Delta y_4}{\Delta x_4} = \frac{\Delta y_5}{\Delta x_5}$$

Next, we focus on the graph of f on the lower left of Figure 6. Suppose we measured the slope of f at the point $(x^*, f(x^*))$. If we extended a step along the dashed line to the right by Δx_1 , the corresponding increment in f would be Δy_1 , as indicated by the leftmost vertical dashed line. If we moved, instead, by the longer Δx_2 to the right, the corresponding increment in f would be Δy_2 . And a horizontal increment of Δx_3 would result in an increase of f by Δy_3 .

In contrast to the graph of g , the graph of f does not exhibit the property of a constant increment Δy in f per unit step Δx to the right. That is, there is no constant slope of f , which results in the fact that the three ratios of the relative increase of f are different. To be precise, we have

$$\frac{\Delta y_1}{\Delta x_1} > \frac{\Delta y_2}{\Delta x_2} > \frac{\Delta y_3}{\Delta x_3}$$

as can be seen in Figure 6. So, the shorter our step Δx to the right, the steeper the slopes of the thin solid lines through $(x^*, f(x^*))$ and the corresponding points on the curve, $(x^* + \Delta x_1, f(x^* + \Delta x_1))$, $(x^* + \Delta x_2, f(x^* + \Delta x_2))$, and $(x^* + \Delta x_2,$

$f(x^* + \Delta x_2))$, respectively. That means that, the smaller the increment Δx , the higher the relative increment Δy of f . So, finally, if we moved only a minuscule step to the right from $(x^*, f(x^*))$, we would obtain the steepest thin line and, consequently, the highest relative increase in f given by

$$\frac{\Delta y}{\Delta x} \quad (1)$$

By letting Δx approach 0, we obtain the marginal increment, in case the limit of (1) exists (i.e., if the ratio has a finite limit). Formally,

$$\frac{\Delta y}{\Delta x} \xrightarrow{\Delta x \rightarrow 0} s(x) \quad \text{with} \quad -\infty < s(x) < \infty$$

This marginal increment $s(x)$ is different, at any point on the graph of f , while we have seen that it is constant for all points on the graph of g .

Construction of the Derivative

The limit analysis of marginal increments now brings us to the notion of a *derivative* that we discuss next. Earlier we introduced the limit growth rate of some continuous function at some point $(x_0, f(x_0))$. To represent the slope of the line through $(x_0, f(x_0))$ and $(x_0 + \Delta x, f(x_0 + \Delta x))$, we define the difference quotient

$$\frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} \quad (2)$$

If we let $\Delta x \rightarrow 0$, we obtain the limit of the difference quotient (2). If this limit is not finite, then we say that it does not exist. Suppose we were not only interested in the behavior of f when moving Δx to the right but also wanted to analyze the reaction by f to a step Δx to the left. We would then obtain two limits of (2). The first with $\Delta x^+ > 0$ (i.e., a step to the right) would be the upper limit L^U

$$\frac{f(x_0 + \Delta x^+) - f(x_0)}{\Delta x^+} \xrightarrow{\Delta x^+ \rightarrow 0} L^U$$

and the second with $\Delta x^- < 0$ (i.e., a step to the left), would be the lower limit L^L

$$\frac{f(x_0 + \Delta x^-) - f(x_0)}{|\Delta x^-|} \xrightarrow{\Delta x^- \rightarrow 0} L^L$$

If L^U and L^L are equal, $L^U = L^L = L$, then f is said to be differentiable at x_0 . The limit L is the *derivative* of f . We commonly write the derivative in the fashion

$$f'(x_0) = \left. \frac{df(x)}{dx} \right|_{x=x_0} = \left. \frac{dy}{dx} \right|_{x=x_0} \quad (3)$$

On the right side of (3), we have replaced $f(x)$ by the variable y as we will often do, for convenience. If the derivative (3) exists for all x , then f is said to be differentiable.

Let us now return to Figure 5. Recall that the graph of the continuous function f is given by the solid line. We start at $x = -1$. Since f is not continuous at $x = -1$, we omit this end point (1,1) from our analysis. For $-1 < x < 0$, we have that f is constant with slope $s = -1$. Consequently, the derivative $f'(x) = -1$, for these x .

At $x = 0$, we observe that f is linear to the left with $f'(x) = -1$ and that it is also linear to the right, however, with $f'(x) = 1$, for $0 < x < 0.5$. So, at $x = 0$, $L^U = 1$ while $L^L = -1$. Since here $L^U \neq L^L$, the derivative of f does not exist at $x = 0$.

For $0 < x < 0.5$, we have the constant derivative $f'(x) = 1$. The corresponding slope of 1 through (0,0) and (0.5,0.5) is indicated by the dashed line. At $x = 0.5$, the left side limit $L^L = 1$ while the right side limit $L^U = 0.8776$. (This value of $\cos(0.5) = 0.8776$ is a result from calculus.) Hence, the two limits are not equal and, consequently, f is not differentiable at $x = 0.5$.

Without formal proof, we state that f is differentiable for all $0.5 < x < 2$. For example, at $x = 1$, $L^L = L^U = 0.5403$ and, thus, the derivative $f'(1) = 0.5403$. The dash-dotted line indicating this derivative is called the tangent of f at $x = 1$. In Figure 5, the arrow indexed $f'(1)$ points at this tangent. As another example, we select $x = 1.571$ where f assumes its maximum value. Here, the derivative $f'(1.571) = 0$ and, hence,

the tangent at $x = 1.571$ is flat as indicated by the horizontal dotted line. In Figure 5, the arrow indexed $f'(1.571)$ points at this tangent.

MONOTONIC FUNCTION

Suppose we have some function $f(x)$ for real-valued x . For example, the graph of f may look like that in Figure 7. We see that on the interval $[0,1]$, the graph is increasing from $f(0) = 0$ to $f(1) = 1$. For $1 \leq x \leq 2$, the graph remains at the level $f(1) = 1$ like a platform. And, finally, between $x = 2$ and $x = 3$, the graph is increasing, again, from $f(2) = 1$ to $f(3) = 2$.

In contrast, we may have another function, $g(x)$. Its graph is given by Figure 8. It looks somewhat similar to the graph in Figure 7, however, without the platform. The graph of g never remains at a level, but increases constantly. Even for the smallest increments from one value of x , say x_1 , to the next higher, say x_2 , there is always an upward slope in the graph.

Both functions, f and g , never decrease. The distinction is that f is *monotonically increasing* since the graph can remain at some level, while g is *strictly monotonic increasing* since its graph never remains at any level. If we can differentiate f and g , we can express this in terms of the derivatives of f and g . Let f' be the derivative of f and g' the derivative of g . Then, we have the following definitions of continuity for continuous functions with existing derivatives:

Monotonically increasing functions: A continuous function f with derivative f' is monotonically increasing if its derivative $f' \geq 0$.

Strictly monotonic increasing functions: A continuous function g with derivative g' is strictly monotonic increasing if its derivative $g' > 0$.

Analogously, a function $f(x)$ is monotonically decreasing if it behaves in the opposite manner. That is, f never increases when moving from some x to any higher value $x_1 > x$. When f is continuous with derivative f' , then we say that f is *monotonically decreasing* if $f'(x) \leq 0$ and that it is *strictly monotonic increasing* if $f'(x) < 0$ for all x . For these two cases, illustrations are given by

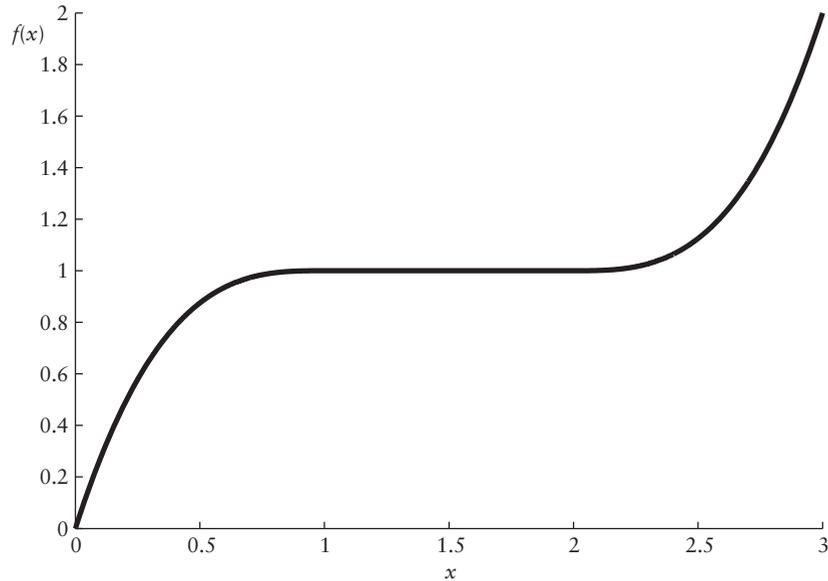


Figure 7 Monotonically Increasing Function f

mirroring the graphs in Figures 7 and 8 against their vertical axes, respectively.

INTEGRAL

Here we derive the concept of integration necessary to understand the probability density and continuous distribution function. The *integral*

of some function over some set of values represents the area between the function values and the horizontal axis. To sketch the idea, we start with an intuitive graphical illustration.

We begin by analyzing the area A between the graph (solid line) of the function $f(t)$ and the horizontal axis between $t = 0$ and $t = T$ in Figure 9. Looking at the graph, it appears

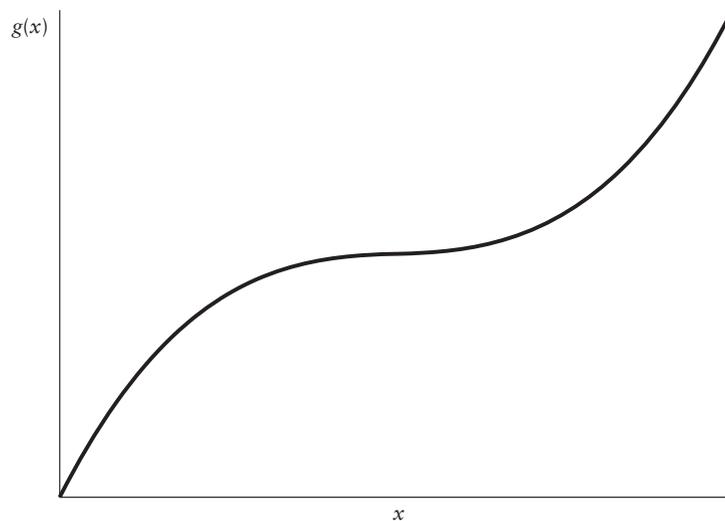


Figure 8 Strictly Monotonic Increasing Function g

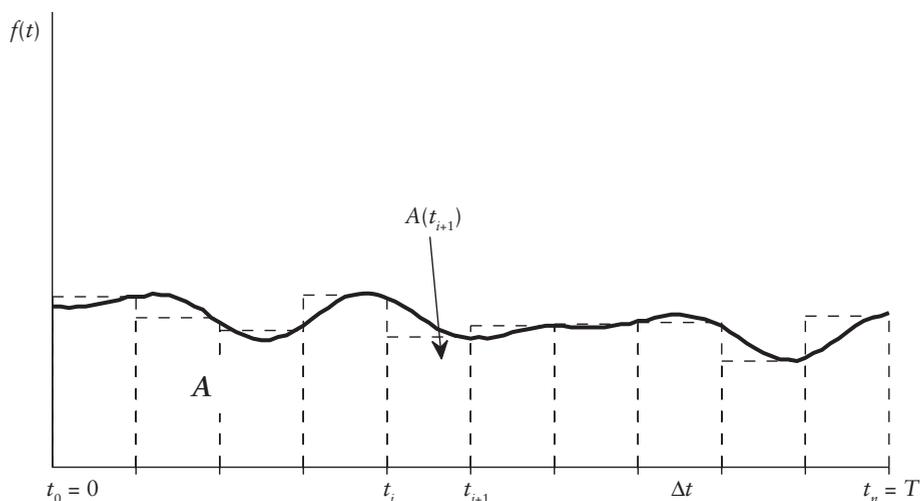


Figure 9 Approximation of the Area A between Graph of $f(t)$ and the Horizontal Axis, for $0 \leq t \leq T$

quite complicated to compute this area A in comparison to, for example, the area of a rectangle where we would only need to know its width and length. However, we can approximate this area by rectangles as will be done next.

Approximation of the Area through Rectangles

Let's approximate the area A under the function graph in Figure 9 as follows. As a first step, we dissect the interval between 0 and T into n equidistant intervals of length $\Delta t = t_{i+1} - t_i$ for $i = 0, 1, \dots, n - 1$. For each such interval, we consider the function value $f(t_{i+1})$ at the rightmost point, t_{i+1} . To obtain an estimate of the area under the graph for the respective interval, we multiply the value $f(t_{i+1})$ at t_{i+1} by the interval width Δt yielding $A(t_{i+1}) = \Delta t \cdot f(t_{i+1})$, which equals the area of the rectangle above interval $i + 1$ as displayed in Figure 9. Finally, we add up the areas $A(t_1), A(t_2), \dots, A(T)$ of all rectangles resulting in the desired estimate of the area A

$$\sum_{i=0}^{n-1} A(t_{i+1}) = \sum_{i=0}^{n-1} \Delta t \cdot f(t_{i+1}) \quad (4)$$

We repeat the just described procedure for decreasing interval widths Δt .

Integral as the Limiting Area

To derive the perfect approximation of the area under the curve in Figure 9, we let the interval width Δt gradually vanish until it almost equals 0, proceeding as before. We denote this infinitesimally small width by the step rate dt . Now, the difference between the function values at either end, that is, $f(t_i)$ and $f(t_{i+1})$, of the interval $i + 1$ will be nearly indistinguishable since t_i and t_{i+1} almost coincide. Hence, the corresponding rectangle with area $A(t_{i+1})$ will turn into a dash with infinitesimally small base dt .

Summation as in equation (4) of the areas of the dashes becomes infeasible. For this purpose, the *integral* has been introduced as the limit of (4) as $\Delta t \rightarrow 0$. (Conditions under which these limits exist are omitted here.) It is denoted by

$$\int_0^T f(t)dt \quad (5)$$

where the limits 0 and T indicate which interval the integration is performed on. In our case, the integration variable is t while the function $f(t)$ is called the integrand. In words, equation (5) is the integral of the function $f(t)$ over t from 0 to T . It is immaterial how we denote the integration variable. The same result as in equation (5)

would result if we wrote

$$\int_0^T f(y)dy$$

instead. The important factors are the integrand and the integral limits.

Note that instead of using the function values of the right boundaries of the intervals $f(t_{i+1})$ in equation (4), referred to as the right-point rule, we might as well have taken the function values of the left boundaries $f(t_i)$, referred to as the left-point rule, which would have led to the same integral. Moreover, we might have taken the function $f(0.5 \cdot (t_{i+1} + t_i))$ values evaluated at the mid-points of the intervals and still obtained the same interval. This latter procedure is called the mid-point rule.

If we keep 0 as the lower limit of the integral in equation (5) and vary T , then equation (5) becomes a function of the variable T . We may denote this function by

$$F(T) = \int_0^T f(t)dt \quad (6)$$

Relationship Between Integral and Derivative

In equation (6) the relationship between $f(t)$ and $F(T)$ is as follows. Suppose we compute the derivative of $F(T)$ with respect to T and assume that $F(T)$ is differentiable, for $T > 0$. The result is

$$F'(T) = \frac{dF(T)}{dT} = f(T) \quad (7)$$

Hence, from equation (7) we see that the marginal increment of the integral at any point (i.e., its derivative) is exactly equal to the integrand evaluated at the according value. This need not generally be true. But in most cases, particularly in financial modeling, this statement is valid.

The implication of this discussion for probability theory is as follows. Let P be a continuous probability measure with probability distribu-

tion function F and (probability) density function f . There is the unique link between f and P given through

$$P(X \leq x) = F(x) = \int_{-\infty}^{\infty} f(x)dx \quad (8)$$

Formally, the integration of f over x is always from $-\infty$ to ∞ , even if the support is not on the entire real line. This is no problem, however, since the density is zero outside the support and, hence, integration over those parts yields 0 contribution to the integral. For example, suppose that some density function were

$$f(x) = \begin{cases} h(x), & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (9)$$

where $h(x)$ is just some function such that f satisfies the requirements for a density function. That is, the support is only on the positive part of the real line. Substituting the function from equation (9) into equation (8) yields the equality

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^{\infty} f(x)dx = \int_0^{\infty} h(x)dx \quad (10)$$

SOME FUNCTIONS

Here we introduce some functions needed in probability theory to describe probability distributions of random variables: factorials, gamma function, beta function, Bessel function of the third kind, and characteristic function. While the first four are functions of very special shape, the characteristic function is of a more general structure. It is the function characterizing the probability distribution of some random variable and, hence, is of unique form for each random variable.

Factorial

Let $k \in \mathbf{N}$ (i.e., $k = 1, 2, \dots$). Then the *factorial* of this natural number k , denoted by the symbol $!$, is given by

$$k! = k \cdot (k - 1) \cdot (k - 2) \cdot \dots \cdot 1 \quad (11)$$

A factorial is the product of this number and all natural numbers smaller than k including 1. By definition, the factorial of zero is one (i.e., $0! \equiv 1$). For example, the factorial of 3 is $3! = 3 \cdot 2 \cdot 1 = 6$.

Gamma Function

The *gamma function* for nonnegative values x is defined by

$$\Gamma(x) = \int_0^{\infty} e^{-t} t^{x-1} dt, \quad x \geq 0 \quad (12)$$

The gamma function has the following properties. If the x correspond with a natural number $n \in N$ (i.e., $n = 1, 2, \dots$), then we have that equation (12) equals the factorial given by equation (11) of $n - 1$. Formally, this is

$$\Gamma(n) = (n - 1)! = (n - 1) \cdot (n - 2) \cdot \dots \cdot 1$$

Furthermore, for any $x \geq 0$, it holds that $\Gamma(x + 1) = x\Gamma(x)$.

In Figure 10, we have displayed part of the gamma function for x values between 0.1 and 5. Note that, for either $x \rightarrow 0$ or $x \rightarrow \infty$, $\Gamma(x)$ goes to infinity.

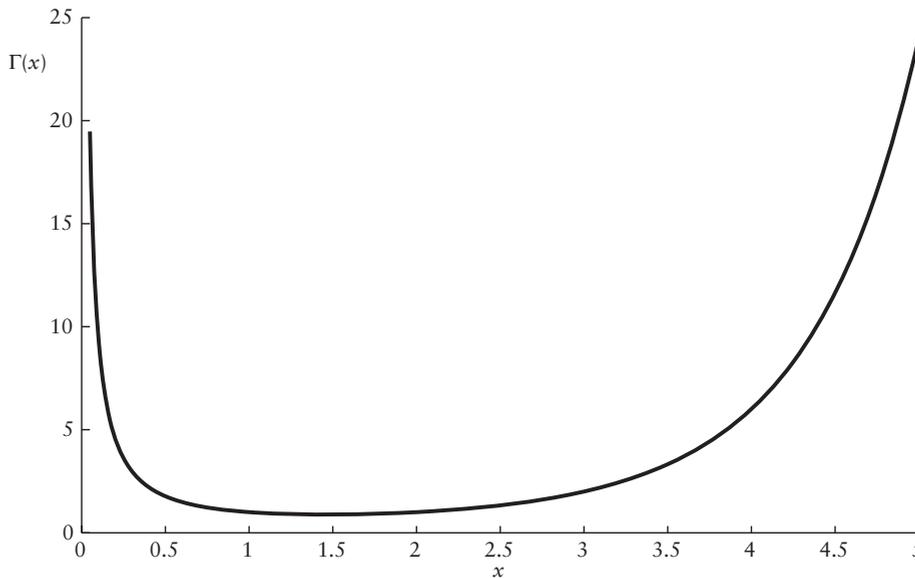


Figure 10 Gamma Function $\Gamma(x)$

Beta Function

The *beta function* with parameters c and d is defined as

$$B(c, d) = \int_0^1 u^{c-1} (1 - u)^{d-1} du = \frac{\Gamma(c)\Gamma(d)}{\Gamma(c + d)}$$

where Γ is the gamma function from equation (12).

Bessel Function of the Third Kind

The *Bessel function* of the third kind is defined as

$$K_1(x) = \frac{1}{2} \int_0^{\infty} \exp \left\{ -\frac{x}{2} \left(y + \frac{1}{y} \right) \right\} dy$$

This function is often a component of other, more complex functions such as the density function of the *NIG* distribution.

Characteristic Function

Before advancing to introduce the *characteristic function*, we briefly explain complex numbers.

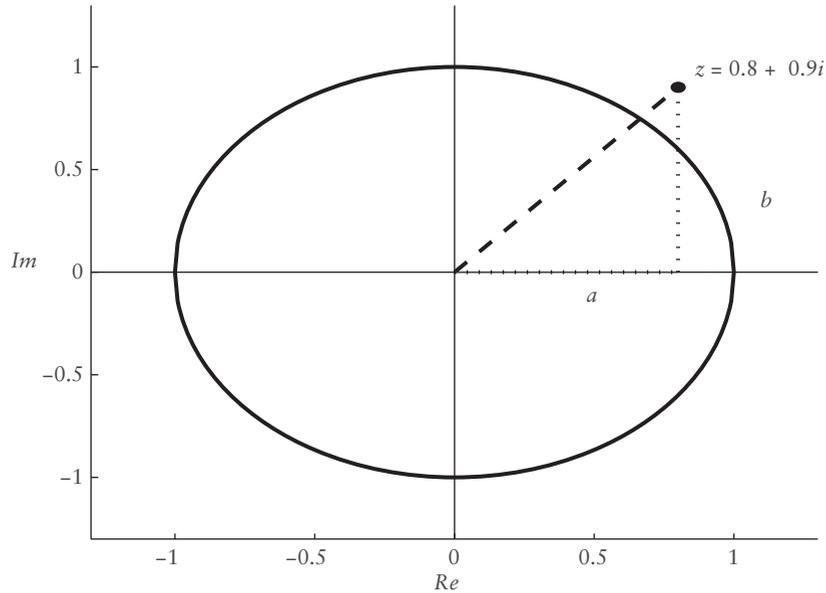


Figure 11 Graphical Representation of the Complex Number $z = 0.8 + 0.9i$

Suppose we were to take the square root of the number -1 , that is, $\sqrt{-1}$. So far, our calculus has no solution for this since the square root of negative numbers has not yet been introduced. However, by introducing the imaginary number i , which is defined as

$$i = \sqrt{-1}$$

we can solve square roots of any real number. Now, we can represent any number as the combination of a real (*Re*) part a plus some units b of i , which we refer to as the imaginary (*Im*) part. Then, any number z will look like

$$z = a + i \cdot b \quad (13)$$

The number given by equation (13) is a complex number. The set of complex numbers is symbolized by \mathbf{C} . This set contains the real numbers that are those complex numbers with $b = 0$. Graphically, we can represent the complex numbers on a two-dimensional space as given in Figure 11.

Now, we can introduce the characteristic function as some function ϕ mapping real numbers into the complex numbers. Formally, we write this as $\phi: \mathbf{R} \rightarrow \mathbf{C}$. Suppose we have some ran-

dom variable X with density function f . The characteristic function is then defined as

$$\phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x) dx \quad (14)$$

which transforms the density f into some complex number at any real position t . Equation (14) is commonly referred to as the Fourier transformation of the density.

The relationship between the characteristic function ϕ and the density function f of some random variable is unique. So, when we state either one, the probability distribution of the corresponding random variable is unmistakably determined.

KEY POINTS

- Continuous functions are an integral component of mathematical analysis. They are useful whenever jumps in the function values are undesirable. This is often the case when financial asset returns are modeled; that is, one assumes that, in particular logarithmic returns, they may assume any value on the

real line such that the related probability distribution is continuous with continuous probability density.

- The indicator function is defined as a function yielding one for certain specified argument values and zero in any other case. It is helpful in expressing so-called exclusive either-or behavior of random variables (i.e., when random variables can only assume exactly one of two values). For example, when one models call option prices where, at maturity, the value of the option is equal to either zero or the difference between the market value of the underlying and the strike price, one resorts to the indicator function.
- The derivative of some function expresses the function's rate of growth at some point for infinitesimally small increments. In words, it expresses by how much the function changes if one takes a very small step. In probability theory, a derivative is used in the context of a continuous probability distribution to express by how much the distribution function increases at a certain value (i.e., the marginal rate of probability at a certain value).
- The integral is the continuous analogue of the sum of discrete values. In probability theory, the probability of individual outcomes is always zero when the distribution is continu-

ous. In order to express the probability of at most a certain value, we cannot sum the individual probabilities of all values less than or equal to the critical value. Instead, at each value, we have the density function which we integrate up to the critical value, yielding the requested probability.

- The characteristic function is the unique representation of a probability distribution. For certain distributions, the probability density function or the distribution function are unknown. Instead, it is necessary to resort to the characteristic function. Technically, the characteristic function is a function involving complex numbers (i.e., numbers including the square root of minus one) to express the behavior of some function at certain frequencies. It is closely linked to the Fourier transform used in engineering.

REFERENCES

- Khuri, A. (2003). *Advanced Calculus with Applications in Statistics*, 2nd ed. Hoboken, NJ: John Wiley & Sons.
- MacCluer, B. D. (2009). *Elementary Functional Analysis*. New York: Springer.
- Richardson, L. F. (2008). *Advanced Calculus: An Introduction to Linear Analysis*. Hoboken, NJ: John Wiley & Sons.

Time Value of Money

PAMELA P. DRAKE, PhD, CFA

J. Gray Ferguson Professor of Finance, College of Business, James Madison University

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Investing decisions require the valuation of investments and the determination of yields on investments. Necessary for the valuation and yield determination are the financial mathematics that involve the time value of money. With these mathematics, future cash flows can be translated to a value in the present, a value today can be converted into a value at some future point in time, and the yield on an investment can be computed.

In this entry, we introduce the mathematical process of translating a value today into a value at some future point in time, and then show how this process can be reversed to determine the value today of some future amount. We then show how to extend the time value of money mathematics to include multiple cash flows and the special cases of annuities and loan amortization. Finally, we demonstrate how these mathematics can be used to calculate the yield on an investment.¹

IMPORTANCE OF THE TIME VALUE OF MONEY

The notion that money has a time value is one of the most basic concepts in investment analysis. Making decisions today regarding future cash flows requires understanding that the value of money does not remain the same throughout time.

A dollar today is worth less than a dollar some time in the future for two reasons:

Reason 1: Cash flows occurring at different points in time have different values relative to any one point in time.

One dollar one year from now is not as valuable as one dollar today. After all, you can invest a dollar today and earn interest so that the value it grows to next year is greater than the one dollar today. This means we have to take into account the *time value of money* to quantify the relation between cash flows at different points in time.

Reason 2: Cash flows are uncertain.

Expected cash flows may not materialize. Uncertainty stems from the nature of forecasts of the timing and/or the amount of cash flows. We do not know for certain when, whether, or how much cash flows will be in the future. This uncertainty regarding future cash flows must somehow be taken

into account in assessing the value of an investment.

Translating a current value into its equivalent future value is referred to as compounding. Translating a future cash flow or value into its equivalent value in a prior period is referred to as discounting. This entry outlines the basic mathematical techniques used in compounding and discounting.

Suppose someone wants to borrow \$100 today and promises to pay back the amount borrowed in one month. Would the repayment of only the \$100 be fair? Probably not. There are two things to consider. First, if the lender didn't lend the \$100, what could he or she have done with it? Second, is there a chance that the borrower may not pay back the loan? So, when considering lending money, we must consider the opportunity cost (that is, what could have been earned or enjoyed), as well as the uncertainty associated with getting the money back as promised.

Let's say that someone is willing to lend the money, but that they require repayment of the \$100 plus some compensation for the opportunity cost and any uncertainty the loan will be repaid as promised. The amount of the loan, the \$100, is the principal. The compensation required for allowing someone else to use the \$100 is the interest.

Looking at this same situation from the perspective of time and value, the amount that you are willing to lend today is the loan's present value. The amount that you require to be paid at the end of the loan period is the loan's future value. Therefore, the future period's value is comprised of two parts:

$$\text{Future value} = \text{Present value} + \text{Interest}$$

The interest is compensation for the use of funds for a specific period. It consists of (1) compensation for the length of time the money is borrowed and (2) compensation for the risk that the amount borrowed will not be repaid exactly as set forth in the loan agreement.

DETERMINING THE FUTURE VALUE

Suppose you deposit \$1,000 into a savings account at the Surety Savings Bank and you are promised 10% interest per period. At the end of one period you would have \$1,100. This \$1,100 consists of the return of your principal amount of the investment (the \$1,000) and the interest or return on your investment (the \$100). Let's label these values:

- \$1,000 is the value today, the present value, PV .
- \$1,100 is the value at the end of one period, the future value, FV .
- 10% is the rate interest is earned in one period, the interest rate, i .

To get to the future value from the present value:

$$FV = \underset{\substack{\uparrow \\ \text{principal}}}{PV} + (PV \times \underset{\substack{\uparrow \\ \text{interest}}}{i})$$

This is equivalent to:

$$FV = PV(1 + i)$$

In terms of our example,

$$\begin{aligned} FV &= \$1,000 + (\$1,000 \times 0.10) \\ &= \$1,000(1 + 0.10) = \$1,100 \end{aligned}$$

If the \$100 interest is withdrawn at the end of the period, the principal is left to earn interest at the 10% rate. Whenever you do this, you earn *simple interest*. It is simple because it repeats itself in exactly the same way from one period to the next as long as you take out the interest at the end of each period and the principal remains the same. If, on the other hand, both the principal and the interest are left on deposit at the Surety Savings Bank, the balance earns interest on the previously paid interest, referred to as *compound interest*. Earning interest on interest is called compounding because the balance at any time is a combination of the principal, interest on principal, and *interest on accumulated interest* (or simply, *interest on interest*).

If you compound interest for one more period in our example, the original \$1,000 grows to \$1,210.00:

$$\begin{aligned}
 FV &= \text{Principal} + \text{First period interest} \\
 &\quad + \text{Second period interest} \\
 &= \$1,000.00 + (\$1,000.00 \times 0.10) \\
 &\quad + (\$1,100.00 \times 0.10) \\
 &= \$1,200.00
 \end{aligned}$$

The present value of the investment is \$1,000, the interest earned over two years is \$210, and the future value of the investment after two years is \$1,210.

The relation between the present value and the future value after two periods, breaking out the second period interest into interest on the principal and interest on interest, is:

$$\begin{aligned}
 FV &= \underset{\substack{\uparrow \\ \text{Principal}}}{PV} + \underset{\substack{\uparrow \\ \text{First} \\ \text{period's} \\ \text{interest on} \\ \text{the principal}}}{(PV \times i)} + \underset{\substack{\uparrow \\ \text{Second} \\ \text{period's} \\ \text{interest on} \\ \text{the principal}}}{(PV \times i)} + \underset{\substack{\uparrow \\ \text{Second} \\ \text{period's} \\ \text{interest on} \\ \text{the first} \\ \text{period's} \\ \text{interest}}}{(PV \times i \times i)}
 \end{aligned}$$

or, collecting the PVs from each term and applying a bit of elementary algebra,

$$FV = PV(1 + 2i + i^2) = PV(1 + i)^2$$

The balance in the account two years from now, \$1,210, comprises three parts:

1. The principal, \$1,000.
2. Interest on principal, \$100 in the first period plus \$100 in the second period.
3. Interest on interest, 10% of the first period's interest, or \$10.

To determine the future value with compound interest for more than two periods, we follow along the same lines:

$$FV = PV(1 + i)^N \tag{1}$$

The value of N is the number of compounding periods, where a compounding period is the unit of time after which interest is paid at the rate i . A period may be any length of time: a minute, a day, a month, or a year. The important thing is to make sure the same compound-

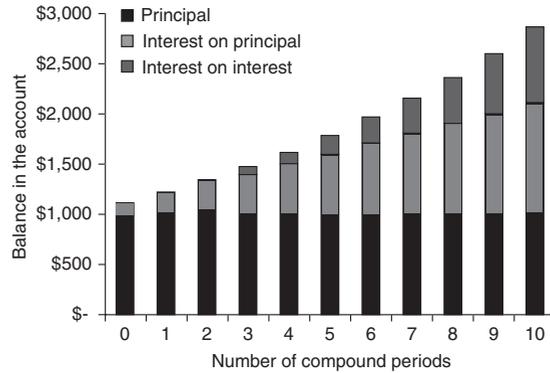


Figure 1 The Value of \$1,000 Invested 10 Years in an Account That Pays 10% Compounded Interest per Year

ing period is reflected throughout the problem being analyzed. The term “ $(1 + i)^N$ ” is referred to as the compound factor. It is the rate of exchange between present dollars and dollars N compounding periods into the future. Equation (1) is the basic valuation equation—the foundation of financial mathematics. It relates a value at one point in time to a value at another point in time, considering the compounding of interest.

The relation between present and future values for a principal of \$1,000 and interest of 10% per period through 10 compounding periods is shown graphically in Figure 1. For example, the value of \$1,000, earning interest at 10% per period, is \$2,593.70 ten periods into the future:

$$\begin{aligned}
 FV &= \$1,000 (1 + 0.10)^{10} = \$1,000 (2.5937) \\
 &= \$2,593.70
 \end{aligned}$$

As you can see in this figure the \$2,593.70 balance in the account at the end of 10 periods is comprised of three parts:

1. The principal, \$1,000.
2. Interest on the principal of \$1,000, which is \$100 per period for 10 periods or \$1,000.
3. Interest on interest totaling \$593.70.

We can express the change in the value of the savings balance (that is, the difference between the ending value and the beginning value) as a growth rate. A *growth rate* is the rate at which a value appreciates (a positive growth) or

depreciates (a negative growth) over time. Our \$1,000 grew at a rate of 10% per year over the 10-year period to \$2,593.70. The average annual growth rate of our investment of \$1,000 is 10%—the value of the savings account balance increased 10% per year.

We could also express the appreciation in our savings balance in terms of a return. A *return* is the income on an investment, generally stated as a change in the value of the investment over each period divided by the amount at the investment at the beginning of the period. We could also say that our investment of \$1,000 provides an average annual return of 10% per year. The average annual return is not calculated by taking the change in value over the entire 10-year period ($\$2,593.70 - \$1,000$) and dividing it by \$1,000. This would produce an *arithmetic average return* of 159.37% over the 10-year period, or 15.937% per year. But the arithmetic average ignores the process of compounding. The correct way of calculating the average annual return is to use a *geometric average return*:

$$i = \sqrt[N]{\frac{FV}{PV}} - 1 \quad (2)$$

which is a rearrangement of equation (1) Using the values from the example,

$$i = \sqrt[10]{\frac{\$2,593.70}{\$1,000.00}} - 1 = \left(\frac{\$2,593.70}{\$1,000.00}\right)^{1/10} - 1 = 1.100 - 1 = 10\%$$

Therefore, the annual return on the investment—sometimes referred to as the *compound average annual return* or the *true return*—is 10% per year.

Here is another example for calculating a future value. A common investment product of a life insurance company is a guaranteed investment contract (GIC). With this investment, an insurance company guarantees a specified interest rate for a period of years. Suppose that the life insurance company agrees to pay 6% annually for a five-year GIC and the amount invested by the policyholder is \$10 million. The amount of the liability (that is, the amount this

life insurance company has agreed to pay the GIC policyholder) is the future value of \$10 million when invested at 6% interest for five years. In terms of equation (1), $PV = \$10,000,000$, $i = 6\%$, and $N = 5$, so that the future value is:

$$FV = \$10,000,000 (1 + 0.06)^5 = \$13,382,256$$

Compounding More Than One Time per Year

An investment may pay interest more than one time per year. For example, interest may be paid semiannually, quarterly, monthly, weekly, or daily, even though the stated rate is quoted on an annual basis. If the interest is stated as, say, 10% per year, compounded semiannually, the nominal rate—often referred to as the *annual percentage rate* (APR)—is 10%. The basic valuation equation handles situations in which there is compounding more frequently than once a year if we translate the nominal rate into a rate per compounding period. Therefore, an APR of 10% with compounding semiannually is 5% per period—where a period is six months—and the number of periods in one year is 2.

Consider a deposit of \$50,000 in an account for five years that pays 8% interest, compounded quarterly. The interest rate per period, i , is $8\%/4 = 2\%$ and the number of compounding periods is $5 \times 4 = 20$. Therefore, the balance in the account at the end of five years is:

$$FV = \$50,000(1 + 0.02)^{20} = \$50,000(1.4859474) = \$74,297.37$$

As shown in Figure 2, through 50 years with both annual and quarterly compounding, the investment's value increases at a faster rate with the increased frequency of compounding.

The last example illustrates the need to correctly identify the “period” because this dictates the interest rate per period and the number of compounding periods. Because interest rates are often quoted in terms of an APR, we need to be able to translate the APR into an interest rate per period and to adjust the number of

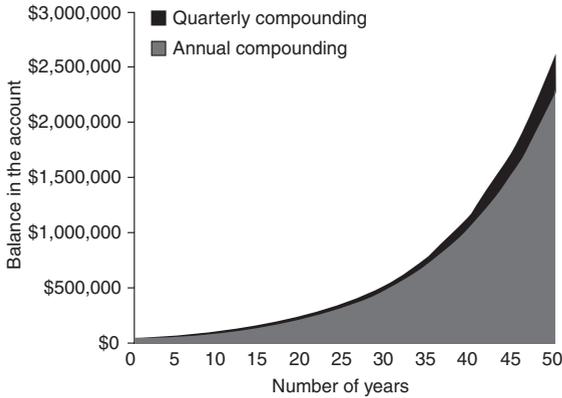


Figure 2 Value of \$50,000 Invested in the Account that Pays 8% Interest per Year: Quarterly versus Annual Compounding

periods. To see how this works, let’s use an example of a deposit of \$1,000 in an account that pays interest at a rate of 12% per year, with interest compounded for different compounding frequencies. How much is in the account after, say, five years depends on the compounding frequency:

Com- pounding Frequency	Period	Rate per Compound- ing Period, i	Number of Periods in Five Years, N	FV at the End of Five Years
Annual	One year	12%	5	\$1,762.34
Semiannual	Six months	6%	10	1,790.85
Quarterly	Three months	3%	20	1,806.11
Monthly	One month	1%	60	1,816.70

As you can see, both the rate per period, i , and the number of compounding periods, N , are adjusted and depend on the frequency of compounding. Interest can be compounded for any frequency, such as daily or hourly.

Let’s work through another example for compounding with compounding more than once a year. Suppose we invest \$200,000 in an investment that pays 4% interest per year, compounded quarterly. What will be the future value of this investment at the end of 10 years?

The given information is $i = 4\%/4 = 1\%$ and $N = 10 \times 4 = 40$ quarters. Therefore,

$$FV = \$200,000(1 + 0.01)^{40} = \$297,772.75$$

Continuous Compounding

The extreme frequency of compounding is *continuous compounding*—interest is compounded instantaneously. The factor for compounding continuously for one year is e^{APR} , where e is 2.71828 . . . , the base of the natural logarithm. And the factor for compounding continuously for two years is $e^{APR} e^{APR}$ or e^{2APR} . The future value of an amount that is compounded continuously for N years is:

$$FV = PVe^{N(APR)} \tag{3}$$

where APR is the annual percentage rate and $e^{N(APR)}$ is the compound factor.

If \$1,000 is deposited in an account for five years with interest of 12% per year, compounded continuously,

$$\begin{aligned} FV &= \$1,000e^{5(0.12)} = \$1,000(e^{0.60}) \\ &= \$1,000(1.82212) = \$1,822.12 \end{aligned}$$

Comparing this future value with that if interest is compounded annually at 12% per year for five years, \$1,762.34, we see the effects of this extreme frequency of compounding.

Multiple Rates

In our discussion thus far, we have assumed that the investment will earn the same periodic interest rate, i . We can extend the calculation of a future value to allow for different interest rates or growth rates for different periods. Suppose an investment of \$10,000 pays 9% during the first year and 10% during the second year. At the end of the first period, the value of the investment is \$10,000 $(1 + 0.09)$, or \$10,900. During the second period, this \$10,900 earns interest at 10%. Therefore, the future value of this \$10,000 at the end of the second period is:

$$FV = \$10,000(1 + 0.09)(1 + 0.10) = \$11,990$$

We can write this more generally as:

$$FV = PV(1 + i_1)(1 + i_2)(1 + i_3) \dots (1 + i_N) \tag{4}$$

where i_N is the interest rate for period N .

Consider a \$50,000 investment in a one-year bank certificate of deposit (CD) today and rolled over annually for the next two years into one-year CDs. The future value of the \$50,000 investment will depend on the one-year CD rate each time the funds are rolled over. Assuming that the one-year CD rate today is 5% and that it is expected that the one-year CD rate one year from now will be 6%, and the one-year CD rate two years from now will be 6.5%, then we know:

$$\begin{aligned} FV &= \$50,000(1 + 0.05)(1 + 0.06)(1 + 0.065) \\ &= \$59,267.25 \end{aligned}$$

Continuing this example, what is the average annual interest rate over this period? We know that the future value is \$59,267.25, the present value is \$50,000, and $N = 3$:

$$i = \sqrt[3]{\frac{\$59,267.25}{\$50,000.00}} - 1 = \sqrt[3]{1.185345} - 1 = 5.8315\%$$

which is also:

$$\begin{aligned} i &= \sqrt[3]{(1 + 0.05) + (1 + 0.06)(1 + 0.065)} - 1 \\ &= 5.8315\% \end{aligned}$$

DETERMINING THE PRESENT VALUE

Now that we understand how to compute future values, let's work the process in reverse. Suppose that for borrowing a specific amount of money today, the Yenom Company promises to pay lenders \$5,000 two years from today. How much should the lenders be willing to lend Yenom in exchange for this promise? This dilemma is different than figuring out a future value. Here we are given the future value and have to figure out the present value. But we can use the same basic idea from the future value problems to solve present value problems.

If you can earn 10% on other investments that have the same amount of uncertainty as the \$5,000 Yenom promises to pay, then:

- The future value, $FV = \$5,000$.
- The number of compounding periods, $N = 2$.
- The interest rate, $i = 10\%$.

We also know the basic relation between the present and future values:

$$FV = PV(1 + i)^N$$

Substituting the known values into this equation:

$$\$5,000 = PV(1 + 0.10)^2$$

To determine how much you are willing to lend now, PV , to get \$5,000 one year from now, FV , requires solving this equation for the unknown present value:

$$\begin{aligned} PV &= \frac{\$5,000}{(1 + 0.10)^2} = \$5,000 \left(\frac{1}{1 + 0.10} \right)^2 \\ &= \$5,000(0.82645) = \$4,132.25 \end{aligned}$$

Therefore, you would be willing to lend \$4,132.25 to receive \$5,000 one year from today if your opportunity cost is 10%. We can check our work by reworking the problem from the reverse perspective. Suppose you invested \$4,132.25 for two years and it earned 10% per year. What is the value of this investment at the end of the year?

We know: $PV = \$4,132.25$, $N = 10\%$ or 0.10, and $i = 2$.

Therefore, the future value is:

$$\begin{aligned} FV &= PV(1 + i)^N = \$4,132.25 (1 + 0.10)^2 \\ &= \$5,000.00 \end{aligned}$$

Compounding translates a value in one point in time into a value at some future point in time. The opposite process translates future values into present values: Discounting translates a value back in time. From the basic valuation equation:

$$FV = PV(1 + i)^N$$

we divide both sides by $(1 + i)^N$ and exchange sides to get the present value,

$$PV = \frac{FV}{(1 + i)^N} \quad (5)$$

$$\text{or } PV = FV \left(\frac{1}{1 + i} \right)^N \quad \text{or } PV = FV \left[\frac{1}{(1 + i)^N} \right]$$

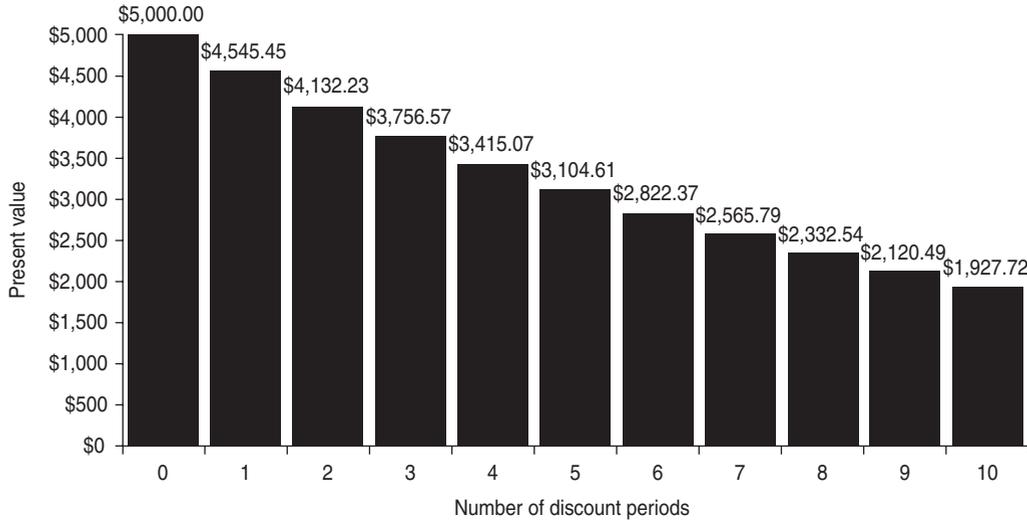


Figure 3 Present Value of \$5,000 Discounted at 10%

The term in brackets [] is referred to as the *discount factor* since it is used to translate a future value to its equivalent present value. The present value of \$5,000 for discount periods ranging from 0 to 10 is shown in Figure 3.

If the frequency of compounding is greater than once a year, we make adjustments to the rate per period and the number of periods as we did in compounding. For example, if the future value five years from today is \$100,000 and the interest is 6% per year, compounded semiannually, $i = 6\%/2 = 3\%$ and $N = 5 \times 2 = 10$, and the present value is:

$$PV = \$100,000(1 + 0.03)^{10} = \$100,000(1.34392) = \$134,392$$

Here is an example of calculating a present value. Suppose that the goal is to have \$75,000 in an account by the end of four years. And suppose that interest on this account is paid at a rate of 5% per year, compounded semiannually. How much must be deposited in the account today to reach this goal? We are given $FV = \$75,000$, $i = 5\%/2 = 2.5\%$ per six months, and $N = 4 \times 2 = 8$ six-month periods. The amount of the required deposit is therefore:

$$PV = \frac{\$75,000}{(1 + 0.025)^8} = \$61,555.99$$

DETERMINING THE UNKNOWN INTEREST RATE

As we saw earlier in our discussion of growth rates, we can rearrange the basic equation to solve for i :

$$i = \sqrt[N]{\frac{FV}{PV}} - 1 = \left(\frac{FV}{PV}\right)^{1/N} - 1$$

As an example, suppose that the value of an investment today is \$100 and the expected value of the investment in five years is expected to be \$150. What is the annual rate of appreciation in value of this investment over the five-year period?

$$i = \sqrt[5]{\frac{\$150}{\$100}} - 1 = \sqrt[5]{1.5} - 1 = 0.0845 \text{ or } 8.45\% \text{ per year}$$

There are many applications in finance where it is necessary to determine the rate of change in values over a period of time. If values are increasing over time, we refer to the rate of change as the growth rate. To make comparisons easier, we usually specify the growth rate as a rate per year.

For example, if we wish to determine the rate of growth in these values, we solve for the unknown interest rate. Consider the growth rate of

dividends for General Electric. General Electric pays dividends each year. In 1996, for example, General Electric paid dividends of \$0.317 per share of its common stock, whereas in 2006 the company paid \$1.03 in dividends per share. This represents a growth rate of 12.507%:

$$\begin{aligned}\text{Growth rate of dividends} &= \sqrt[10]{\frac{\$1.03}{\$0.317}} - 1 \\ &= \sqrt[10]{3.2492} - 1 \\ &= 12.507\%\end{aligned}$$

The 12.507% is the average annual rate of the growth during this 10-year span.

DETERMINING THE NUMBER OF COMPOUNDING PERIODS

Given the present and future values, calculating the number of periods when we know the interest rate is a bit more complex than calculating the interest rate when we know the number of periods. Nevertheless, we can develop an equation for determining the number of periods, beginning with the valuation formula given by equation (1) and rearranging to solve for N ,

$$N = \frac{\ln FV - \ln PV}{\ln(1 + i)} \quad (6)$$

where \ln indicates the natural logarithm, which is the log of the base e . (e is approximately equal to 2.718. The natural logarithm function can be found on most calculators, usually indicated by “ \ln ”.)

Suppose that the present value of an investment is \$100 and you wish to determine how long it will take for the investment to double in value if the investment earns 6% per year, compounded annually:

$$\begin{aligned}N &= \frac{\ln 200 - \ln 100}{\ln 1.06} = \frac{5.2983 - 4.6052}{0.0583} \\ &= 11.8885 \text{ or approximately } 12 \text{ years}\end{aligned}$$

You’ll notice that we round off to the next whole period. To see why, consider this last example. After 11.8885 years, we have doubled our money if interest were paid 88.85% the way through the 12th year. But, we stated earlier that interest is paid at the end of each period—not part of the way through. At the end of the 11th year, our investment is worth \$189.93, and at the end of the 12th year, our investment is worth \$201.22. So, our investment’s value doubles by the 12th period—with a little extra, \$1.22.

THE TIME VALUE OF A SERIES OF CASH FLOWS

Applications in finance may require the determination of the present or future value of a series of cash flows rather than simply a single cash flow. The principles of determining the future value or present value of a series of cash flows are the same as for a single cash flow, yet the math becomes a bit more cumbersome.

Suppose that the following deposits are made in a Thrifty Savings and Loan account paying 5% interest, compounded annually:

Time When Deposit Is Made	Amount of Deposit
Today	\$1,000
At the end of the first year	2,000
At the end of the second year	1,500

What is the balance in the savings account at the end of the second year if no withdrawals are made and interest is paid annually?

Let’s simplify any problem like this by referring to today as the end of period 0, and identifying the end of the first and each successive period as 1, 2, 3, and so on. Represent each end-of-period cash flow as “ CF ” with a subscript specifying the period to which it corresponds. Thus, CF_0 is a cash flow today, CF_{10} is a cash flow at the end of period 10, and CF_{25} is a cash flow at the end of period 25, and so on.

Representing the information in our example using cash flow and period notation:

Period	Cash Flow	End-of-Period Cash Flow
0	CF_0	\$1,000
1	CF_1	\$2,000
2	CF_2	\$1,500

The future value of the series of cash flows at the end of the second period is calculated as follows:

Period	End-of-Period Cash Flow	Number of Periods Interest Is Earned	Compounding Factor	Future Value
0	\$1,000	2	1.1025	\$1,102.50
1	2,000	1	1.0500	2,100.00
2	1,500	0	1.0000	1,500.00
				\$4,702.50

The last cash flow, \$1,500, was deposited at the very end of the second period—the point of time at which we wish to know the future value of the series. Therefore, this deposit earns no interest. In more formal terms, its future value is precisely equal to its present value.

Today, the end of period 0, the balance in the account is \$1,000 since the first deposit is made but no interest has been earned. At the end of period 1, the balance in the account is \$3,050, made up of three parts:

1. The first deposit, \$1,000.
2. \$50 interest on the first deposit.
3. The second deposit, \$2,000.

The balance in the account at the end of period 2 is \$4,702.50, made up of five parts:

1. The first deposit, \$1,000.
2. The second deposit, \$2,000.
3. The third deposit, \$1,500.
4. \$102.50 interest on the first deposit, \$50 earned at the end of the first period, \$52.50 more earned at the end of the second period.
5. \$100 interest earned on the second deposit at the end of the second period.

These cash flows can also be represented in a time line. A time line is used to help graph-

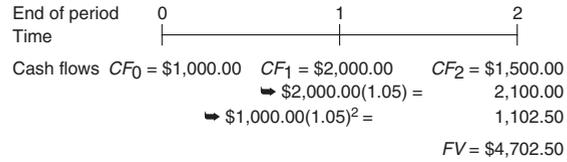


Figure 4 Time Line for the Future Value of a Series of Uneven Cash Flows Deposited to Earn 5% Compounded Interest per Period

ically depict and sort out each cash flow in a series. The time line for this example is shown in Figure 4. From this example, you can see that the future value of the entire series is the sum of each of the compounded cash flows comprising the series. In much the same way, we can determine the future value of a series comprising any number of cash flows. And if we need to, we can determine the future value of a number of cash flows before the end of the series.

For example, suppose you are planning to deposit \$1,000 today and at the end of each year for the next ten years in a savings account paying 5% interest annually. If you want to know the future value of this series after four years, you compound each cash flow for the number of years it takes to reach four years. That is, you compound the first cash flow over four years, the second cash flow over three years, the third over two years, the fourth over one year, and the fifth you don't compound at all because you will have just deposited it in the bank at the end of the fourth year.

To determine the present value of a series of future cash flows, each cash flow is discounted back to the present, where the beginning of the first period, today, is designated as 0. As an example, consider the Thrifty Savings & Loan problem from a different angle. Instead of calculating what the deposits and the interest on these deposits will be worth in the future, let's calculate the present value of the deposits. The present value is what these future deposits are worth today.

In the series of cash flows of \$1,000 today, \$2,000 at the end of period 1, and \$1,500 at

the end of period 2, each are discounted to the present, 0, as follows:

Period	End-of-Period Cash Flow	Number of Periods of Discounting	Discount Factor	Present Value
0	\$1,000	0	1.00000	\$1,000.00
1	\$2,000	1	0.95238	1,904.76
2	\$1,500	2	0.90703	1,360.54
				$FV = \$4,265.30$

The present value of the series is the sum of the present value of these three cash flows, \$4,265.30. For example, the \$1,500 cash flow at the end of period 2 is worth \$1,428.57 at the end of the first period and is worth \$1,360.54 today.

The present value of a series of cash flows can be represented in notation form as:

$$PV = CF_0 \left(\frac{1}{1+i} \right)^0 + CF_1 \left(\frac{1}{1+i} \right)^1 + CF_2 \left(\frac{1}{1+i} \right)^2 + \cdots + CF_N \left(\frac{1}{1+i} \right)^N$$

For example, if there are cash flows today and at the end of periods 1 and 2, today's cash flow is not discounted, the first period cash flow is discounted one period, and the second period cash flow is discounted two periods.

We can represent the present value of a series using summation notation as shown below:

$$PV = \sum_{t=0}^N CF_t \left(\frac{1}{1+i} \right)^t \quad (7)$$

This equation tells us that the present value of a series of cash flows is the sum of the products of each cash flow and its corresponding discount factor.

Shortcuts: Annuities

There are valuation problems that require us to evaluate a series of level cash flows—each cash flow is the same amount as the others—received at regular intervals. Let's suppose you expect to deposit \$2,000 at the end of each of the next four years in an account earning 8% compounded

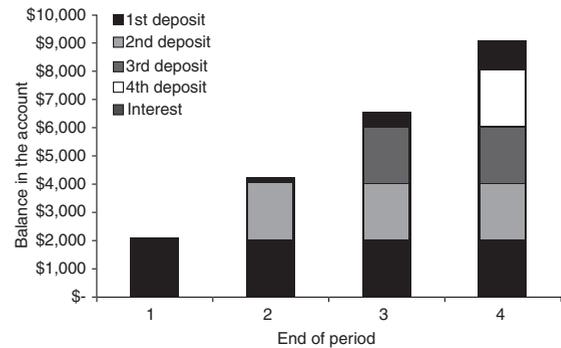


Figure 5 Balance in an Account in Which Deposits of \$2,000 Each Are Made Each Year (The Balance in the Account Earns 8%)

interest. How much will you have available at the end of the fourth year?

As we just did for the future value of a series of uneven cash flows, we can calculate the future value (as of the end of the fourth year) of each \$2,000 deposit, compounding interest at 8%:

$$\begin{aligned} FV &= \$2,000(1 + 0.08)^3 + \$2,000(1 + 0.08)^2 \\ &\quad + \$2,000(1 + 0.08)^1 + \$2,000(1 + 0.08)^0 \\ &= \$2,519.40 + \$2,332.80 + \$2,160.00 \\ &\quad + \$2,000 = \$9,012.20 \end{aligned}$$

Figure 5 shows the contribution of each deposit and the accumulated interest at the end of each period.

- At the end of the first year, there is \$2,000.00 in the account because you have just made your first deposit.
- At the end of the second, there is \$4,160.00 in the account: two deposits of \$2,000 each, plus \$160 interest (8% of \$2,000).
- At the end of the third year, there is \$6,492.80 in the account: three deposits of \$2,000.00 each, plus accumulated interest of \$492.80 [$\$160.00 + (0.08 \times \$4,000) + (0.08 \times \$160)$].
- At the end of the fourth year, you would have \$9,012.20 available: four deposits of \$2,000 each, plus \$1,012.20 accumulated interest [$\$160.00 + \$492.80 + (0.08 \times \$6,000) + (0.08 \times (\$160.00 + 492.80))$].

Notice that in our calculations, each deposit of \$2,000 is multiplied by a factor that corresponds to an interest rate of 8% and the number of periods that the deposit has been in the savings account. Since the deposit of \$2,000 is common to each multiplication, we can simplify the math a bit by multiplying the \$2,000 by the sum of the factors to get the same answer:

$$\begin{aligned} FV &= \$2,000(1.2597) + \$2,000(1.1664) \\ &\quad + \$2,000(1.0800) + \$2,000(1.0000) \\ &= \$9,012.20 \end{aligned}$$

A series of cash flows of equal amount occurring at even intervals is referred to as an annuity. Determining the value of an annuity, whether compounding or discounting, is simpler than valuing uneven cash flows. If each CF_t is equal (that is, all the cash flows are the same value) and the first one occurs at the end of the first period ($t = 1$), we can express the future value of the series as:

$$FV = \sum_{t=1}^N CF_t(1+i)^{N-t}$$

N is last and t indicates the time period corresponding to a particular cash flow, starting at 1 for an ordinary annuity. Since CF_t is shorthand for: $CF_1, CF_2, CF_3, \dots, CF_N$, and we know that $CF_1 = CF_2 = CF_3 = \dots = CF_N$, let's make things simple by using CF to indicate the same value for the periodic cash flows. Rearranging the future value equation we get:

$$FV = CF \sum_{t=1}^N (1+i)^{N-t} \tag{8}$$

This equation tells us that the future value of a level series of cash flows, occurring at regular intervals beginning one period from today (notice that t starts at 1), is equal to the amount of cash flow multiplied by the sum of the compound factors.

In a like manner, the equation for the present value of a series of level cash flows beginning

after one period simplifies to:

$$PV = \sum_{t=1}^N CF_t \left(\frac{1}{1+i} \right)^t = CF \sum_{t=1}^N \left(\frac{1}{1+i} \right)^t$$

or

$$PV = CF \sum_{t=1}^N \frac{1}{(1+i)^t} \tag{9}$$

This equation tells us that the present value of an annuity is equal to the amount of one cash flow multiplied by the sum of the discount factors.

Equations (8) and (9) are the valuation—future and present value—formulas for an *ordinary annuity*. An ordinary annuity is a special form of annuity, where the first cash flow occurs at the end of the first period.

To calculate the future value of an annuity we multiply the amount of the annuity (that is, the amount of one periodic cash flow) by the sum of the compound factors. The sum of these compounding factors for a given interest rate, i , and number of periods, N , is referred to as the *future value annuity factor*. Likewise, to calculate the present value of an annuity we multiply one cash flow of the annuity by the sum of the discount factors. The sum of the discounting factors for a given i and N is referred to as the *present value annuity factor*.

Suppose you wish to determine the future value of a series of deposits of \$1,000, deposited each year in the No Fault Vault Bank for five years, with the first deposit made at the end of the first year. If the NFV Bank pays 5% interest on the balance in the account at the end of each year and no withdrawals are made, what is the balance in the account at the end of the five years?

Each \$1,000 is deposited at a different time, so it contributes a different amount to the future value. For example, the first deposit accumulates interest for four periods, contributing \$1,215.50 to the future value (at the end of period 5), whereas the last deposit contributes only \$1,000 to the future value since it is deposited at exactly the point in time when we

are determining the future value, hence there is no interest on this deposit.

The future value of an annuity is the sum of the future value of each deposit:

Period	Amount of Deposit	Number of Periods Interest Is Earned	Compounding Factor	Future Value
1	\$1,000	4	1.2155	\$1,215.50
2	1,000	3	1.1576	1,157.60
3	1,000	2	1.1025	1,102.50
4	1,000	1	1.0500	1,050.00
5	1,000	0	1.0000	1,000.00
Total			5.5256	\$5,525.60

The future value of the series of \$1,000 deposits, with interest compounded at 5%, is \$5,525.60. Since we know the value of one of the level period flows is \$1,000, and the future value of the annuity is \$5,525.60, and looking at the sum of the individual compounding factors, 5.5256, we can see that there is an easier way to calculate the future value of an annuity. If the sum of the individual compounding factors for a specific interest rate and a specific number of periods were available, all we would have to do is multiply that sum by the value of one cash flow to get the future value of the entire annuity.

In this example, the shortcut is multiplying the amount of the annuity, \$1,000, by the sum of the compounding factors, 5.5256:

$$FV = \$1,000 \times 5.5256 = \$5,525.60$$

For large numbers of periods, summing the individual factors can be a bit clumsy—with possibilities of errors along the way. An alternative formula for the sum of the compound factors—that is, the future value annuity factor—is:

$$\text{Future value annuity factor} = \frac{(1+i)^N - 1}{i} \quad (10)$$

In the last example, $N = 5$ and $i = 5\%$:

$$\begin{aligned} \text{Future value annuity factor} &= \frac{(1+0.05)^5 - 1}{0.05} \\ &= \frac{1.2763 - 1.000}{0.05} \\ &= 5.5256 \end{aligned}$$

Let's use the long method to find the present value of the series of five deposits of \$1,000 each, with the first deposit at the end of the first period. Then we'll do it using the shortcut method. The calculations are similar to the future value of an ordinary annuity, except we are taking each deposit back in time, instead of forward:

Period	Amount of Deposit	Discounting Periods	Discounting Factor	Present Value
1	\$1,000	1	0.9524	\$952.40
2	1,000	2	0.9070	907.00
3	1,000	3	0.8638	863.80
4	1,000	4	0.8227	822.70
5	1,000	5	0.7835	783.50
Total			4.3294	\$4,329.40

The present value of this series of five deposits is \$4,329.40.

This same value is obtained by multiplying the annuity amount of \$1,000 by the sum of the discounting factors, 4.3294:

$$PV = \$1,000 \times 4.3294 = \$4,329.40$$

Another, more convenient way of solving for the present value of an annuity is to rewrite the factor as:

$$\text{Present value annuity factor} = \frac{1 - \frac{1}{(1+i)^N}}{i} \quad (11)$$

If there are many discount periods, this formula is a bit easier to calculate. In our last example,

$$\begin{aligned} \text{Present value annuity factor} &= \frac{\left[1 - \frac{1}{(1+0.05)^5}\right]}{0.05} \\ &= \frac{1 - 0.7835}{0.05} \\ &= 4.3295 \end{aligned}$$

which is different from the sum of the factors, 4.3294, due to rounding.

We can turn this present value of an annuity problem around to look at it from another angle. Suppose you borrow \$4,329.40 at an interest rate of 5% per period and are required to pay back this loan in five installments ($N = 5$): one payment per period for five periods, starting

one period from now. The payments are determined by equating the present value with the product of the cash flow and the sum of the discount factors:

$$\begin{aligned}
 PV &= CF(\text{sum of discount factors}) \\
 &= CF \sum_{t=1}^5 \frac{1}{(1 + 0.05)^t} \\
 &= CF (0.9524 + 0.9070 + 0.8638 + 0.8227 \\
 &\quad + 0.7835) \\
 &= CF (4.3294)
 \end{aligned}$$

substituting the known present value,

$$\$4,329.40 = CF (4.3294)$$

and rearranging to solve for the payment:

$$CF = \$4,329.40/4.3290 = \$1,000.00$$

We can convince ourselves that five installments of \$1,000 each can pay off the loan of \$4,329.40 by carefully stepping through the calculation of interest and the reduction of the principal:

Beginning of Periods Loan Balance	Payment	Interest (Principal × 5%)	Reduction in Loan Balance (Payment – Interest)	End-of-Period Loan Balance
\$4,329.40	\$1,000.00	\$216.47	\$783.53	\$3,545.87
3,545.87	1,000.00	177.29	822.71	2,723.16
2,723.16	1,000.00	136.16	863.84	1,859.32
1,859.32	1,000.00	92.97	907.03	952.29
952.29	1,000.00	47.61	952.29 ^a	0

^aThe small difference between calculated reduction (\$952.38) and reported reduction is due to rounding differences.

For example, the first payment of \$1,000 is used to: (1) pay interest on the loan at 5% (\$4,329.40 × 0.05 = \$216.47) and (2) pay down the principal or loan balance (\$1,000.00 – \$216.47 = \$783.53 paid off). Each successive payment pays off a greater amount of the loan—as the principal amount of the loan is reduced, less of each payment goes to paying off interest and more goes to reducing the loan principal. This analysis of the repayment of a loan is referred to as loan amortization. *Loan amortization*

is the repayment of a loan with equal payments, over a specified period of time. As we can see from the example of borrowing \$4,329.40, each payment can be broken down into its interest and principal components.

VALUING CASH FLOWS WITH DIFFERENT TIME PATTERNS

Valuing a Perpetual Stream of Cash Flows

There are some circumstances where cash flows are expected to continue forever. For example, a corporation may promise to pay dividends on preferred stock forever, or, a company may issue a bond that pays interest every six months, forever. How do you value these cash flow streams? Recall that when we calculated the present value of an annuity, we took the amount of one cash flow and multiplied it by the sum of the discount factors that corresponded to the interest rate and number of payments. But what if the number of payments extends forever—into infinity?

A series of cash flows that occur at regular intervals, forever, is a *perpetuity*. Valuing a perpetual cash flow stream is just like valuing an ordinary annuity. It looks like this:

$$\begin{aligned}
 PV &= CF_1 \left(\frac{1}{1+i} \right)^1 + CF_2 \left(\frac{1}{1+i} \right)^2 \\
 &\quad + CF_3 \left(\frac{1}{1+i} \right)^3 + \dots + CF_\infty \left(\frac{1}{1+i} \right)^\infty
 \end{aligned}$$

Simplifying, recognizing that the cash flows CF_t are the same in each period, and using summation notation,

$$PV = CF \sum_{t=1}^{\infty} \left(\frac{1}{1+i} \right)^t$$

As the number of discounting periods approaches infinity, the summation approaches $1/i$. To see why, consider the present value

annuity factor for an interest rate of 10%, as the number of payments goes from 1 to 200:

Number of Discounting Periods, N	Present Value Annuity Factor
1	0.9091
10	6.1446
40	9.7791
100	9.9993
200	9.9999

For greater numbers of payments, the factor approaches 10, or $1/0.10$. Therefore, the present value of a perpetual annuity is very close to:

$$PV = \frac{CF}{i} \quad (12)$$

Suppose you are considering an investment that promises to pay \$100 each period forever, and the interest rate you can earn on alternative investments of similar risk is 5% per period. What are you willing to pay today for this investment?

$$PV = \frac{\$100}{0.05} = \$2,000$$

Therefore, you would be willing to pay \$2,000 today for this investment to receive, in return, the promise of \$100 each period forever.

Let's look at the value of a perpetuity another way. Suppose that you are given the opportunity to purchase an investment for \$5,000 that promises to pay \$50 at the end of every period forever. What is the periodic interest per period—the return—associated with this investment?

We know that the present value is $PV = \$5,000$ and the periodic, perpetual payment is $CF = \$50$. Inserting these values into the formula for the present value of a perpetuity:

$$\$5,000 = \frac{\$50}{i}$$

Solving for i ,

$$i = \frac{\$50}{\$5,000} = 0.01 \text{ or } 1\% \text{ per period}$$

Therefore, an investment of \$5,000 that generates \$50 per period provides 1% compounded interest per period.

Valuing an Annuity Due

The ordinary annuity cash flow analysis assumes that cash flows occur at the end of each period. However, there is another fairly common cash flow pattern in which level cash flows occur at regular intervals, but the first cash flow occurs immediately. This pattern of cash flows is called an *annuity due*. For example, if you win the Florida Lottery Lotto grand prize, you will receive your winnings in 20 installments (after taxes, of course). The 20 installments are paid out annually, beginning immediately. The lottery winnings are therefore an annuity due.

Like the cash flows we have considered thus far, the future value of an annuity due can be determined by calculating the future value of each cash flow and summing them. And, the present value of an annuity due is determined in the same way as a present value of any stream of cash flows.

Let's consider first an example of the future value of an annuity due, comparing the values of an ordinary annuity and an annuity due, each comprising three cash flows of \$500, compounded at the interest rate of 4% per period. The calculation of the future value of both the ordinary annuity and the annuity due at the end of three periods is:

$$\begin{array}{ll} \text{Ordinary annuity} & \text{Annuity due} \\ FV = \$500 \sum_{t=1}^3 (1 + 0.04)^{3-t} & FV_{\text{due}} = \$500 \sum_{t=1}^3 (1 + 0.04)^{3-t+1} \end{array}$$

The future value of each of the \$500 payments in the annuity due calculation is compounded for one more period than for the ordinary annuity. For example, the first deposit of \$500 earns interest for two periods in the ordinary annuity situation [$\$500 (1 + 0.04)^2$], whereas the first \$500 in the annuity due case earns interest for three periods [$\$500 (1 + 0.04)^3$].

In general terms,

$$FV_{\text{due}} = CF \sum_{t=1}^N (1+i)^{N-t+1} \quad (13)$$

which is equal to the future value of an ordinary annuity multiplied by a factor of $1+i$:

$$FV_{\text{due}} = CF[\text{Future value annuity factor (ordinary) for } N \text{ and } i](1+i)$$

The present value of the annuity due is calculated in a similar manner, adjusting the ordinary annuity formula for the different number of discount periods:

$$PV_{\text{due}} = CF \sum_{t=1}^N \frac{1}{(1+i)^{t-1}} \quad (14)$$

Since the cash flows in the annuity due situation are each discounted one less period than the corresponding cash flows in the ordinary annuity, the present value of the annuity due is greater than the present value of the ordinary annuity for an equivalent amount and number of cash flows. Like the future value an annuity due, we can specify the present value in terms of the ordinary annuity factor:

$$PV_{\text{due}} = CF[\text{Present value annuity factor (ordinary) for } N \text{ and } i](1+i)$$

Valuing a Deterred Annuity

A *deferred annuity* has a stream of cash flows of equal amounts at regular periods starting at some time after the end of the first period. When we calculated the present value of an annuity, we brought a series of cash flows back to the beginning of the first period—or, equivalently the end of the period 0. With a deferred annuity, we determine the present value of the ordinary annuity and then discount this present value to an earlier period.

To illustrate the calculation of the present value of an annuity due, suppose you deposit \$20,000 per year in an account for 10 years, starting today, for a total of 10 deposits. What will

be the balance in the account at the end of 10 years if the balance in the account earns 5% per year? The future value of this annuity due is:

$$\begin{aligned} FV_{\text{due},10} &= \$20,000 \sum_{t=1}^{10} (1+0.05)^{10-t+1} \\ &= \$20,000 \left(\begin{array}{l} \text{Future value annuity} \\ \text{factor (ordinary) for} \\ \text{10 periods and 5\%} \end{array} \right) \\ &\quad \times (1+0.05) \\ &= \$20,000(12.5779)(1+0.05) = \$264,135.74 \end{aligned}$$

Suppose you want to deposit an amount today in an account such that you can withdraw \$5,000 per year for four years, with the first withdrawal occurring five years from today. We can solve this problem in two steps:

Step 1: Solve for the present value of the withdrawals.

Step 2: Discount this present value to the present.

The first step requires determining the present value of a four-cash-flow ordinary annuity of \$5,000. This calculation provides the present value as of the end of the fourth year (one period prior to the first withdrawal):

$$\begin{aligned} PV_4 &= \$5,000 \sum_{t=1}^4 \frac{1}{(1+0.04)^t} \\ &= \$5,000 (\text{present value annuity factor} \\ &\quad N = 4, i = 4\%) \\ &= \$18,149.48 \end{aligned}$$

This means that there must be a balance in the account of \$18,149.48 at the end of the fourth period to satisfy the withdrawals of \$5,000 per year for four years.

The second step requires discounting the \$18,149.48—the savings goal—to the present, providing the deposit today that produces the goal:

$$PV_0 = \frac{\$18,149.48}{(1+0.04)^4} = \$15,514.25$$

The balance in the account throughout the entire eight-year period is shown in Figure 6 with

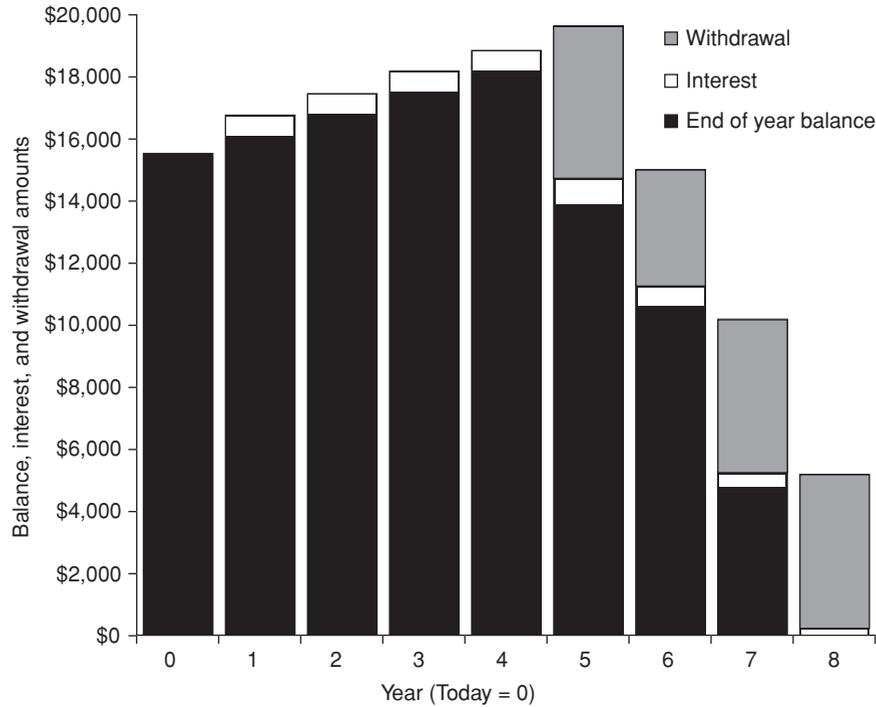


Figure 6 Balance in the Account that Requires a Deposit Today (Year 0) that Permits Withdrawals of \$5,000 Each Starting at the End of Year 4

the balance indicated both before and after the \$5,000 withdrawals.

Let's look at a more complex deferred annuity. Consider making a series of deposits, beginning today, to provide for a steady cash flow beginning at some future time period. If interest is earned at a rate of 4% compounded per year, what amount must be deposited in a savings account each year for four years, starting today, so that \$1,000 may be withdrawn each year for five years, beginning five years from today? As with any deferred annuity, we need to perform this calculation in steps:

Step 1: Calculate the present value of the \$1,000 per year five-year ordinary annuity as of the end of the fourth year:

The present value of the annuity deferred to the end of the fourth period is

$$\begin{aligned}
 PV_4 &= \$1,000 \sum_{t=1}^5 \frac{1}{(1 + 0.04)^t} = \$1,000(4.4518) \\
 &= \$4,451.80
 \end{aligned}$$

Therefore, there must be \$4,451.80 in the account at the end of the fourth year to permit five \$1,000 withdrawals at the end of each of the years 5, 6, 7, 8, and 9.

Step 2: Calculate the cash flow needed to arrive at the future value of that annuity due comprising four annual deposits earning 4% compounded interest, starting today.

The present value of the annuity at the end of the fourth year, \$4,451.80, is the future value of the annuity due of four payments of an unknown amount. Using the formula for the future value of an annuity due,

$$\begin{aligned}
 \$4,451.80 &= CF \sum_{t=1}^4 (1 + 0.04)^{4-t+1} \\
 &= CF (4.2465)(1.04)
 \end{aligned}$$

and rearranging,

$$CF = \$4,451.80 / 4.4164 = \$1,008.02$$

Therefore, by depositing \$1,008.02 today and the same amount on the same date each of the

next three years, we will have a balance in the account of \$4,451.80 at the end of the fourth period. With this period 4 balance, we will be able to withdraw \$1,000 at the end of the following five periods.

LOAN AMORTIZATION

There are securities backed by various types of loans. These include asset-backed securities, residential mortgage-backed securities, and commercial mortgage-backed securities. Consequently, it is important to understand the mathematics associated with loan amortization.

If an amount is loaned and then repaid in installments, we say that the loan is amortized. Therefore, *loan amortization* is the process of calculating the loan payments that amortize the loaned amount. We can determine the amount of the loan payments once we know the frequency of payments, the interest rate, and the number of payments.

Consider a loan of \$100,000. If the loan is repaid in 24 annual installments (at the end of each year) and the interest rate is 5% per year, we calculate the amount of the payments by applying the relationship:

$$\begin{aligned}
 PV &= \sum_{t=1}^N \frac{CF}{(1+i)^t} \\
 \text{Amount loaned} &= \sum_{t=1}^N \frac{\text{Loan payment}}{(1+i)^t} \\
 \$100,000 &= \sum_{t=1}^{24} \frac{\text{Loan payment}}{(1+0.05)^t}
 \end{aligned}$$

We want to solve for the loan payment, that is, the amount of the annuity. Using a financial calculator or spreadsheet, the periodic loan payment is \$7,247.09 ($PV = \$100,000$; $N = 24$; $i = 5\%$). Therefore, the monthly payments are \$7,247.09 each. In other words, if payments of \$7,247.09 are made each year for 24 years (at the end of each year), the \$100,000 loan will be repaid and the lender earns a return that is equivalent to a 5% interest on this loan.

We can calculate the amount of interest and principal repayment associated with each loan

payment using a loan amortization schedule, as shown in Table 1.

The loan payments are determined such that after the last payment is made there is no loan balance outstanding. Thus, the loan is referred to as a *fully amortizing loan*. Even though the loan payment each year is the same, the proportion of interest and principal differs with each payment: The interest is 5% of the principal amount of the loan that remains at the beginning of the period, whereas the principal repaid with each payment is the difference between the payment and the interest. As the payments are made, the remainder is applied to repayment of the principal; this is referred to as the scheduled principal repayment or the *amortization*. As the principal remaining on the loan declines, less interest is paid with each payment. We show the decline in the loan's principal graphically in Figure 7. The decline in the remaining principal is not a linear, but is curvilinear due to the compounding of interest.

Loan amortization works the same whether this is a mortgage loan, a term loan, or any other loan in which the interest paid is determined on the basis of the remaining amount of the loan. The calculation of the loan amortization can be modified to suit different principal repayments, such as additional lump-sum payments, known as *balloon payments*. For example, if there is a \$10,000 balloon payment at the end of the loan in the loan of \$100,000 repaid over 24 years, the calculation of the payment is modified as:

$$\begin{aligned}
 \text{Amount loaned} &= \left[\sum_{t=1}^N \frac{\text{Loan payment}}{(1+i)^t} \right] + \frac{\text{balloon payment}}{(1+i)^N} \\
 \$100,000 &= \left[\sum_{t=1}^{24} \frac{\text{Loan payment}}{(1+0.05)^t} \right] + \frac{\$10,000}{(1+i)^{24}}
 \end{aligned}$$

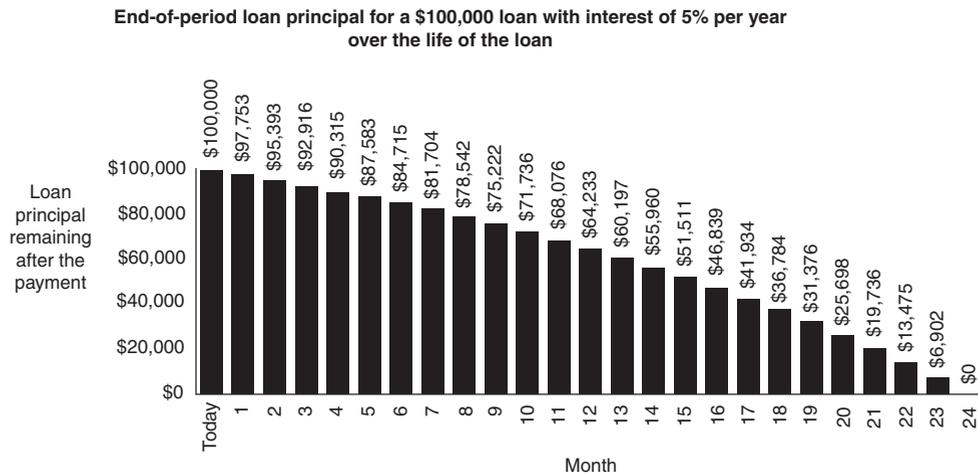
The loan payment that solves this equation is \$7,022.38 ($PV = \$100,000$; $N = 24$; $i = 5\%$; $FV = \$10,000$). The last payment (that is, at the

Table 1 Loan Amortization on a \$100,000 Loan for 24 Years and an Interest Rate of 5% per Year

Payment	Loan Payment	Beginning-of-the-Year Principal	Interest on Loan	Principal Paid Off = Payment – Interest	Remaining Principal
0					\$100,000.00
1	\$7,247.09	\$100,000.00	\$5,000.00	\$2,247.09	\$97,752.91
2	\$7,247.09	\$97,752.91	\$4,887.65	\$2,359.44	\$95,393.47
3	\$7,247.09	\$95,393.47	\$4,769.67	\$2,477.42	\$92,916.05
4	\$7,247.09	\$92,916.05	\$4,645.80	\$2,601.29	\$90,314.76
5	\$7,247.09	\$90,314.76	\$4,515.74	\$2,731.35	\$87,583.41
6	\$7,247.09	\$87,583.41	\$4,379.17	\$2,867.92	\$84,715.49
7	\$7,247.09	\$84,715.49	\$4,235.77	\$3,011.32	\$81,704.17
8	\$7,247.09	\$81,704.17	\$4,085.21	\$3,161.88	\$78,542.29
9	\$7,247.09	\$78,542.29	\$3,927.11	\$3,319.98	\$75,222.32
10	\$7,247.09	\$75,222.32	\$3,761.12	\$3,485.97	\$71,736.34
11	\$7,247.09	\$71,736.34	\$3,586.82	\$3,660.27	\$68,076.07
12	\$7,247.09	\$68,076.07	\$3,403.80	\$3,843.29	\$64,232.78
13	\$7,247.09	\$64,232.78	\$3,211.64	\$4,035.45	\$60,197.33
14	\$7,247.09	\$60,197.33	\$3,009.87	\$4,237.22	\$55,960.11
15	\$7,247.09	\$55,960.11	\$2,798.01	\$4,449.08	\$51,511.03
16	\$7,247.09	\$51,511.03	\$2,575.55	\$4,671.54	\$46,839.49
17	\$7,247.09	\$46,839.49	\$2,341.97	\$4,905.12	\$41,934.37
18	\$7,247.09	\$41,934.37	\$2,096.72	\$5,150.37	\$36,784.00
19	\$7,247.09	\$36,784.00	\$1,839.20	\$5,407.89	\$31,376.11
20	\$7,247.09	\$31,376.11	\$1,568.81	\$5,678.28	\$25,697.83
21	\$7,247.09	\$25,697.83	\$1,284.89	\$5,962.20	\$19,735.63
22	\$7,247.09	\$19,735.63	\$986.78	\$6,260.31	\$13,475.32
23	\$7,247.09	\$13,475.32	\$673.77	\$6,573.32	\$6,901.99
24	\$7,247.09	\$6,901.99	\$345.10	\$6,901.99	\$0.00

end of the 24th year) is the regular payment of \$7,022.38, plus the balloon payment, for a total of \$17,022.38. As you can see in Figure 8, the loan amortization is slower when compared to the loan without the balloon payment.

The same mathematics work with term loans. Term loans are usually repaid in installments either monthly, quarterly, semiannually, or annually. Let's look at the typical repayment schedule for a term loan. Suppose that BigRock

**Figure 7** Loan Amortization

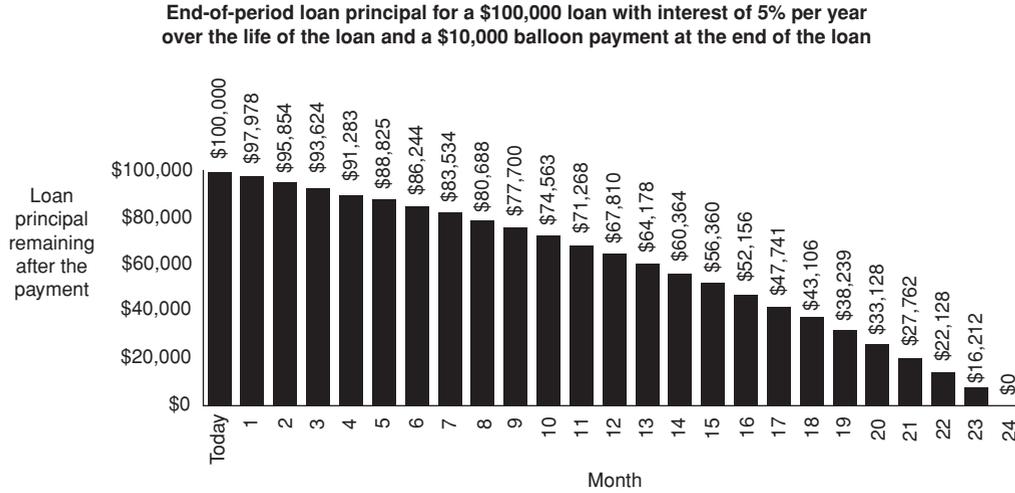


Figure 8 Loan Amortization with Balloon Payment

Corporation seeks a four-year term loan of \$100 million. Let’s assume for now that the term loan carries a fixed interest rate of 8% and that level payments are made monthly. If the annual interest rate is 8%, the rate per month is $8\% \div 12 = 0.6667\%$ per month. In a typical term loan, the payments are structured such that the loan is fully amortizing.

For this four-year, \$100 million term loan with an 8% interest rate, the monthly payment is \$2,441,292.23 ($PV = \$100,000,000; N = 48; i = 0.6667\%$). This amount is determined by solving for the annuity payment that equates the present value of the payments with the amount of the loan, considering a discount rate of 0.6667%. In Table 2 we show for each month the beginning monthly balance, the interest payment for the month, the amount of the monthly, and the ending loan balance. Notice that in our illustration, the ending loan balance is zero. That is, it is a fully amortizing loan.

In the loan amortization examples so far, we have assumed that the interest rate is fixed throughout the loan. However, in many loans the interest rate may change during the loan, as in the case of a floating-rate loan. The new loan rate at the reset date is determined by a formula. The formula is typically composed of two parts. The first is the reference rate. For

example, in a monthly pay loan, the loan rate might be one-month London Interbank Offered Rate (LIBOR). The second part is a spread that is added to the reference rate. This spread is referred to as the quoted margin and depends on the credit of the borrower.

A floating-rate loan requires a recalculation of the loan payment and payment schedule at each time the loan rate is reset. Suppose in the case of BigRock’s term loan that the rate remains constant for the first three years, but is reset to 9% in the fourth year. This requires BigRock to pay off the principal remaining at the end of three years, the \$28,064,562.84, in the remaining 12 payments. The revised schedule of payments and payoff for the fourth year require a payment of \$2,454,287.47 ($PV = \$27,064,562.84; N = 12; i = 0.09 \div 12 = 0.75\%$), as shown in Table 3.

THE CALCULATION OF INTEREST RATES AND YIELDS

The calculation of the present or future value of a lump-sum or set of cash flows requires information on the timing of cash flows and the compound or discount rate. However, there are

Table 2 Term Loan Schedule: Fixed Rate, Fully Amortized

Monthly Payment	Beginning-of-the-Year Principal	Interest on Loan	Principal Paid Off = Payment – Interest	Remaining Principal
	\$100,000,000			
	Interest rate	8% per year		
	Number of years	4		
	Monthly payment	\$2,441,292.33		
1	\$100,000,000.00	\$666,666.67	\$1,774,625.57	\$98,225,374.43
2	\$98,225,374.43	\$654,835.83	\$1,786,456.40	\$96,438,918.03
3	\$96,438,918.03	\$642,926.12	\$1,798,366.11	\$94,640,551.91
4	\$94,640,551.91	\$630,937.01	\$1,810,355.22	\$92,830,196.69
5	\$92,830,196.69	\$618,867.98	\$1,822,424.26	\$91,007,772.44
6	\$91,007,772.44	\$606,718.48	\$1,834,573.75	\$89,173,198.69
7	\$89,173,198.69	\$594,487.99	\$1,846,804.24	\$87,326,394.44
8	\$87,326,394.44	\$582,175.96	\$1,859,116.27	\$85,467,278.17
9	\$85,467,278.17	\$569,781.85	\$1,871,510.38	\$83,595,767.79
10	\$83,595,767.79	\$557,305.12	\$1,883,987.12	\$81,711,780.68
11	\$81,711,780.68	\$544,745.20	\$1,896,547.03	\$79,815,233.65
12	\$79,815,233.65	\$532,101.56	\$1,909,190.68	\$77,906,042.97
13	\$77,906,042.97	\$519,373.62	\$1,921,918.61	\$75,984,124.36
14	\$75,984,124.36	\$506,560.83	\$1,934,731.41	\$74,049,392.95
15	\$74,049,392.95	\$493,662.62	\$1,947,629.61	\$72,101,763.34
16	\$72,101,763.34	\$480,678.42	\$1,960,613.81	\$70,141,149.52
17	\$70,141,149.52	\$467,607.66	\$1,973,684.57	\$68,167,464.95
18	\$68,167,464.95	\$454,449.77	\$1,986,842.47	\$66,180,622.49
19	\$66,180,622.49	\$441,204.15	\$2,000,088.08	\$64,180,534.40
20	\$64,180,534.40	\$427,870.23	\$2,013,422.00	\$62,167,112.40
21	\$62,167,112.40	\$414,447.42	\$2,026,844.82	\$60,140,267.58
22	\$60,140,267.58	\$400,935.12	\$2,040,357.12	\$58,099,910.46
23	\$58,099,910.46	\$387,332.74	\$2,053,959.50	\$56,045,950.96
24	\$56,045,950.96	\$373,639.67	\$2,067,652.56	\$53,978,298.40
25	\$53,978,298.40	\$359,855.32	\$2,081,436.91	\$51,896,861.49
26	\$51,896,861.49	\$345,979.08	\$2,095,313.16	\$49,801,548.33
27	\$49,801,548.33	\$332,010.32	\$2,109,281.91	\$47,692,266.42
28	\$47,692,266.42	\$317,948.44	\$2,123,343.79	\$45,568,922.63
29	\$45,568,922.63	\$303,792.82	\$2,137,499.42	\$43,431,423.21
30	\$43,431,423.21	\$289,542.82	\$2,151,749.41	\$41,279,673.80
31	\$41,279,673.80	\$275,197.83	\$2,166,094.41	\$39,113,579.39
32	\$39,113,579.39	\$260,757.20	\$2,180,535.04	\$36,933,044.35
33	\$36,933,044.35	\$246,220.30	\$2,195,071.94	\$34,737,972.42
34	\$34,737,972.42	\$231,586.48	\$2,209,705.75	\$32,528,266.66
35	\$32,528,266.66	\$216,855.11	\$2,224,437.12	\$30,303,829.54
36	\$30,303,829.54	\$202,025.53	\$2,239,266.70	\$28,064,562.84
37	\$28,064,562.84	\$187,097.09	\$2,254,195.15	\$25,810,367.69
38	\$25,810,367.69	\$172,069.12	\$2,269,223.12	\$23,541,144.57
39	\$23,541,144.57	\$156,940.96	\$2,284,351.27	\$21,256,793.30
40	\$21,256,793.30	\$141,711.96	\$2,299,580.28	\$18,957,213.02
41	\$18,957,213.02	\$126,381.42	\$2,314,910.81	\$16,642,302.21
42	\$16,642,302.21	\$110,948.68	\$2,330,343.55	\$14,311,958.66
43	\$14,311,958.66	\$95,413.06	\$2,345,879.18	\$11,966,079.48
44	\$11,966,079.48	\$79,773.86	\$2,361,518.37	\$9,604,561.11
45	\$9,604,561.11	\$64,030.41	\$2,377,261.83	\$7,227,299.28
46	\$7,227,299.28	\$48,182.00	\$2,393,110.24	\$4,834,189.04
47	\$4,834,189.04	\$32,227.93	\$2,409,064.31	\$2,425,124.74
48	\$2,425,124.74	\$16,167.50	\$2,425,124.74	\$0.00

Table 3 Term Loan Schedule: Reset Rate, Fully Amortized

Amount of loan	\$100,000,000			
Interest rate	8% per year for the first 3 years, 9% thereafter			
Number of years	4			
Monthly payment	\$2,441,292.33 for the first 3 years, \$2,454,287.47 for the fourth year and beyond			
Monthly Payment	Beginning-of-the-Year Principal	Interest on Loan	Principal Paid Off = Payment – Interest	Remaining Principal
37	\$28,064,562.84	\$210,484.22	\$2,243,803.24	\$25,820,759.59
38	\$25,820,759.59	\$193,655.70	\$2,260,631.77	\$23,560,127.82
39	\$23,560,127.82	\$176,700.96	\$2,277,586.51	\$21,282,541.32
40	\$21,282,541.32	\$159,619.06	\$2,294,668.41	\$18,987,872.91
41	\$18,987,872.91	\$142,409.05	\$2,311,878.42	\$16,675,994.49
42	\$16,675,994.49	\$125,069.96	\$2,329,217.51	\$14,346,776.99
43	\$14,346,776.99	\$107,600.83	\$2,346,686.64	\$12,000,090.35
44	\$12,000,090.35	\$90,000.68	\$2,364,286.79	\$9,635,803.56
45	\$9,635,803.56	\$72,268.53	\$2,382,018.94	\$7,253,784.62
46	\$7,253,784.62	\$54,403.38	\$2,399,884.08	\$4,853,900.54
47	\$4,853,900.54	\$36,404.25	\$2,417,883.21	\$2,436,017.33
48	\$2,436,017.33	\$18,270.13	\$2,436,017.34	\$0.00

many applications in which we are presented with values and cash flows, and wish to calculate the yield or implied interest rate associated with these values and cash flows. By calculating the yield or implied interest rate, we can then compare investment or financing opportunities. We first look at how interest rates are stated and how the effective interest rate can be calculated based on this stated rate, and then we look at how to calculate the yield, or rate of return, on a set of cash flows.

Annual Percentage Rate versus Effective Annual Rate

A common problem in finance is comparing alternative financing or investment opportunities when the interest rates are specified in a way that makes it difficult to compare terms. The Truth in Savings Act requires institutions to provide the annual percentage yield for savings accounts. As a result of this law, consumers can compare the yields on different savings arrangements. But this law does not apply beyond savings accounts. One investment may pay 10% interest compounded semiannually, whereas another investment may pay 9% interest compounded daily. One financing ar-

agement may require interest compounding quarterly, whereas another may require interest compounding monthly. To compare investments or financing with different frequencies of compounding, we must first translate the stated interest rates into a common basis. There are two ways to convert interest rates stated over different time intervals so that they have a common basis: the annual percentage rate and the effective annual interest rate.

One obvious way to represent rates stated in various time intervals on a common basis is to express them in the same unit of time—so we annualize them. The annualized rate is the product of the stated rate of interest per compound period and the number of compounding periods in a year. Let i be the rate of interest per period and n be the number of compounding periods in a year. The annualized rate, also referred to as the *nominal interest rate* or the annual percentage rate (APR) is:

$$\text{APR} = i \times n$$

Consider the following example. Suppose the Lucky Break Loan Company has simple loan terms: Repay the amount borrowed, plus 50%, in six months. Suppose you borrow \$10,000 from Lucky. After six months, you must pay

back the \$10,000 plus \$5,000. The APR on financing with Lucky is the interest rate per period (50% for six months) multiplied by the number of compound periods in a year (two six-month periods in a year). For the Lucky Break financing arrangement:

$$\text{APR} = 0.50 \times 2 = 1.00 \text{ or } 100\% \text{ per year}$$

But what if you cannot pay Lucky back after six months? Lucky will let you off this time, but you must pay back the following at the end of the next six months:

- The \$10,000 borrowed.
- The \$5,000 interest from the first six months.
- The 50% of interest on both the unpaid \$10,000 and the unpaid \$5,000 interest (\$15,000 (0.50) = \$7,500).

So, at the end of the year, knowing what is good for you, you pay off Lucky:

Amount of the original loan	\$10,000
Interest from first six months	5,000
Interest on second six months	7,500
	\$22,500
Total payment at end of the year	\$22,500

Using the Lucky Break method of financing, you have to pay \$12,500 interest to borrow \$10,000 for one year's time. Because you have to pay \$12,500 interest to borrow \$10,000 over one year's time, you pay not 100% interest, but rather 125% interest per year (\$12,500/\$10,000 = 1.25 = 125%). What's going on here? It looks like the APR in the Lucky Break example ignores the compounding (interest on interest) that takes place after the first six months. And that's the way it is with all APRs. The APR ignores the effect of compounding. Therefore, this rate understates the true annual rate of interest if interest is compounded at any time prior to the end of the year. Nevertheless, APR is an acceptable method of disclosing interest on many lending arrangements, since it is easy to understand and simple to compute. However, because it ignores compounding, it is not the best way to convert interest rates to a common basis.

Another way of converting stated interest rates to a common basis is the effective rate of interest. The *effective annual rate* (EAR) is the true economic return for a given time period—it takes into account the compounding of interest—and is also referred to as the effective rate of interest.

Using our Lucky Break example, we see that we must pay \$12,500 interest on the loan of \$10,000 for one year. Effectively, we are paying 125% annual interest. Thus, 125% is the effective annual rate of interest. In this example, we can easily work through the calculation of interest and interest on interest. But for situations where interest is compounded more frequently, we need a direct way to calculate the effective annual rate. We can calculate it by resorting once again to our basic valuation equation:

$$FV = PV(1 + i)^n$$

Next, we consider that a return is the change in the value of an investment over a period and an annual return is the change in value over a year. Using our basic valuation equation, the relative change in value is the difference between the future value and the present value, divided by the present value:

$$\text{EAR} = \frac{FV - PV}{PV} = \frac{PV(1 + i)^n}{PV}$$

Canceling PV from both the numerator and the denominator,

$$\text{EAR} = (1 + i)^n - 1 \quad (15)$$

Let's look how the EAR is affected by the compounding. Suppose that the Safe Savings and Loan promises to pay 6% interest on accounts, compounded annually. Since interest is paid once, at the end of the year, the effective annual return, EAR, is 6%. If the 6% interest is paid on a semiannual basis—3% every six months—the effective annual return is larger than 6% since interest is earned on the 3% interest earned at the end of the first six months. In this case, to calculate the EAR, the interest rate per compounding period—six months—is 0.03

(that is, $0.06/2$) and the number of compounding periods in an annual period is 2:

$$\text{EAR} = (1 + 0.03)^2 - 1 = 1.0609 - 1 = 0.0609$$

or 6.09%

Extending this example to the case of quarterly compounding with a nominal interest rate of 6%, we first calculate the interest rate per period, i , and the number of compounding periods in a year, n :

$$i = 0.06/4 = 0.015 \text{ per quarter}$$

$$n = 4 \text{ quarters in a year}$$

The EAR is:

$$\text{EAR} = (1 + 0.015)^4 - 1 = 1.0614 - 1 = 0.0614$$

or 6.14%

As we saw earlier, the extreme frequency of compounding is continuous compounding. Continuous compounding is when interest is compounded at the smallest possible increment of time. In continuous compounding, the rate per period becomes extremely small:

$$i = \frac{\text{APR}}{\infty}$$

And the number of compounding periods in a year, n , is infinite. The EAR is therefore:

$$\text{EAR} = e^{\text{APR}} - 1 \tag{16}$$

where e is the natural logarithmic base.

For the stated 6% annual interest rate compounded continuously, the EAR is:

$$\text{EAR} = e^{0.06} - 1 = 1.0618 - 1 = 0.0618 \text{ or } 6.18\%$$

The relation between the frequency of compounding for a given stated rate and the effective annual rate of interest for this example indicates that the greater the frequency of compounding, the greater the EAR.

Frequency of Compounding	Calculation	Effective Annual Rate
Annual	$(1 + 0.060)^1 - 1$	6.00%
Semiannual	$(1 + 0.030)^2 - 1$	6.09%
Quarterly	$(1 + 0.015)^4 - 1$	6.14%
Continuous	$e^{0.06} - 1$	6.18%

Figuring out the effective annual rate is useful when comparing interest rates for different investments. It doesn't make sense to compare the APRs for different investments having a different frequency of compounding within a year. But since many investments have returns stated in terms of APRs, we need to understand how to work with them.

To illustrate how to calculate effective annual rates, consider the rates offered by two banks, Bank A and Bank B. Bank A offers 9.2% compounded semiannually and Bank B other offers 9% compounded daily. We can compare these rates using the EARs. Which bank offers the highest interest rate? The effective annual rate for Bank A is $(1 + 0.046)^2 - 1 = 9.4\%$. The effective annual rate for Bank B is $(1 + 0.000247)^{365} - 1 = 9.42\%$. Therefore, Bank B offers the higher interest rate.

Yields on Investments

Suppose an investment opportunity requires an investor to put up \$1 million and offers cash inflows of \$500,000 after one year and \$600,000 after two years. The return on this investment, or *yield*, is the discount rate that equates the present values of the \$500,000 and \$600,000 cash inflows to equal the present value of the \$1 million cash outflow. This yield is also referred to as the *internal rate of return* (IRR) and is calculated as the rate that solves the following:

$$\$1,000,000 = \frac{\$500,000}{(1 + \text{IRR})^1} + \frac{\$600,000}{(1 + \text{IRR})^2}$$

Unfortunately, there is no direct mathematical solution (that is, closed-form solution) for the IRR, but rather we must use an iterative procedure. Fortunately, financial calculators and financial software ease our burden in this calculation. The IRR that solves this equation is 6.3941%:

$$\$1,000,000 = \frac{\$500,000}{(1.063941)^1} + \frac{\$600,000}{(1.063941)^2}$$

In other words, if you invest \$1 million today and receive \$500,000 in one year and \$600,000

in two years, the return on your investment is 6.3941%.

Another way of looking at this same yield is to consider that an investment's IRR is the discount rate that makes the present value of all expected future cash flows—both the cash outflows for the investment and the subsequent inflows—equal to zero. We can represent the IRR as the rate that solves:

$$\$0 = \sum_{t=1}^N \frac{CF_t}{(1 + \text{IRR})^t}$$

Consider another example. Suppose an investment of \$1 million produces no cash flow in the first year but cash flows of \$200,000, \$300,000, and \$900,000 two, three, and four years from now, respectively. The IRR for this investment is the discount rate that solves:

$$0 = \frac{\$1,000,000}{(1 + \text{IRR})^0} + \frac{0}{(1 + \text{IRR})^1} + \frac{\$200,000}{(1 + \text{IRR})^2} + \frac{\$300,000}{(1 + \text{IRR})^3} + \frac{\$900,000}{(1 + \text{IRR})^4}$$

Using a calculator or a computer, we get the precise answer of 10.172% per year.

We can use this approach to calculate the yield on any type of investment, as long as we know the cash flows—both positive and negative—and the timing of these flows. Consider the case of the yield to maturity on a bond. Most bonds pay interest semiannually—that is, every six months. Therefore, when calculating the yield on a bond, we must consider the timing of the cash flows to be such that the discount period is six months.

Consider a bond that has a current price of 90; that is, if the par value of the bond is \$1,000, the bond's price is 90% of \$1,000 or \$900. And suppose that this bond has five years remaining to maturity and an 8% coupon rate. With five years remaining to maturity, the bond has 10 six-month periods remaining. With a coupon rate of 8%, this means that the cash flows for

interest is \$40 every six months. For a given bond, we therefore have the following information:

1. Present value = \$900
2. Number of periods to maturity = 10
3. Cash flow every six months = \$40
4. Additional cash flow at maturity = \$1,000

The six-month yield, r_d , is the discount rate that solves the following:

$$\$900 = \left[\sum_{t=1}^{10} \frac{\$40}{(1 + r_d)^t} \right] + \frac{\$1,000}{(1 + r_d)^{10}}$$

Using a calculator or spreadsheet, the six-month yield is 5.315%. Bond yields are generally stated on the basis of an annualized yield, referred to as the yield to maturity (YTM) on a bond-equivalent basis. This YTM is analogous to the APR with semiannual compounding. Therefore, yield to maturity is 10.63%.

KEY POINTS

- A present value can be translated into a value in the future through compounding. The extreme frequency of compounding is continuous compounding.
- A future value can be converted into an equivalent value today through discounting.
- Applications in finance may require the determination of the present or future value of a series of cash flows rather than simply a single cash flow. The principles of determining the future value or present value of a series of cash flows are the same as for a single cash flow. That is, any number of cash flows can be translated into a present or future value.
- When faced with a series of cash flows, a financial modeler must value each cash flow individually, and then sum these individual values to arrive at the present value of the series.

- The tools of the time value of money can be used to value many different patterns of cash flows, including perpetuities, annuities due, and deferred annuities. Applying the tools to these different patterns of cash flows requires specifying the timing of the various cash flows.
- The interest on alternative investments is stated in different terms, so these interest rates must be placed on a common basis so that investment alternatives can be compared. Typically, an interest rate on an annual basis is specified, using either the annual percentage rate or the effective annual rate. The latter method is preferred since it takes into consideration the compounding of interest within a year.
- The yield on an investment (also referred to as internal rate of return) is the interest rate that makes the present value of the future cash flows equal to the cost of the investment.

NOTE

1. For a more detailed treatment of this topic, see Drake and Fabozzi (2009). The topic is covered in finite mathematics textbooks. See, for example, Barnett, Ziegler, and Byleen (2002), Mizrahi and Sullivan (1999), and Rolf (2007).

REFERENCES

- Barnett, R.A., Ziegler, M.R., and Byleen, K.E. (2002). *Finite Mathematics for Business, Economics, Life Sciences, and Social Sciences*, 9th ed. Upper Saddle River, NJ: Prentice Hall.
- Drake, P.P. and Fabozzi, F.J. (2009). *Foundations and Applications of the Time Value of Money*. Hoboken, NJ: John Wiley & Sons.
- Mizrahi, A. and Sullivan, M. (2003). *Finite Mathematics: An Applied Approach*, 9th ed. New York: John Wiley & Sons.
- Rolf, H.L. (2008). *Finite Mathematics*, 7th ed. Brooks Cole, Belmont, CA: Cengage Learning.

Fundamentals of Matrix Algebra

SERGIO M. FOCARDI, PhD
Partner, The Intertek Group

FRANK J. FABOZZI, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

Abstract: Ordinary algebra deals with operations such as addition and multiplication performed on individual numbers. In many applications, however, it is useful to consider operations performed on ordered arrays of numbers. This is the domain of matrix algebra. Ordered arrays of numbers are called vectors and matrices while individual numbers are called scalars.

In financial modeling, it is useful to consider operations performed on ordered arrays of numbers. Ordered arrays of numbers are called vectors and matrices while individual numbers are called scalars. In this entry, we will discuss some concepts, operations, and results of matrix algebra used in financial modeling.

VECTORS AND MATRICES DEFINED

We begin by defining the concepts of *vector* and *matrix*. Though vectors can be thought of as particular matrices, in many cases it is useful to keep the two concepts—vectors and matrices—distinct. In particular, a number of important concepts and properties can be defined for vectors but do not generalize easily to matrices.¹

Vectors

An n -dimensional *vector* is an ordered array of n numbers. Vectors are generally indicated

with boldface lowercase letters, although we do not always follow that convention in this book. Thus a vector \mathbf{x} is an array of the form:

$$\mathbf{x} = [x_1, \dots, x_n].$$

The numbers a_i are called the components of the vector \mathbf{x} .

A vector is identified by the set of its components. Vectors can be row vectors or column vectors. If the vector components appear in a horizontal row, then the vector is called a row vector, as for instance the vector:

$$\mathbf{x} = [1, 2, 8, 7]$$

Here are two examples. Suppose that we let w_n be a risky asset's weight in a portfolio. Assume that there are N risky assets. Then the following vector, \mathbf{w} , is a row vector that represents a portfolio's holdings of the N risky assets:

$$\mathbf{w} = [w_1 \ w_2 \ \dots \ w_N]$$

As a second example of a row vector, suppose that we let r_n be the excess return for a

risky asset. (The excess return is the difference between the return on a risky asset and the risk-free rate.) Then the following row vector is the excess return vector:

$$\mathbf{r} = [r_1 \ r_2 \ \dots \ r_N]$$

If the vector components are arranged in a column, then the vector is called a column vector.

For example, we know that a portfolio's excess return will be affected by what can be different characteristics or attributes that affect all asset prices. A few examples would be the price-earnings ratio, market capitalization, and industry. Let us denote for a particular attribute a column vector, \mathbf{a} , that shows the exposure of each risky asset to that attribute, denoted a_n :

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}$$

Matrices

An $n \times m$ matrix is a bidimensional ordered array of $n \times m$ numbers. Matrices are usually indicated with boldface uppercase letters. Thus, the generic matrix \mathbf{A} is an $n \times m$ array of the form:

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & \cdot & a_{1,j} & \cdot & a_{1,m} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{i,1} & \cdot & a_{i,j} & \cdot & a_{i,m} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n,1} & \cdot & a_{n,j} & \cdot & a_{n,m} \end{bmatrix}$$

Note that the first subscript indicates rows while the second subscript indicates columns. The entries a_{ij} —called the elements of the matrix \mathbf{A} —are the numbers at the crossing of the i -th row and the j -th column. The commas between the subscripts of the matrix entries are omitted when there is no risk of confusion: $a_{i,j} \equiv a_{ij}$. A matrix \mathbf{A} is often indicated by its generic element between brackets:

$$\mathbf{A} = \{a_{ij}\}_{nm} \quad \text{or} \quad \mathbf{A} = [a_{ij}]_{nm}$$

where the subscripts nm are the dimensions of the matrix.

There are several types of matrices. First there is a broad classification of square and rectangular matrices. A *rectangular matrix* can have different numbers of rows and columns; a *square matrix* is a rectangular matrix with the same number n of rows as of columns. Because of the important role that they play in applications, we focus on square matrices in the next section.

SQUARE MATRICES

The $n \times n$ *identity matrix*, indicated as the matrix \mathbf{I}_n , is a square matrix whose diagonal elements (i.e., the entries with the same row and column suffix) are equal to one while all other entries are zero:

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 1 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & 1 \end{bmatrix}$$

A matrix whose entries are all zero is called a zero matrix.

A *diagonal matrix* is a square matrix whose elements are all zero except the ones on the diagonal:

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & a_{22} & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & 0 & a_{nn} \end{bmatrix}$$

Given a square $n \times n$ matrix \mathbf{A} , the matrix $\text{dg } \mathbf{A}$ is the diagonal matrix extracted from \mathbf{A} . The diagonal matrix $\text{dg } \mathbf{A}$ is a matrix whose elements are all zero except the elements on the

diagonal that coincide with those of the matrix **A**:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1n} \\ a_{21} & a_{22} & \cdot & \cdot & \cdot & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \cdot & \cdot & \cdot & a_{nm} \end{bmatrix} \Rightarrow$$

$$\text{dg}\mathbf{A} = \begin{bmatrix} a_{11} & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & a_{22} & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & a_{nn} \end{bmatrix}$$

The *trace* of a square matrix **A** is the sum of its diagonal elements:

$$\text{tr}\mathbf{A} = \sum_{i=1}^n a_{ii}$$

A square matrix is called *symmetric* if the elements above the diagonal are equal to the corresponding elements below the diagonal: $a_{ij} = a_{ji}$. A matrix is said to be *skew-symmetric* if the diagonal elements are zero and the elements above the diagonal are the opposite of the corresponding elements below the diagonal: $a_{ij} = -a_{ji}, i \neq j, a_{ii} = 0$.

The most commonly used symmetric matrix in financial economics and econometrics is the covariance matrix, also referred to as the variance-covariance matrix. For example, suppose that there are N risky assets and that the variance of the excess return for each risky asset and the covariances between each pair of risky assets are estimated. As the number of risky assets is N , there are N^2 elements, consisting of N variances (along the diagonal) and $N^2 - N$ covariances. Symmetry restrictions reduce the number of independent elements. In fact, the covariance between risky asset i and risky asset j will be equal to the covariance between risky asset j and risky asset i . Notice that the variance-covariance matrix is a symmetric matrix.

DETERMINANTS

Consider a square, $n \times n$, matrix **A**. The *determinant* of **A**, denoted $|\mathbf{A}|$, is defined as follows:

$$|\mathbf{A}| = \sum (-1)^{t(j_1, \dots, j_n)} \prod_{i=1}^n a_{ij}$$

where the sum is extended over all permutations (j_1, \dots, j_n) of the set $(1, 2, \dots, n)$ and $t(j_1, \dots, j_n)$ is the number of transpositions (or inversions of positions) required to go from $(1, 2, \dots, n)$ to (j_1, \dots, j_n) . Otherwise stated, a determinant is the sum of all products formed taking exactly one element from each row with each product multiplied by $(-1)^{t(j_1, \dots, j_n)}$. Consider, for instance, the case $n = 2$, where there is only one possible transposition: $1, 2 \Rightarrow 2, 1$. The determinant of a 2×2 matrix is therefore computed as follows:

$$|\mathbf{A}| = (-1)^0 a_{11}a_{22} + (-1)^1 a_{12}a_{21}$$

$$= a_{11}a_{22} - a_{12}a_{21}.$$

Consider a square matrix **A** of order n . Consider the matrix \mathbf{M}_{ij} obtained by removing the i th row and the j th column. The matrix \mathbf{M}_{ij} is a square matrix of order $(n - 1)$. The determinant $|\mathbf{M}_{ij}|$ of the matrix \mathbf{M}_{ij} is called the *minor* of a_{ij} . The signed minor $(-1)^{(i+j)}|\mathbf{M}_{ij}|$ is called the *cofactor* of a_{ij} and is generally denoted as α_{ij} .

A square matrix **A** is said to be *singular* if its determinant is equal to zero. An $n \times m$ matrix **A** is of *rank* r if at least one of its (square) r -minors is different from zero while all $(r + 1)$ -minors, if any, are zero. A nonsingular square matrix is said to be of full rank if its rank r is equal to its order n .

SYSTEMS OF LINEAR EQUATIONS

A system of n linear equations in m unknown variables is a set of n simultaneous equations of the following form:

$$a_{1,1}x_1 + \dots + a_{1,m}x_m = b_1$$

$$\dots \dots \dots$$

$$a_{n,1}x_1 + \dots + a_{n,m}x_m = b_n$$

It can be demonstrated that in any matrix the number p of linearly independent columns is the same as the number q of linearly independent rows. This number is equal, in turn, to the rank r of the matrix. Recall that a $n \times m$ matrix \mathbf{A} is said to be of rank r if at least one of its (square) r -minors is different from zero while all $(r + 1)$ -minors, if any, are zero. The constant p , is the same for rows and for columns. We can now give an alternative definition of the rank of a matrix:

Given an $n \times m$ matrix \mathbf{A} , its *rank*, denoted $\text{rank}(\mathbf{A})$, is the number r of linearly independent rows or columns as the row rank is always equal to the column rank.

VECTOR AND MATRIX OPERATIONS

Let's now introduce the most common operations performed on vectors and matrices. An operation is a mapping that operates on scalars, vectors, and matrices to produce new scalars, vectors, or matrices. The notion of operations performed on a set of objects to produce another object of the same set is the key concept of algebra. Let's start with vector operations.

Vector Operations

The following three operations are usually defined on vectors: transpose, addition, and multiplication.

Transpose

The *transpose* operation transforms a row vector into a column vector and vice versa. Given the row vector $\mathbf{x} = [x_1, \dots, x_n]$, its transpose, denoted as \mathbf{x}^T or \mathbf{x}' , is the column vector:

$$\mathbf{x}^T = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}.$$

Clearly the transpose of the transpose is the original vector: $(\mathbf{x}^T)^T = \mathbf{x}$.

Addition

Two row (or column) vectors $\mathbf{x} = [x_1, \dots, x_n]$, $\mathbf{y} = [y_1, \dots, y_n]$ with the same number n of components can be added. The addition of two vectors is a new vector whose components are the sums of the components:

$$\mathbf{x} + \mathbf{y} = [x_1 + y_1, \dots, x_n + y_n]$$

This definition can be generalized to any number N of summands:

$$\sum_{i=1}^N \mathbf{x}_i = \left[\sum_{i=1}^N x_{1i}, \dots, \sum_{i=1}^N y_{ni} \right]$$

The summands must be both column or row vectors; it is not possible to add row vectors to column vectors.

It is clear from the definition of addition that addition is a commutative operation in the sense that the order of the summands does not matter: $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$. Addition is also an associative operation in the sense that $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$.

Multiplication

We define two types of multiplication: (1) multiplication of a scalar and a vector, and (2) scalar multiplication of two vectors (inner product).²

The multiplication of a scalar a and a row (or column) vector \mathbf{x} , denoted as $a\mathbf{x}$, is defined as the multiplication of each component of the vector by the scalar:

$$a\mathbf{x} = [ax_1, \dots, ax_n].$$

A similar definition holds for column vectors. It is clear from this definition that multiplication by a scalar is associative as:

$$a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$$

The *scalar product* (also called the inner product), of two vectors \mathbf{x}, \mathbf{y} , denoted as $\mathbf{x} \cdot \mathbf{y}$, is defined between a row vector and a column vector. The scalar product between two vectors produces a scalar according to the following rule:

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^n x_i y_i.$$

Two vectors \mathbf{x}, \mathbf{y} are said to be orthogonal if their scalar product is zero.

MATRIX OPERATIONS

Let's now define operations on matrices. The following five operations on matrices are usually defined: transpose, addition, multiplication, inverse, and adjoint.

Transpose

The definition of the *transpose of a matrix* is an extension of the transpose of a vector. The transpose operation consists in exchanging rows with columns. Consider the $n \times m$ matrix $\mathbf{A} = \{a_{ij}\}_{nm}$. The transpose of \mathbf{A} , denoted \mathbf{A}^T or \mathbf{A}' is the $m \times n$ matrix whose i th row is the i th column of \mathbf{A} :

$$\mathbf{A}^T = \{a_{ji}\}_{mn}$$

The following should be clear from this definition:

$$(\mathbf{A}^T)^T = \mathbf{A}$$

and that a matrix is symmetric if and only if

$$\mathbf{A}^T = \mathbf{A}$$

Addition

Consider two $n \times m$ matrices $\mathbf{A} = \{a_{ij}\}_{nm}$ and $\mathbf{B} = \{b_{ij}\}_{nm}$. The sum of the matrices \mathbf{A} and \mathbf{B} is defined as the $n \times m$ matrix obtained by adding the respective elements:

$$\mathbf{A} + \mathbf{B} = \{a_{ij} + b_{ij}\}_{nm}.$$

Note that it is essential for the definition of addition that the two matrices have the same order $n \times m$.

The operation of addition can be extended to any number N of summands as follows:

$$\sum_{s=1}^N \mathbf{A}_s = \left\{ \sum_{s=1}^N a_{sij} \right\}_{nm}$$

where a_{sij} is the generic i, j element of the s th summand.

Multiplication

Consider a scalar c and a matrix $\mathbf{A} = \{a_{ij}\}_{nm}$. The product $c\mathbf{A} = \mathbf{A}c$ is the $n \times m$ matrix obtained by multiplying each element of the matrix by c :

$$c\mathbf{A} = \mathbf{A}c = \{ca_{ij}\}_{nm}.$$

Multiplication of a matrix by a scalar is distributive with respect to matrix addition:

$$c(\mathbf{A} + \mathbf{B}) = c\mathbf{A} + c\mathbf{B}.$$

Let's now define the product of two matrices. Consider two matrices $\mathbf{A} = \{a_{it}\}_{np}$ and $\mathbf{B} = \{b_{sj}\}_{pm}$. The product $\mathbf{C} = \mathbf{A}\mathbf{B}$ is defined as follows:

$$\mathbf{C} = \mathbf{A}\mathbf{B} = \{c_{ij}\} = \left\{ \sum_{t=1}^p a_{it} b_{tj} \right\}.$$

The product $\mathbf{C} = \mathbf{A}\mathbf{B}$ is therefore a matrix whose generic element $\{c_{ij}\}$ is the scalar product of the i th row of the matrix \mathbf{A} and the j th column of the matrix \mathbf{B} . This definition generalizes the definition of scalar product of vectors: The scalar product of two n -dimensional vectors is the product of an $n \times 1$ matrix (a row vector) for a $1 \times n$ matrix (the column vector).

Inverse and Adjoint

Consider two square matrices of order n \mathbf{A} and \mathbf{B} . If $\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A} = \mathbf{I}$, then the matrix \mathbf{B} is called the *inverse* of \mathbf{A} and is denoted as \mathbf{A}^{-1} . It can be demonstrated that the two following properties hold:

Property 1: A square matrix \mathbf{A} admits an inverse \mathbf{A}^{-1} if and only if it is nonsingular, that is, if and only if its determinant is different from zero. Otherwise stated, a matrix \mathbf{A} admits an inverse if and only if it is of full rank.

Property 2: The inverse of a square matrix, if it exists, is unique. This property is a consequence of the property that, if \mathbf{A} is nonsingular, then $\mathbf{AB} = \mathbf{AC}$ implies $\mathbf{B} = \mathbf{C}$.

Consider now a square matrix of order n $\mathbf{A} = \{a_{ij}\}$ and consider its cofactors α_{ij} . Recall that the cofactors α_{ij} are the signed minors $(-1)^{i+j}|M_{ij}|$ of the matrix \mathbf{A} . The *adjoint* of the matrix \mathbf{A} , denoted as $\text{Adj}(\mathbf{A})$, is the following matrix:

$$\begin{aligned} \text{Adj}(\mathbf{A}) &= \begin{bmatrix} \alpha_{1,1} & \cdot & \alpha_{1,j} & \cdot & \alpha_{1,n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \alpha_{i,1} & \cdot & \alpha_{i,j} & \cdot & \alpha_{i,n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \alpha_{n,1} & \cdot & \alpha_{n,j} & \cdot & \alpha_{n,n} \end{bmatrix}^T \\ &= \begin{bmatrix} \alpha_{1,1} \cdot \alpha_{2,1} \cdot \alpha_{n,1} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \alpha_{1,i} \cdot \alpha_{2,i} \cdot \alpha_{n,i} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \alpha_{1,n} \cdot \alpha_{2,n} \cdot \alpha_{n,n} \end{bmatrix} \end{aligned}$$

The adjoint of a matrix \mathbf{A} is therefore the transpose of the matrix obtained by replacing the elements of \mathbf{A} with their cofactors.

If the matrix \mathbf{A} is nonsingular, and therefore admits an inverse, it can be demonstrated that:

$$\mathbf{A}^{-1} = \frac{\text{Adj}(\mathbf{A})}{|\mathbf{A}|}$$

A square matrix of order n \mathbf{A} is said to be orthogonal if the following property holds:

$$\mathbf{AA}' = \mathbf{A}'\mathbf{A} = \mathbf{I}_n$$

Because in this case \mathbf{A} must be of full rank, the transpose of an orthogonal matrix coincides with its inverse: $\mathbf{A}^{-1} = \mathbf{A}'$.

EIGENVALUES AND EIGENVECTORS

Consider a square matrix \mathbf{A} of order n and the set of all n -dimensional vectors. The matrix \mathbf{A} is a linear operator on the space of vectors. This means that \mathbf{A} operates on each vector producing another vector subject to the following restriction:

$$\mathbf{A}(a\mathbf{x} + b\mathbf{y}) = a\mathbf{A}\mathbf{x} + b\mathbf{A}\mathbf{y}$$

Consider now the set of vectors \mathbf{x} such that the following property holds:

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}.$$

Any vector such that the above property holds is called an *eigenvector* of the matrix \mathbf{A} and the corresponding value of λ is called an *eigenvalue*.

To determine the eigenvectors of a matrix and the relative eigenvalues, consider that the equation $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ can be written as:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$$

which can, in turn, be written as a system of linear equations:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \begin{bmatrix} a_{1,1} - \lambda & \cdot & a_{1,j} & \cdot & a_{1,n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{i,1} & \cdot & a_{i,i} - \lambda & \cdot & a_{i,n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n,1} & \cdot & a_{n,j} & \cdot & a_{n,n} - \lambda \end{bmatrix} \begin{bmatrix} x_1 \\ \cdot \\ x_i \\ \cdot \\ x_n \end{bmatrix} = 0$$

This system of equations has nontrivial solutions only if the matrix $\mathbf{A} - \lambda\mathbf{I}$ is singular. To determine the eigenvectors and the eigenvalues of the matrix \mathbf{A} we must therefore solve the equation:

$$|\mathbf{A} - \lambda\mathbf{I}| = \begin{vmatrix} a_{1,1} - \lambda & \cdot & a_{1,j} & \cdot & a_{1,n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{i,1} & \cdot & a_{i,i} - \lambda & \cdot & a_{i,n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n,1} & \cdot & a_{n,j} & \cdot & a_{n,n} - \lambda \end{vmatrix} = 0$$

The expansion of this determinant yields a polynomial $\phi(\lambda)$ of degree n known as the *characteristic polynomial* of the matrix \mathbf{A} . The equation $\phi(\lambda) = 0$ is known as the *characteristic equation* of the matrix \mathbf{A} . In general, this equation will have n roots λ_s which are the eigenvalues of the matrix \mathbf{A} . To each of these eigenvalues corresponds a solution of the system of linear equations as illustrated below:

$$\begin{bmatrix} a_{1,1} - \lambda_s & \cdot & a_{1,j} & \cdot & a_{1,n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{i,1} & \cdot & a_{i,i} - \lambda_s & \cdot & a_{i,n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n,1} & \cdot & a_{n,j} & \cdot & a_{n,n} - \lambda_s \end{bmatrix} \begin{bmatrix} x_{1s} \\ \cdot \\ x_{is} \\ \cdot \\ x_{ns} \end{bmatrix} = 0$$

Each solution represents the eigenvector \mathbf{x}_s corresponding to the eigenvalue λ_s . The determination of eigenvalues and eigenvectors is the basis for principal component analysis.

KEY POINTS

- An n -dimensional vector is an ordered array of n numbers with the numbers referred to as the components. An $n \times m$ matrix is a bidimensional ordered array of $n \times m$ numbers.
- A rectangular matrix can have different numbers of rows and columns; a square matrix is a rectangular matrix with the same number of rows and columns. An identity matrix is a square matrix whose diagonal elements are equal to one while all other entries are zero.

A diagonal matrix is a square matrix whose elements are all zero except the ones on the diagonal.

- The trace of a square matrix is the sum of its diagonal elements. A symmetric matrix is a square matrix where the elements above the diagonal are equal to the corresponding elements below the diagonal. The most commonly used symmetric matrix in finance is the covariance matrix (or variance-covariance matrix).
- The rank of a matrix is used to determine the number of solutions of a system of linear equations.
- An operation is a mapping that operates on scalars, vectors, and matrices to produce new scalars, vectors, or matrices. The notion of operations performed on a set of objects to produce another object of the same set is the key concept of algebra. Five vector operations on matrices are transpose, addition, multiplication, inverse, and adjoint.

NOTES

1. Vectors can be thought of as the elements of an abstract linear space while matrices are operators that operate on linear spaces.
2. A third type of product between vectors—the vector (or outer) product between vectors—produces a third vector. We do not define it here as it is not typically used in economics, though widely used in the physical sciences.

Difference Equations

SERGIO M. FOCARDI, PhD

Partner, The Intertek Group

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: The theory of linear difference equations has found applications in many areas in finance. A difference equation is an equation that involves differences between successive values of a function of a discrete variance. The theory of linear difference equations covers three areas: solving difference equations, describing the behavior of difference equations, and identifying the equilibrium (or critical value) and stability of difference equations.

Linear difference equations are important in the context of dynamic econometric models. Stochastic models in finance are expressed as linear difference equations with random disturbances added. Understanding the behavior of solutions of linear difference equations helps develop intuition about the behavior of these models. The relationship between *difference equations* (the subject of this entry) and differential equations is as follows. The latter are great for modeling situations in finance where there is a continually changing value. The problem is that not all changes in value occur continuously. If the change in value occurs incrementally rather than continuously, then differential equations have their limitations. Instead, a financial modeler can use difference equations, which are recursively defined sequences.

In this entry we explain the theory of linear difference equations and describe how to

compute explicit solutions of different types of equations.

THE LAG OPERATOR L

The *lag operator* L is a linear operator that acts on doubly infinite time series by shifting positions by one place:

$$Lx_t = x_{t-1}$$

The difference operator $\Delta x_t = x_t - x_{t-1}$ can be written in terms of the lag operator as

$$\Delta x_t = (1 - L)x_t$$

Products and thus powers of the lag operator are defined as follows:

$$(L \times L)x_t = L^2x_t = L(Lx_t) = x_{t-2}$$

From the previous definition, we can see that the i -th power of the lag operator shifts the

series by i places:

$$L^i x_t = x_{t-i}$$

The lag operator is linear, that is, given scalars a and b we have

$$(aL^i + bL^j)x_t = ax_{t-i} + bx_{t-j}$$

Hence we can define the polynomial operator:

$$A(L) = (1 - a_1L - \dots - a_pL^p) \equiv \left(1 - \sum_{i=1}^p a_iL^i\right)$$

HOMOGENEOUS DIFFERENCE EQUATIONS

Homogeneous difference equations are linear conditions that link the values of variables at different time lags. Using the lag operator L , they can be written as follows:

$$\begin{aligned} A(L)x_t &= (1 - a_1L - \dots - a_pL^p)x_t \\ &= (1 - \lambda_1L) \times \dots \times (1 - \lambda_pL)x_t = 0 \end{aligned}$$

where the $\lambda_i, i = 1, 2, \dots, p$ are the solutions of the characteristic equation:

$$\begin{aligned} z^p - a_1z^{p-1} - \dots - a_{p-1}z - a_p &= 0 \\ &= (z - \lambda_1) \times \dots \times (z - \lambda_p) = 0 \end{aligned}$$

Suppose that time extends from $0 \Rightarrow \infty, t = 0, 1, 2, \dots$ and that the initial conditions $(x_{-1}, x_{-2}, \dots, x_{-p})$ are given.

Real Roots

Consider first the case of real roots. In this case, as we see later in this entry, solutions are sums of exponentials. First suppose that the roots of the characteristic equation are all real and distinct. It can be verified by substitution that any series of the form

$$x_t = C(\lambda_i)^t$$

where C is a constant, solves the homogeneous difference equation. In fact, we can write

$$(1 - \lambda_iL)(C\lambda_i^t) = C\lambda_i^t - \lambda_iC\lambda_i^{t-1} = 0$$

In addition, given the linearity of the lag operator, any linear combination of solutions of the

homogeneous difference equation is another solution. We can therefore state that the following series solves the homogeneous difference equation:

$$x_t = \sum_{i=1}^p C_i \lambda_i^t$$

By solving the linear system

$$\begin{aligned} x_{-1} &= \sum_{i=1}^p C_i \lambda_i^{-1} \\ x_{-p} &= \sum_{i=1}^p C_i \lambda_i^{-p} \end{aligned}$$

that states that the p initial conditions are satisfied, we can determine the p constants C_s .

Suppose now that all m roots of the characteristic equation are real and coincident. In this case, we can represent a difference equation in the following way:

$$A(L) = 1 - a_1L - \dots - a_pL^p = (1 - \lambda L)^p$$

It can be demonstrated by substitution that, in this case, the general solution of the process is the following:

$$x_t = C_1(\lambda)^t + C_2t(\lambda)^t + \dots + C_p t^{p-1}(\lambda)^t$$

In the most general case, assuming that all roots are real, there will be $m < p$ distinct roots $\varphi_i, i = 1, 2, \dots, m$ each of order $n_i \geq 1$,

$$\sum_{i=1}^m n_i = p$$

and the general solution of the process will be

$$\begin{aligned} x_t &= C_1^1(\lambda_1)^t + C_2^1t(\lambda_1)^t + \dots + C_{n_1}^1 t^{n_1-1}(\lambda_1)^t + \dots \\ &\quad + C_1^m(\lambda_m)^t + C_2^m t(\lambda_m)^t + \dots + C_{n_m}^m t^{n_m-1}(\lambda_m)^t \end{aligned}$$

We can therefore conclude that the solutions of a homogeneous difference equation whose characteristic equation has only real roots is formed by a sum of exponentials. If these roots have modulus greater than unity, then solutions are diverging exponentials; if they have modulus smaller than unity, solutions are exponentials that go to zero. If the roots are unity, solutions are either constants or, if the roots have multiplicity greater than 1, polynomials.

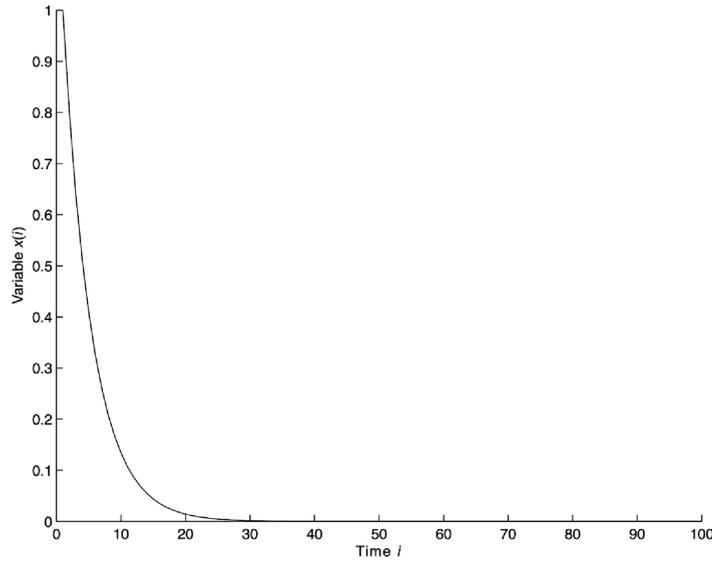


Figure 1 Solution of the Equation $(1 - 0.8L)x_t = 0$ with Initial Condition $x_1 = 1$

Figure 1 illustrates the simple equation

$$A(L)x_t = (1 - 0.8L)x_t = 0, \lambda = 0.8,$$

$$t = 1, 2, \dots, n, \dots$$

whose solution, with initial condition $x_1 = 1$, is

$$x_t = 1.25(0.8)^t$$

The behavior of the solution is that of an exponential decay.

Figure 2 illustrates the equation

$$A(L)x_t = (1 + 0.8L)x_t = 0, \lambda = -0.8,$$

$$t = 1, 2, \dots, n, \dots$$

Simulations were run for 100 time steps

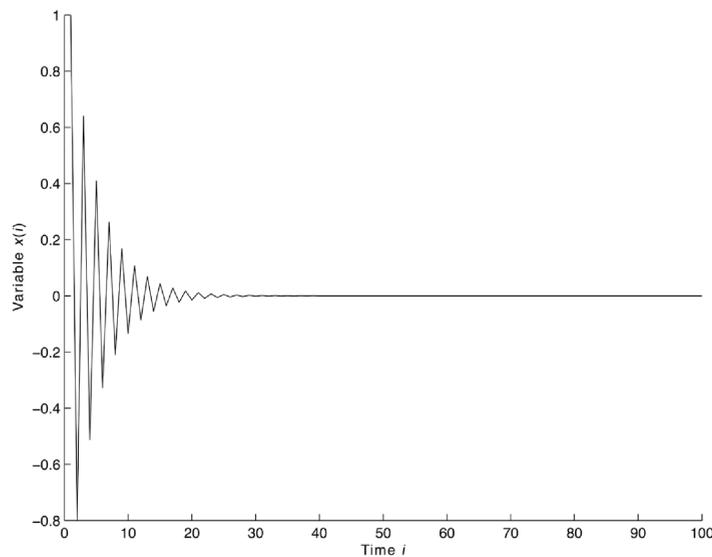


Figure 2 Solution of the Equation $(1 + 0.8L)x_t = 0$ with Initial Condition $x_1 = 1$

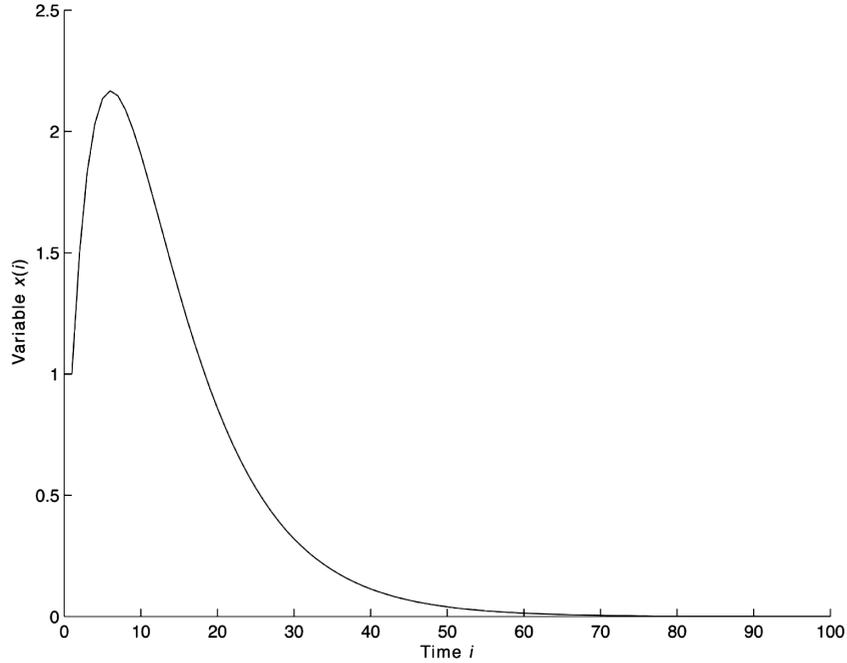


Figure 3 Solution of the Equation $(1 - 1.7L + 0.72L^2)x_t = 0$ with Initial Conditions $x_1 = 1, x_2 = 1.5$

whose solution, with initial condition $x_1 = 1$, is

$$x_t = -1.25(-0.8)^t$$

The behavior of the solution is that of an exponential decay with oscillations at each step. The oscillations are due to the change in sign of the exponential at odd and even time steps.

If the equation has more than one real root, then the solution is a sum of exponentials. Figure 3 illustrates the equation

$$A(L)x_t = (1 - 1.7L + 0.72L^2)x_t = 0, \lambda_1 = 0.8, \\ \lambda_2 = 0.9, \quad t = 1, 2, \dots, n, \dots$$

whose solution, with initial condition $x_1 = 1, x_2 = 1.5$, is

$$x_t = -7.5(0.8)^t + 7.7778(0.9)^t$$

The behavior of the solution is that of an exponential decay after a peak.

Figure 4 illustrates the equation

$$A(L)x_t = (1 - 1.9L + 0.88L^2)x_t = 0, \\ \lambda_1 = 0.8, \lambda_2 = 1.1, \quad t = 1, 2, \dots, n, \dots$$

whose solution, with initial condition $x_1 = 1, x_2 = 1.5$, is

$$x_t = -1.6667(0.8)^t + 2.1212(1.1)^t$$

The behavior is that of exponential explosion due to the exponential with modulus greater than 1.

Complex Roots

Now suppose that some of the roots are complex. In this case, solutions exhibit an oscillating behavior with a period that depends on the model coefficients. For simplicity, consider initially a second-order homogeneous difference equation:

$$A(L)x_t = (1 - a_1L - a_2L^2)x_t$$

Suppose that its characteristic equation given by

$$A(z) = z^2 - a_1z - a_2 = 0$$

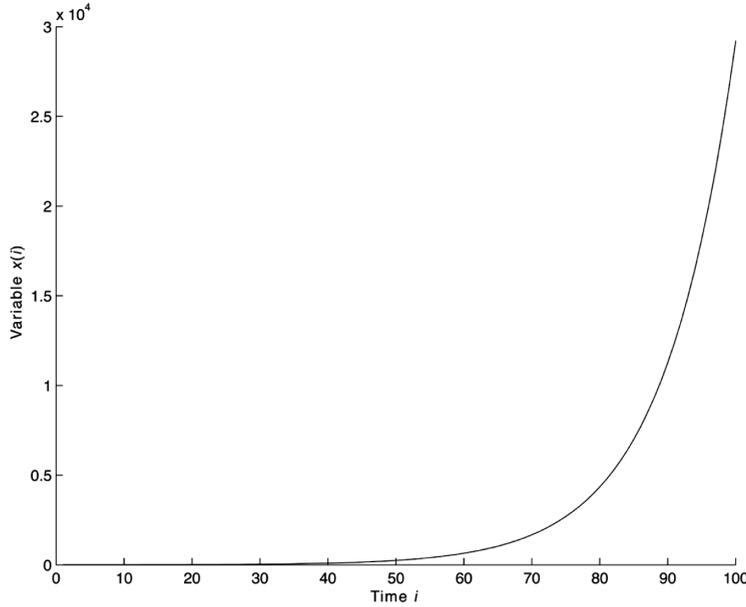


Figure 4 Solution of the Equation $(1 - 1.9L + 0.88L^2)x_t = 0$ with Initial Conditions $x_1 = 1, x_2 = 1.5$

admits the two complex conjugate roots:

$$\lambda_1 = a + ib, \quad \lambda_2 = a - ib$$

Let's write the two roots in polar notation:

$$\lambda_1 = r e^{i\omega}, \quad \lambda_2 = r e^{-i\omega}$$

$$r = \sqrt{a^2 + b^2}, \quad \omega = \arctan \frac{b}{a}$$

It can be demonstrated that the general solution of the above difference equation has the following form:

$$x_t = r^t (C_1 \cos(\omega t) + C_2 \sin(\omega t)) = C r^t \cos(\omega t + \vartheta)$$

where the C_1 and C_2 or C and ϑ are constants to be determined in function of initial conditions. If the imaginary part of the roots vanishes, then ω vanishes and $a = r$, the two complex conjugate roots become a real root, and we find again the expression $x_t = C r^t$.

Consider now a homogeneous difference equation of order $2n$. Suppose that the characteristic equation has only two distinct complex conjugate roots with multiplicity n . We can write the difference equation as follows:

$$A(L)x_t = (1 - a_1 L - \dots - a_{2n} L^{2n})x_t = 0$$

$$= [(1 - \lambda L)^n (1 - \bar{\lambda} L)^n] x_t = 0$$

and its general solution as follows:

$$x_t = r^t (C_1^1 \cos(\omega t) + C_2^1 \sin(\omega t)) + \dots$$

$$+ t^n r^t (C_1^n \cos(\omega t) + C_2^n \sin(\omega t))$$

The general solution of a homogeneous difference equation that admits both real and complex roots with different multiplicities is a sum of the different types of solutions. The above formulas show that real roots correspond to a sum of exponentials while complex roots correspond to oscillating series with exponential dumping or explosive behavior. The above formulas confirm that in both the real and the complex case, solutions decay if the modulus of the roots of the inverse characteristic equation is outside the unit circle and explode if it is inside the unit circle.

Figure 5 illustrates the equation

$$A(L)x_t = (1 - 1.2L + 1.0L^2)x_t = 0,$$

$$t = 1, 2, \dots, n, \dots$$

which has two complex conjugate roots,

$$\lambda_1 = 0.6 + i0.8, \quad \lambda_2 = 0.6 - i0.8$$

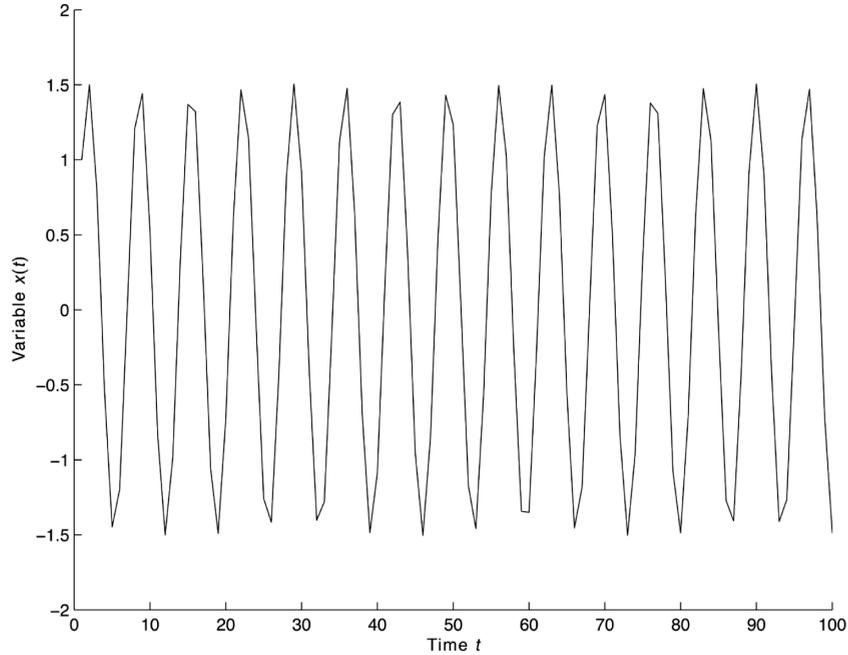


Figure 5 Solutions of the Equation $(1 - 1.2L + 1.0L^2)x_t = 0$ with Initial Conditions $x_1 = 1, x_2 = 1.5$

or in polar form,

$$\lambda_1 = e^{i0.9273}, \quad \lambda_2 = e^{-i0.9273}$$

and whose solution, with initial condition $x_1 = 1, x_2 = 1.5$, is

$$x_t = -0.3 \cos(0.9273t) + 1.475 \sin(0.9273t)$$

The behavior of the solutions is that of undamped oscillations with frequency determined by the model.

Figure 6 illustrates the equation

$$A(L)x_t = (1 - 1.0L + 0.89L^2)x_t = 0, \\ t = 1, 2, \dots, n, \dots$$

which has two complex conjugate roots,

$$\lambda_1 = 0.5 + i0.8, \quad \lambda_2 = 0.5 - i0.8$$

or in polar form,

$$\lambda_1 = 0.9434e^{i1.0122}, \quad \lambda_2 = 0.9434e^{-i1.0122}$$

and whose solution, with initial condition $x_1 = 1, x_2 = 1.5$, is

$$x_t = 0.9434^t (-0.5618 \cos(1.0122t) \\ + 1.6011 \sin(1.0122t))$$

The behavior of the solutions is that of damped oscillations with frequency determined by the model.

NONHOMOGENEOUS DIFFERENCE EQUATIONS

Consider now the following n -th order difference equation:

$$A(L)x_t = (1 - a_1L - \dots - a_pL^p)x_t = y_t$$

where y_t is a given sequence of real numbers. Recall that we are in a deterministic setting, that is, the y_t are given. The general solution of the above difference equation will be the sum of two solutions $x_{1,t} + x_{2,t}$ where $x_{1,t}$ is the solution of the associated homogeneous equation,

$$A(L)x_t = (1 - a_1L - \dots - a_pL^p)x_t = 0$$

and $X_{2,t}$ solves the given *nonhomogeneous equation*.

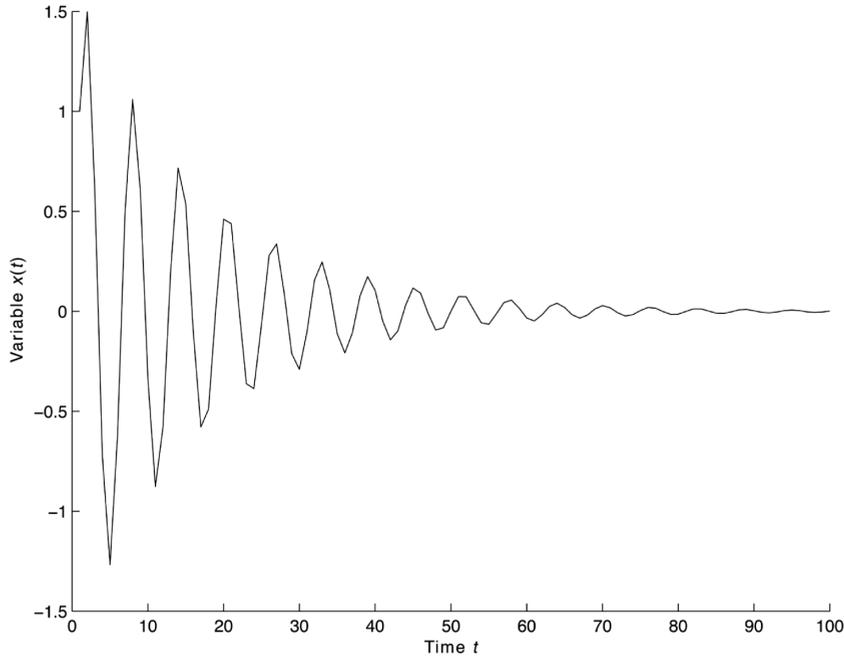


Figure 6 Solutions of the Equation $(1 - 1.0L + 0.89L^2)x_t = 0$ with Initial Conditions $x_1 = 1, x_2 = 1.5$

Real Roots

To determine the general form of $x_{2,t}$ in the case of real roots, we begin by considering the case of a first-order equation:

$$A(L)x_t = (1 - a_1L)x_t = y_t$$

We can compute the solution as follows:

$$x_{2,t} = \frac{1}{(1 - a_1L)}y_t = \left(\sum_{j=0}^{\infty} (a_1L)^j \right) y_t$$

which is meaningful only for $|a_1| < 1$. If, however, y_t starts at $t = -1$, that is, if $y_t = 0$ for $t = -2, -3, \dots, n$, we can rewrite the above formula as

$$x_{2,t} = \frac{1}{(1 - a_1L)}y_t = \left(\sum_{j=0}^{t+1} (a_1L)^j \right) y_t$$

This latter formula, which is valid for any real value of a_1 , yields

$$\begin{aligned} x_{2,0} &= y_0 + a_1y_{-1} \\ x_{2,1} &= y_1 + a_1y_0 + a_1^2y_{-1} \\ x_{2,t} &= y_t + a_1y_{t-1} + \dots + a_1^{t+1}y_{-1} \end{aligned}$$

and so on. These formulas can be easily verified by direct substitution. If $y_t = y = \text{constant}$, then

$$x_{2,t} = y(1 + a_1^2 + \dots + a_1^{t+1})$$

Consider now the case of a second-order equation:

$$\begin{aligned} A(L)x_t &= (1 - a_1L - a_2L^2)x_t \\ &= (1 - \lambda_1L)(1 - \lambda_2L)x_t = y_t \end{aligned}$$

where λ_1, λ_2 are the solutions of the characteristic equation (the reciprocal of the solutions of the inverse characteristic equation). We can write the solution of the above equation as

$$x_{2,t} = \frac{1}{(1 - a_1L - a_2L^2)}y_t = \frac{1}{(1 - \lambda_1L)(1 - \lambda_2L)}y_t$$

Recall that, if $|\lambda_i| < 1, i = 1, 2$, we can write:

$$\begin{aligned} &\frac{1}{(1 - \lambda_1L)(1 - \lambda_2L)} \\ &= \frac{1}{\lambda_1 - \lambda_2} \left(\frac{\lambda_1}{(1 - \lambda_1L)} - \frac{\lambda_2}{(1 - \lambda_2L)} \right) \\ &= \frac{\lambda_1}{\lambda_1 - \lambda_2} \left(\sum_{j=0}^{\infty} (\lambda_1L)^j \right) - \frac{\lambda_2}{\lambda_1 - \lambda_2} \left(\sum_{j=0}^{\infty} (\lambda_2L)^j \right) \end{aligned}$$

so that the solution can be written as

$$x_{2,t} = \frac{\lambda_1}{\lambda_1 - \lambda_2} \left(\sum_{j=0}^{\infty} (\lambda_1 L)^j \right) y_t - \frac{\lambda_2}{\lambda_1 - \lambda_2} \left(\sum_{j=0}^{\infty} (\lambda_2 L)^j \right) y_t$$

If the two solutions are coincident, reasoning as in the homogeneous case, we can establish that the general solutions can be written as follows:

$$x_{2,t} = \frac{1}{(1 - a_1 L)^2} y_t = \left(\sum_{j=0}^{\infty} (a_1 L)^j \right) y_t + t \left(\sum_{j=0}^{\infty} (a_1 L)^j \right) y_t$$

If y_t starts at $t = -2$, that is, if $y_t = 0$ for $t = -3, -4, \dots, -n, \dots$, we can rewrite the above formula respectively as

$$x_{2,t} = \frac{\lambda_1}{\lambda_1 - \lambda_2} \left(\sum_{j=0}^{t+2} (\lambda_1 L)^j \right) y_t - \frac{\lambda_2}{\lambda_1 - \lambda_2} \left(\sum_{j=0}^{t+2} (\lambda_2 L)^j \right) y_t$$

if the solutions are distinct, and as

$$x_{2,t} = \frac{1}{(1 - a_1 L)^2} y_t = \left(\sum_{j=0}^{t+2} (a_1 L)^j \right) y_t + t \left(\sum_{j=0}^{t+2} (a_1 L)^j \right) y_t$$

if the solutions are coincident. These formulas are valid for any real value of λ_1 .

The above formulas can be generalized to cover the case of an n -th order difference equation. In the most general case of an n -th order difference equation, assuming that all roots are real, there will be $m < n$ distinct roots $\lambda_i, i = 1, 2, \dots, m$, each of order $n_i \geq 1$,

$$\sum_{i=1}^m n_i = n$$

and the general solution of the process will be

$$x_{2,t} = \sum_{i=0}^{\infty} ((\lambda_1 L)^i + i(\lambda_1 L)^i + \dots + i^{n_1-1}(\lambda_1 L)^i + \dots + (\lambda_m L)^i + i(\lambda_m L)^i + \dots + i^{n_m-1}(\lambda_m L)^i) y_t$$

if $|\lambda_i| < 1, i = 1, 2, \dots, m$, and

$$x_{2,t} = \sum_{i=0}^{t+m} ((\lambda_1 L)^i + i(\lambda_1 L)^i + \dots + i^{n_1-1}(\lambda_1 L)^i + \dots + (\lambda_m L)^i + i(\lambda_m L)^i + \dots + i^{n_m-1}(\lambda_m L)^i) y_t$$

if y_t starts at $t = -n$, that is, if $y_t = 0$ for $t = -(n+1), -(n+2), \dots$ for any real value of the λ_i .

Therefore, if the roots are all real, the general solution of a difference equation is a sum of exponentials. Figure 7 illustrates the case of the same difference equation as in Figure 3 with the same initial conditions $x_1 = 1, x_2 = 1.5$ but with an exogenous forcing sinusoidal variable:

$$(1 - 1.7L + 0.72L^2)x_t = 0.1 \times \sin(0.4 \times t)$$

The solution of the equation is the sum of $x_{1,t} = -7.5(0.8)^t + 7.7778(0.9)^t$ plus

$$x_{2,t} = \sum [((0.8)^i + (0.9)^i)0.1 \times \sin(0.4 \times (t - i))]$$

After the initial phase dominated by the solution of the homogeneous equation, the forcing term dictates the shape of the solution.

Complex Roots

Consider now the case of complex roots. For simplicity, consider initially a second-order difference equation:

$$A(L)x_t = (1 - a_1 L - a_2 L^2)x_t = y_t$$

Suppose that its characteristic equation,

$$A(z) = z^2 - a_1 z - a_2 = 0$$

admits the two complex conjugate roots,

$$\lambda_1 = a + ib, \quad \lambda_2 = a - ib$$

We write the two roots in polar notation:

$$\lambda_1 = r e^{i\omega}, \quad \lambda_2 = r e^{-i\omega} \\ r = \sqrt{a^2 + b^2}, \quad \omega = \arctan \frac{b}{a}$$

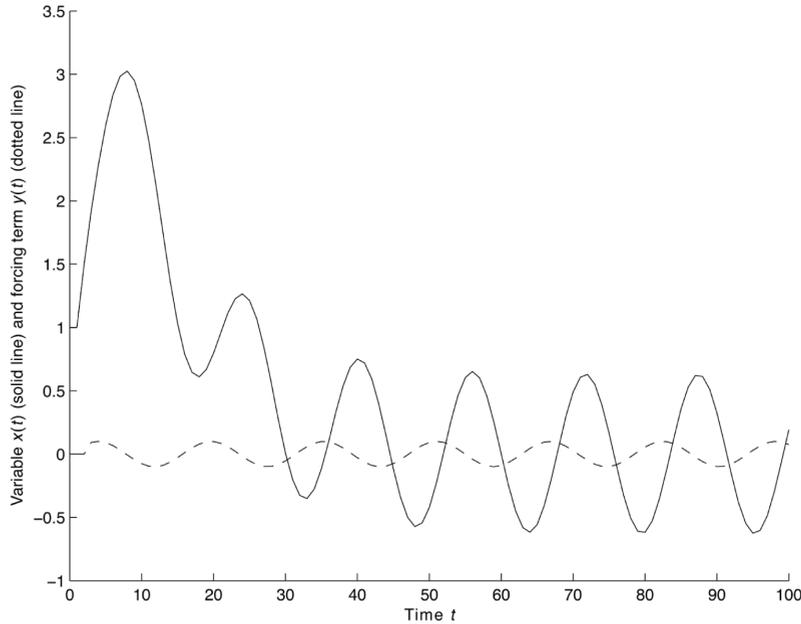


Figure 7 Solutions of the Equation $(1 - 1.7L + 0.72L^2)x_t = 0.1 \times \sin(0.4 \times t)$ with Initial Conditions $x_1 = 1, x_2 = 1.5$

It can be demonstrated that the general form of the $x_{2,t}$ of the above difference equation has the following form:

$$x_{2,t} = \sum_{i=1}^{\infty} (r^i (\cos(\omega i) + \sin(\omega i)) y_{t-i})$$

which is meaningful only if $|r| < 1$. If y_t starts at $t = -2$, that is, if $y_t = 0$ for $t = -3, -4, \dots, -n, \dots$ we can rewrite the previous formula as

$$x_{2,t} = \sum_{i=1}^{t+2} (r^i (\cos(\omega i) + \sin(\omega i)) y_{t-i})$$

This latter formula is meaningful for any real value of r . Note that the constant ω is determined by the structure of the model while the constants C_1, C_2 that appear in $x_{1,t}$ need to be determined in the function of initial conditions. If the imaginary part of the roots vanishes, then ω vanishes and $a = r$, the two complex conjugate roots become a real root, and we again find the expression $x_t = Cr^t$.

Figure 8 illustrates the case of the same difference equation as in Figure 7 with the same initial conditions $x_1 = 1, x_2 = 1.5$ but with an

exogenous forcing sinusoidal variable:

$$(1 - 1.2L + 1.0L^2)x_t = 0.5 \times \sin(0.4 \times t)$$

The solution of the equation is the sum of $x_{1,t} = -0.3\cos(0.9273t) + 1.475 \sin(0.9273t)$ plus

$$x_{2,t} = \sum_{i=0}^{t-1} [(\cos(0.9273i) + \sin(0.9273i))0.5 \sin(0.4 \times (t - i))]$$

After the initial phase dominated by the solution of the homogeneous equation, the forcing term dictates the shape of the solution. Note the model produces amplification and phase shift of the forcing term $0.1 \times \sin(0.4 \times t)$ represented by a dotted line.

SYSTEMS OF LINEAR DIFFERENCE EQUATIONS

In this section, we discuss *systems of linear difference equations* of the type

$$\begin{aligned} x_{1,t} &= a_{11}x_{1,t-1} + \dots + a_{1k}x_{k,t-1} + y_{1,t} \\ x_{k,t} &= a_{k1}x_{1,t-1} + \dots + a_{kk}x_{k,t-1} + y_{k,t} \end{aligned}$$

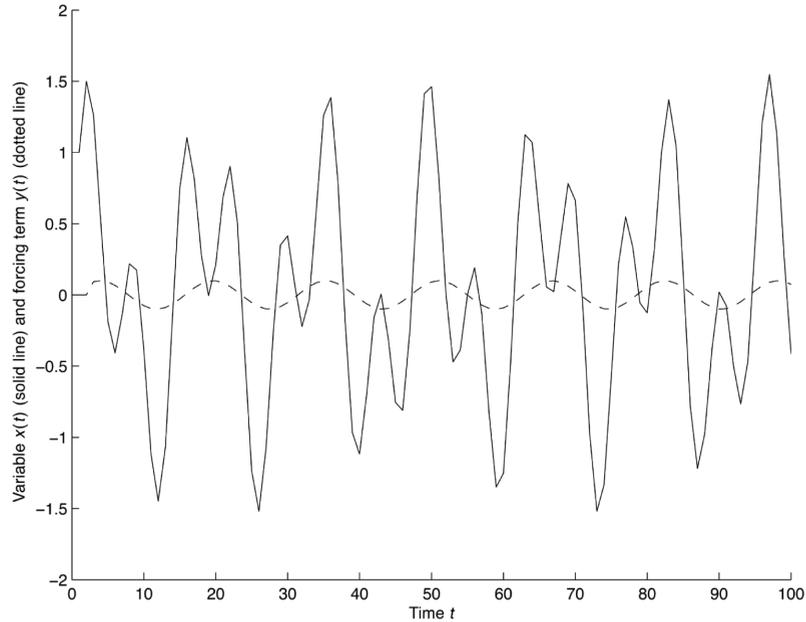


Figure 8 Solutions of the Equation $(1 - 1.2L + 1.0L^2)x_t = 0.5 \times \sin(0.4 \times t)$ with Initial Conditions $x_1 = 1, x_2 = 1.5$

or in vector notation:

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{y}_t$$

Observe that we need to consider only first-order systems, that is, systems with only one lag. In fact, a system of an arbitrary order can be transformed into a first-order system by adding one variable for each additional lag. For example, a second-order system of two difference equations,

$$\begin{aligned} x_{1,t} &= a_{11}x_{1,t-1} + a_{12}x_{2,t-1} + b_{11}x_{1,t-2} \\ &\quad + b_{12}x_{2,t-2} + y_{1,t} \\ x_{2,t} &= a_{21}x_{1,t-1} + a_{22}x_{2,t-1} + b_{21}x_{1,t-2} \\ &\quad + b_{22}x_{2,t-2} + y_{2,t} \end{aligned}$$

can be transformed in a first-order system adding two variables:

$$\begin{aligned} x_{1,t} &= a_{11}x_{1,t-1} + a_{12}x_{2,t-1} + b_{11}x_{1,t-1} \\ &\quad + b_{12}x_{2,t-1} + y_{1,t} \\ x_{2,t} &= a_{21}x_{1,t-1} + a_{22}x_{2,t-1} + b_{21}x_{1,t-1} \\ &\quad + b_{22}x_{2,t-1} + y_{2,t} \\ z_{1,t} &= x_{1,t-1} \\ z_{2,t} &= x_{2,t-1} \end{aligned}$$

Transformations of this type can be generalized to systems of any order and any number of equations.

A system of difference equations is called *homogeneous* if the exogenous variable \mathbf{y}_t is zero, that is, if it can be written as

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1}$$

while it is called *nonhomogeneous* if the exogenous term is present.

There are different ways to solve first-order systems of difference equations. One method consists in eliminating variables as in ordinary algebraic systems. In this way, the original first-order system in k equations is solved by solving a single difference equation of order k with the methods explained above. This observation implies that solutions of systems of linear difference equations are of the same nature as those of difference equations (i.e., sums of exponential and/or sinusoidal functions). In the following section we will show a direct method for solving systems of linear difference equations. This method could be used to solve equations of any order, as they are equivalent to first-order

systems. In addition, it gives a better insight into vector autoregressive processes.

SYSTEMS OF HOMOGENEOUS LINEAR DIFFERENCE EQUATIONS

Consider a homogeneous system of the following type:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t - 1), \quad t = 0, 1, \dots, n, \dots$$

where \mathbf{A} is a $k \times k$, real-valued, nonsingular matrix of constant coefficients. Using the lag operator notation, we can also write the above systems in the following form:

$$(\mathbf{I} - \mathbf{A}L)\mathbf{x}_t = 0, \quad t = 1, \dots, n, \dots$$

If a vector of initial conditions $\mathbf{x}(0)$ is given, the above system is called an *initial value problem*.

Through recursive computation, that is, starting at $t = 0$ and computing forward, we can write

$$\begin{aligned} \mathbf{x}(1) &= \mathbf{A}\mathbf{x}(0) \\ \mathbf{x}(2) &= \mathbf{A}\mathbf{x}(1) = \mathbf{A}^2\mathbf{x}(0) \\ \mathbf{x}(t) &= \mathbf{A}^t\mathbf{x}(0) \end{aligned}$$

The following theorem can be demonstrated: Any homogeneous system of the type $\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t - 1)$, where \mathbf{A} is a $k \times k$, real-valued, nonsingular matrix, coupled with given initial conditions $\mathbf{x}(0)$ admits one and only one solution.

A set of k solutions $\mathbf{x}_i(t), i = 1, \dots, k, t = 0, 1, 2, \dots$ are said to be linearly independent if

$$\sum_{i=1}^k c_i \mathbf{x}_i(t) = 0$$

$t = 0, 1, 2, \dots$ implies $c_i = 0, i = 1, \dots, k$. Suppose now that k linearly independent solutions $\mathbf{x}_i(t), i = 1, \dots, k$ are given. Consider the matrix

$$\Phi(t) = [\mathbf{x}_1(t) \cdots \mathbf{x}_k(t)]$$

The following matrix equation is clearly satisfied:

$$\Phi(t) = \mathbf{A}\Phi(t - 1)$$

The solutions $\mathbf{x}_i(t), i = 1, \dots, n$ are linearly independent if and only if the matrix $\Phi(t)$ is

nonsingular for every value $t \geq 0$, that is, if $\det[\Phi(t)] \neq 0, t = 0, 1, \dots$. Any nonsingular matrix $\Phi(t), t = 0, 1, \dots$ such that the matrix equation

$$\Phi(t) = \mathbf{A}\Phi(t - 1)$$

is satisfied is called a *fundamental matrix* of the system $\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t - 1), t = 1, \dots, n, \dots$ and it satisfies the equation

$$\Phi(t) = \mathbf{A}^t\Phi(0)$$

In order to compute an explicit solution of this system, we need an efficient algorithm to compute the matrix sequence \mathbf{A}^t . We will discuss one algorithm for this computation.¹ Recall that an eigenvalue of the $k \times k$ real valued matrix $\mathbf{A} = (a_{ij})$ is a real or complex number λ that satisfies the matrix equation:

$$(\mathbf{A} - \lambda\mathbf{I})\xi = 0$$

where $\xi \in \mathbb{C}^k$ is a k -dimensional complex vector. The above equation has a nonzero solution if and only if

$$|(\mathbf{A} - \lambda\mathbf{I})| = 0$$

or

$$\det \begin{pmatrix} a_{11} - \lambda & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} - \lambda \end{pmatrix} = 0$$

The above condition can be expressed by the following algebraic equation:

$$z^k + a_1z^{k-1} + \cdots + a_{k-1}z + a_k$$

which is called the characteristic equation of the matrix $\mathbf{A} = (a_{ij})$.

To see the relationship of this equation with the characteristic equations of single equations, consider the k -order equation:

$$\begin{aligned} (1 - a_1L - \cdots - a_kL^k)x(t) &= 0 \\ x_t &= a_1x(t - 1) + \cdots + a_kx(t - k) \end{aligned}$$

which is equivalent to the first-order system,

$$\begin{aligned} x_t &= a_1x_{t-1} + \cdots + a_kz_{t-1}^{k-1} \\ z_t^1 &= x_{t-1} \\ &\vdots \\ z_{t-1}^{k-1} &= x_{t-k} \end{aligned}$$

The matrix

$$\mathbf{A} = \begin{bmatrix} a_1 & a_2 & \cdots & a_{k-1} & a_k \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

is called the *companion matrix*. By induction, it can be demonstrated that the characteristic equation of the system $\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t-1)$, $t = 1, \dots, n, \dots$ and of the k -order equation above coincide.

Given a system $\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t-1)$, $t = 1, \dots, n, \dots$, we now consider separately two cases: (1) All, possibly complex, eigenvalues of the real-valued matrix \mathbf{A} are distinct, and (2) two or more eigenvalues coincide.

Recall that if λ is a complex eigenvalue with corresponding complex eigenvector ξ , the complex conjugate number $\bar{\lambda}$ is also an eigenvalue with corresponding complex eigenvector $\bar{\xi}$.

If the eigenvalues of the real-valued matrix \mathbf{A} are all distinct, then the matrix can be diagonalized. This means that \mathbf{A} is similar to a diagonal matrix, according to the matrix equation

$$\mathbf{A} = \Xi \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \Xi^{-1}$$

$$\Xi = [\xi_1 \cdots \xi_n]$$

and

$$\mathbf{A}^t = \Xi \begin{bmatrix} \lambda_1^t & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n^t \end{bmatrix} \Xi^{-1}$$

We can therefore write the general solution of the system $\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t-1)$ as follows:

$$\mathbf{x}(t) = c_1 \lambda_1^t \xi_1 + \cdots + c_n \lambda_n^t \xi_n$$

The c_i are complex numbers that need to be determined for the solutions to be real and to satisfy initial conditions. We therefore see the parallel between the solutions of first-order systems of difference equations and the solutions of k -order difference equations that we have determined above. In particular, if solutions are all real they exhibit exponential decay if their modulus is less than 1 or exponential growth if their modulus is greater than 1. If the solutions of the characteristic equation are real, they can produce oscillating damped or undamped behavior with period equal to two time steps. If the solutions of the characteristic equation are complex, then solutions might exhibit damped or undamped oscillating behavior with any period.

To illustrate the above, consider the following second-order system:

$$x_{1,t} = 0.6x_{1,t-1} - 0.1x_{2,t-1} - 0.7x_{1,t-2} + 0.15x_{2,t-2}$$

$$x_{2,t} = -0.12x_{1,t-1} + 0.7x_{2,t-1} + 0.22x_{1,t-2} - 0.85x_{2,t-2}$$

This system can be transformed in the following first-order system:

$$x_{1,t} = 0.6x_{1,t-1} - 0.1x_{2,t-1} - 0.7x_{1,t-2} + 0.15x_{2,t-2}$$

$$x_{2,t} = -0.12x_{1,t-1} + 0.7x_{2,t-1} + 0.22x_{1,t-2} - 0.85x_{2,t-2}$$

$$z_{1,t} = x_{1,t-1}$$

$$z_{2,t} = x_{2,t-1}$$

with matrix

$$\mathbf{A} = \begin{bmatrix} 0.6 & -0.1 & -0.7 & 0.15 \\ -0.12 & 0.7 & 0.22 & -0.8 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

The eigenvalues of the matrix \mathbf{A} are distinct and complex:

$$\lambda_1 = 0.2654 + 0.7011i, \quad \lambda_2 = \bar{\lambda}_1 = 0.2654 - 0.7011i$$

$$\lambda_3 = 0.3846 + 0.8887i, \quad \lambda_4 = \bar{\lambda}_3 = 0.3846 - 0.8887i$$

The corresponding eigenvector matrix Ξ is

$$\Xi = \begin{bmatrix} 0.1571 + 0.4150i & 0.1571 - 0.4150i & -0.1311 - 0.3436i & -0.1311 + 0.3436i \\ -0.0924 + 0.3928i & 0.0924 - 0.3928i & 0.2346 + 0.5419i & 0.2346 - 0.5419i \\ 0.5920 & 0.5920 & -0.3794 - 0.0167i & -0.3794 + 0.0167i \\ 0.5337 + 0.0702i & 0.5337 - 0.0702i & 0.6098 & 0.6098 \end{bmatrix}$$

Each column of the matrix is an eigenvector.
The solution of the system is given by

$$\begin{aligned}
 \mathbf{x}(t) &= c_1 \lambda_1^t \xi_1 + c_2 \overline{\lambda_1^t} \overline{\xi_1} + c_3 \lambda_3^t \xi_3 + c_4 \overline{\lambda_3^t} \overline{\xi_3} \\
 &= c_1 (0.2654 + 0.7011i)^t \begin{pmatrix} 0.1571 + 0.4150i \\ 0.0924 + 0.3928i \\ 0.5920 \\ 0.5337 + 0.0702i \end{pmatrix} \xi_1 \\
 &+ c_2 (0.2654 - 0.7011i)^t \begin{pmatrix} 0.1571 - 0.4150i \\ 0.0924 - 0.3928i \\ 0.5920 \\ 0.5337 - 0.0702i \end{pmatrix} \\
 &+ c_3 (0.3846 + 0.8887i)^t \begin{pmatrix} -0.1311 + 0.3436i \\ 0.2346 + 0.5419i \\ -0.3794 + 0.0167i \\ 0.6098 \end{pmatrix} \xi_3 \\
 &+ c_4 (0.3846 - 0.8887i)^t \begin{pmatrix} -0.1311 - 0.3436i \\ 0.2346 - 0.5419i \\ -0.3794 - 0.0167i \\ 0.6098 \end{pmatrix}
 \end{aligned}$$

The four constants c can be determined using the initial conditions: $x(1) = 1; x(2) = 1.2;$

$y(1) = 1.5; y(2) = -2.$ Figure 9 illustrates the behavior of solutions.

Now consider the case in which two or more solutions of the characteristic equation are coincident. In this case, it can be demonstrated that the matrix \mathbf{A} can be diagonalized only if it is normal, that is if

$$\mathbf{A}^T \mathbf{A} = \mathbf{A} \mathbf{A}^T$$

If the matrix \mathbf{A} is not normal, it cannot be diagonalized. However, it can be put in *Jordan canonical form*. In fact, it can be demonstrated that any nonsingular real-valued matrix \mathbf{A} is similar to a matrix in Jordan canonical form,

$$\mathbf{A} = \mathbf{P} \mathbf{J} \mathbf{P}^{-1}$$

where the matrix \mathbf{J} has the form $\mathbf{J} = \text{diag}[\mathbf{J}_1, \dots, \mathbf{J}_k]$, that is, it is formed by *Jordan diagonal blocks*:

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{J}_k \end{bmatrix}$$

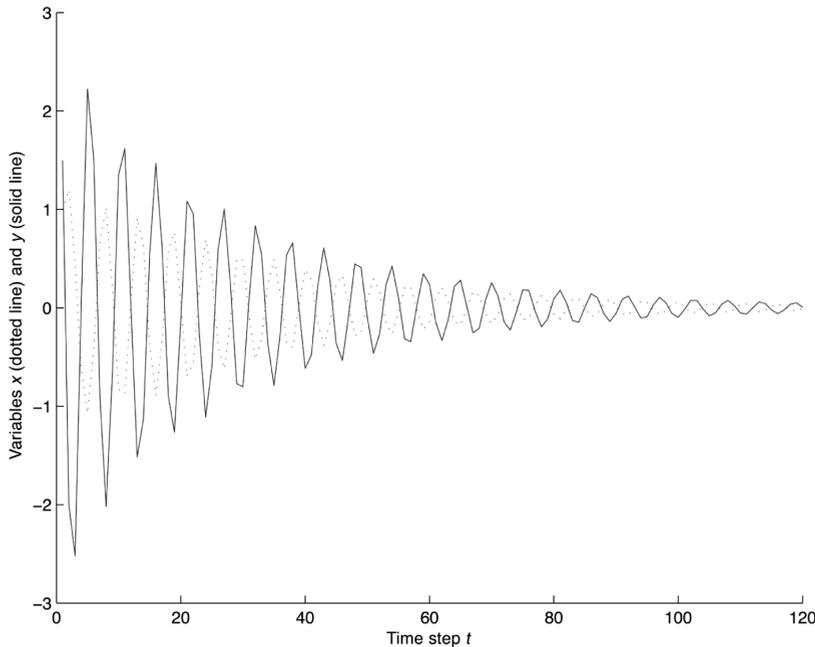


Figure 9 Solution of the System

$$\begin{aligned}
 x_{1,t} &= 0.6x_{1,t-1} - 0.1x_{2,t-1} - 0.7x_{1,t-2} + 0.15x_{2,t-2} \\
 x_{2,t} &= -0.12x_{1,t-1} + 0.7x_{2,t-1} + 0.22x_{1,t-2} - 0.85x_{2,t-2}
 \end{aligned}$$

where each Jordan block has the form

$$J_i = \begin{bmatrix} \lambda_1 & 1 & \cdots & 0 \\ 0 & \lambda_i & \cdots & \vdots \\ \vdots & \vdots & \ddots & 1 \\ 0 & 0 & \cdots & \lambda_i \end{bmatrix}$$

The Jordan canonical form is characterized by two sets of multiplicity parameters, the algebraic multiplicity and the geometric multiplicity. The geometric multiplicity of an eigenvalue is the number of Jordan blocks corresponding to that eigenvalue, while the algebraic multiplicity of an eigenvalue is the number of times the eigenvalue is repeated. An eigenvalue that is repeated s times can have from 1 to s Jordan blocks. For example, suppose a matrix has only one eigenvalue $\lambda = 5$ that is repeated three times. There are four possible matrices with the following Jordan representation:

$$\begin{pmatrix} 5 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{pmatrix}, \begin{pmatrix} 5 & 1 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 5 \end{pmatrix}, \begin{pmatrix} 5 & 0 & 0 \\ 0 & 5 & 1 \\ 0 & 0 & 5 \end{pmatrix}, \begin{pmatrix} 5 & 1 & 0 \\ 0 & 5 & 1 \\ 0 & 0 & 5 \end{pmatrix}$$

These four matrices have all algebraic multiplicity 3 but geometric multiplicity from left to right 1, 2, 2, 3, respectively.

KEY POINTS

- Homogeneous difference equations are linear conditions that link the values of variables at different time lags.
- In the case of real roots, solutions are sums of exponentials. Any linear combination of solu-

tions of the homogeneous difference equation is another solution.

- When some of the roots are complex, the solutions of a homogeneous difference equation exhibit an oscillating behavior with a period that depends on the model coefficients.
- The general solution of a homogeneous difference equation that admits both real and complex roots with different multiplicities is a sum of the different types of solutions.
- A system of difference equations is called homogeneous if the system's exogenous variable is zero, and nonhomogeneous if the exogenous term is present.
- One method of solving first-order systems of difference equations is by eliminating variables as in ordinary algebraic systems; another way is a direct method that can be used to solve systems of linear difference equations of any order.

NOTE

1. This discussion of systems of difference equations draws on Elaydi (2002).

REFERENCES

- Elaydi, S. (2002). *An Introduction to Difference Equations*. New York: Springer Verlag.
- Goldberg, S. (2010). *Introduction to Difference Equations*. New York: Dover Publications.
- Kelley, W. G., and Peterson, A. C. (1991). *Difference Equations: An Introduction with Applications*, 2nd ed. San Diego, CA: Academic Press.

Differential Equations

SERGIO M. FOCARDI, PhD

Partner, The Intertek Group

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: In financial modeling, the goal is to be able to represent the problem at hand as a mathematical function. In a mathematical function, the dependent variable depends on one or more variables that are referred to as independent variables. In standard calculus, there are two basic operations with mathematical functions: differentiation and integration. The differentiation operation leads to derivatives. When a mathematical function has only one independent variable, then the derivative is referred to as an ordinary derivative. Typically in financial applications, the independent variable is time. The derivative of a mathematical function that has more than one independent variable (one of which is typically time) is called a partial derivative. A differential equation is an equation that contains derivatives. When it contains only an ordinary derivative, it is referred to as an ordinary differential equation; when the differential equation contains partial derivatives, the differential equation is called a partial differential equation.

In nontechnical terms, *differential equations* are equations that express a relationship between a function and one or more *derivatives* (or differentials) of that function. The highest order of derivatives included in a differential equation is referred to as its *order*. In financial modeling, differential equations are used to specify the laws governing the evolution of price distributions, deriving solutions to simple and complex options, and estimating term structure models. In most applications in finance, only first- and second-order differential equations are found.

Differential equations are classified as ordinary differential equations and partial differential equations depending on the type of

derivatives included in the differential equation. When there is only an ordinary derivative (i.e., a derivative of a mathematical function with only one independent variable), the differential equation is called an *ordinary differential equation*. For differential equations where there are partial derivatives (i.e., a derivative of a mathematical function with more than one independent variable), then the differential equation is called a *partial differential equation*. Typically in differential equations, one of the independent variables is time. A differential equation may have a derivative of a mathematical function where one or more of the independent variables is a random variable or a

stochastic process. In such instances, the differential equation is referred to as a *stochastic differential equation*.

The solutions to a differential equation or system of differential equations can be as simple as explicit formulas. When an explicit formula is not possible to obtain, various numerical methods can be used to approximate a solution. Even in the absence of an exact solution, properties of solutions of a differential equation can be determined. A large number of properties of differential equations have been established over the last three centuries. This entry provides only a brief introduction to the concept of differential equations and their properties, limiting our discussion to the principal concepts. We do not cover stochastic differential equations.

DIFFERENTIAL EQUATIONS DEFINED

A differential equation is a condition expressed as a functional link between one or more functions and their derivatives. It is expressed as an equation (that is, as an equality between two terms).

A solution of a differential equation is a function that satisfies the given condition. For example, the condition

$$Y''(x) + \alpha Y'(x) + \beta Y(x) - b(x) = 0$$

equates to zero a linear relationship between an unknown function $Y(x)$, its first and second derivatives $Y'(x), Y''(x)$, and a known function $b(x)$. (In some equations we will denote the first and second derivatives by a single and double prime, respectively.) The unknown function $Y(x)$ is the solution of the equation that is to be determined.

There are two broad types of differential equations: ordinary differential equations and partial differential equations. Ordinary differential equations are equations or systems of equations involving only one independent variable. Another way of saying this is that ordinary

differential equations involve only total derivatives. In contrast, partial differential equations are differential equations or systems of equations involving partial derivatives. That is, there is more than one independent variable.

ORDINARY DIFFERENTIAL EQUATIONS

In full generality, an ordinary differential equation (ODE) can be expressed as the following relationship:

$$F[x, Y(x), Y^1(x), \dots, Y^{(n)}(x)] = 0$$

where $Y^{(m)}(x)$ denotes the m -th derivative of an unknown function $Y(x)$. If the equation can be solved for the n -th derivative, it can be put in the form:

$$Y^{(n)}(x) = G[x, Y(x), Y^{(1)}(x), \dots, Y^{(n-1)}(x)]$$

Order and Degree of an ODE

A differential equation is classified in terms of its order and its degree. The order of a differential equation is the order of the highest derivative in the equation. For example, the above differential equation is of order n since the highest order derivative is $Y^{(n)}(x)$. The degree of a differential equation is determined by looking at the highest derivative in the differential equation. The degree is the power to which that derivative is raised.

For example, the following ordinary differential equations are first-degree differential equations of different orders:

$$Y^{(1)}(x) - 10Y(x) + 40 = 0 \quad (\text{order } 1)$$

$$4Y^{(3)}(x) + Y^{(2)}(x) + Y^{(1)}(x) - 0.5Y(x) + 100 = 0 \quad (\text{order } 3)$$

The following ordinary differential equations are of order 3 and fifth degree:

$$4[Y^{(3)}(x)]^5 + [Y^{(2)}(x)]^2 + Y^{(1)}(x) - 0.5Y(x) + 100 = 0$$

$$4[Y^{(3)}(x)]^5 + [Y^{(2)}(x)]^3 + Y^{(1)}(x) - 0.5Y(x) + 100 = 0$$

When an ordinary differential equation is of the first degree, it is said to be a *linear ordinary differential equation*.

Solution to an ODE

Let's return to the general ODE. A solution of this equation is any function $y(x)$ such that:

$$F[x, y(x), y^{(1)}(x), \dots, y^{(n)}(x)] = 0$$

In general there will be not one but an infinite family of solutions. For example, the equation

$$Y^{(1)}(x) = \alpha Y(x)$$

admits, as a solution, all the functions of the form

$$y(x) = C \exp(\alpha x)$$

To identify one specific solution among the possible infinite solutions that satisfy a differential equation, additional restrictions must be imposed. Restrictions that uniquely identify a solution to a differential equation can be of various types. For instance, one could impose that a solution of an n -th order differential equation passes through n given points. A common type of restriction—called an *initial condition*—is obtained by imposing that the solution and some of its derivatives assume given initial values at some initial point.

Given an ODE of order n , to ensure the uniqueness of solutions it will generally be necessary to specify a starting point and the initial value of $n-1$ derivatives. It can be demonstrated, given the differential equation

$$F[x, Y(x), Y^{(1)}(x), \dots, Y^{(n)}(x)] = 0$$

that if the function F is continuous and all of its partial derivatives up to order n are continuous in some region containing the values $y_0, \dots, y_0^{(n-1)}$, then there is a unique solution $y(x)$ of the equation in some interval $I = (M \leq x \leq L)$ such that $y_0 = Y(x_0), \dots, y_0^{(n-1)} = Y^{(n-1)}(x_0)$.¹ Note that this theorem states that there is an interval in which the solution exists. Existence and uniqueness of solutions in a given interval

is a more delicate matter and must be examined for different classes of equations.

The *general solution* of a differential equation of order n is a function of the form

$$y = \varphi(x, C_1, \dots, C_n)$$

that satisfies the following two conditions:

- *Condition 1.* The function $y = \varphi(x, C_1, \dots, C_n)$ satisfies the differential equation for any n -tuple of values (C_1, \dots, C_n) .
- *Condition 2.* Given a set of initial conditions $y(x_0) = y_0, \dots, y^{(n-1)}(x_0) = y_0^{(n-1)}$ that belong to the region where solutions of the equation exist, it is possible to determine n constants in such a way that the function $y = \varphi(x, C_1, \dots, C_n)$ satisfies these conditions.

The coupling of differential equations with initial conditions embodies the notion of universal determinism of classical physics. Given initial conditions, the future evolution of a system that obeys those equations is completely determined. This notion was forcefully expressed by Pierre-Simon Laplace in the eighteenth century: A supernatural mind who knows the laws of physics and the initial conditions of each atom could perfectly predict the future evolution of the universe with unlimited precision.

In the twentieth century, the notion of universal determinism was challenged twice in the physical sciences. First in the 1920s the development of quantum mechanics introduced the so-called indeterminacy principle which established explicit bounds to the precision of measurements. Later, in the 1970s, the development of *nonlinear dynamics* and chaos theory showed how arbitrarily small initial differences might become arbitrarily large: The flapping of a butterfly's wings in the southern hemisphere might cause a tornado in the northern hemisphere.

SYSTEMS OF ORDINARY DIFFERENTIAL EQUATIONS

Differential equations can be combined to form systems of differential equations. These are sets

of differential conditions that must be satisfied simultaneously. A *first-order system of differential equations* is a system of the following type:

$$\begin{cases} \frac{dy_1}{dx} = f_1(x, y_1, \dots, y_n) \\ \frac{dy_2}{dx} = f_2(x, y_1, \dots, y_n) \\ \cdot \\ \cdot \\ \frac{dy_n}{dx} = f_n(x, y_1, \dots, y_n) \end{cases}$$

Solving this system means finding a set of functions y_1, \dots, y_n that satisfy the system as well as the initial conditions:

$$y_1(x_0) = y_{10}, \dots, y_n(x_0) = y_{n0}$$

Systems of orders higher than one can be reduced to first-order systems in a straightforward way by adding new variables defined as the derivatives of existing variables. As a consequence, an n -th order differential equation can be transformed into a first-order system of n equations. Conversely, a system of first-order differential equations is equivalent to a single n -th order equation.

To illustrate this point, let's differentiate the first equation to obtain

$$\frac{d^2 y_1}{dx^2} = \frac{\partial f_1}{\partial x} + \frac{\partial f_1}{\partial y_1} \frac{dy_1}{dx} + \dots + \frac{\partial f_1}{\partial y_n} \frac{dy_n}{dx}$$

Replacing the derivatives

$$\frac{dy_1}{dx}, \dots, \frac{dy_n}{dx}$$

with their expressions f_1, \dots, f_n from the system's equations, we obtain

$$\frac{d^2 y_1}{dx^2} = F_2(x, y_1, \dots, y_n)$$

If we now reiterate this process, we arrive at the n -th order equation:

$$\frac{d^{(n)} y_1}{dx^{(n)}} = F_n(x, y_1, \dots, y_n)$$

We can thus write the following system:

$$\begin{cases} \frac{dy_1}{dx} = f_1(x, y_1, \dots, y_n) \\ \frac{d^2 y_1}{dx^2} = F_2(x, y_1, \dots, y_n) \\ \cdot \\ \cdot \\ \frac{d^{(n)} y_1}{dx^{(n)}} = F_n(x, y_1, \dots, y_n) \end{cases}$$

We can express y_2, \dots, y_n as functions of $x, y_1, y_1', \dots, y_1^{(n-1)}$ by solving, if possible, the system formed with the first $n-1$ equations:

$$\begin{cases} y_2 = \varphi_2(x, y_1, y_1', \dots, y_1^{(n-1)}) \\ y_3 = \varphi_3(x, y_1, y_1', \dots, y_1^{(n-1)}) \\ \cdot \\ \cdot \\ y_n = \varphi_n(x, y_1, y_1', \dots, y_1^{(n-1)}) \end{cases}$$

Substituting these expressions into the n -th equation of the previous system, we arrive at the single equation:

$$\frac{d^{(n)} y_1}{dx^{(n)}} = \Phi(x, y_1', \dots, y_1^{(n-1)})$$

Solving, if possible, this equation, we find the general solution

$$y_1 = y_1(x, C_1, \dots, C_n)$$

Substituting this expression for y_1 into the previous system, y_2, \dots, y_n can be computed.

CLOSED-FORM SOLUTIONS OF ORDINARY DIFFERENTIAL EQUATIONS

Let's now consider the methods for solving two types of common differential equations: equations with separable variables and equations of linear type. Let's start with equations with separable variables. Consider the equation

$$\frac{dy}{dx} = f(x)g(y)$$

This equation is said to have separable variables because it can be written as an equality between two sides, each depending on only y or only x . We can rewrite our equation in the following way:

$$\frac{dy}{g(y)} = f(x)dx$$

This equation can be regarded as an equality between two differentials in y and x respectively. Their indefinite integrals can differ only by a constant. Integrating the left side with respect to y and the right side with respect to x , we obtain the general solution of the equation:

$$\int \frac{dy}{g(y)} = \int f(x)dx + C$$

For example, if $g(y) \equiv y$, the previous equation becomes

$$\frac{dy}{y} = f(x)dx$$

whose solution is

$$\int \frac{dy}{y} = \int f(x)dx + C \Rightarrow \log y = \int f(x)dx + C \Rightarrow y = A \exp\left(\int f(x)dx\right)$$

where $A = \exp(C)$.

A differential equation of this type describes the continuous compounding of time-varying interest rates. Consider, for example, the growth of capital C deposited in a bank account that earns the variable but deterministic rate $r = f(t)$. When interest rates R_i are constant for discrete periods of time Δt_i , compounding is obtained by purely algebraic formulas as follows:

$$R_i \Delta t_i = \frac{C(t_i) - C(t_{i-\Delta t_i})}{C(t_{i-\Delta t_i})}$$

Solving for $C(t_i)$:

$$C(t_i) = (1 + R_i \Delta t_i)C(t_{i-\Delta t_i})$$

By recursive substitution we obtain

$$C(t_i) = (1 + R_i \Delta t_i)(1 + R_{i-1} \Delta t_{i-1}) \dots (1 + R_1 \Delta t_1)C(t_0)$$

However, market interest rates are subject to rapid change. In the limit of very short time intervals, the instantaneous rate $r(t)$ would be defined as the limit, if it exists, of the discrete interest rate:

$$r(t) = \lim_{\Delta t \rightarrow 0} \frac{C(t + \Delta t) - C(t)}{\Delta t C(t)}$$

The above expression can be rewritten as a simple first-order differential equation in C :

$$r(t)C(t) = \frac{dC(t)}{dt}$$

In a simple intuitive way, the above equation can be obtained considering that in the elementary time dt the bank account increments by the amount $dC = C(t)r(t)dt$. In this equation, variables are separable. It admits the family of solutions:

$$C = A \exp\left(\int r(t)dt\right)$$

where A is the initial capital.

Linear Differential Equation

Linear differential equations are equations of the following type:

$$a_n(x)y^{(n)} + a_{n-1}(x)y^{(n-1)} + \dots + a_1(x)y^{(1)} + a_0(x)y + b(x) = 0$$

If the function b is identically zero, the equation is said to be homogeneous.

In cases where the coefficients a 's are constant, *Laplace transforms* provide a powerful method for solving linear differential equations. (Laplace transforms are one of two popular integral transforms—the other being Fourier transforms—used in financial modeling. Integral transforms are operations that take any function into another function of a different variable through an improper integral.) Consider, without loss of generality, the following linear equation with constant coefficients:

$$a_n y^{(n)} + a_{n-1} y^{(n-1)} + \dots + a_1 y^{(1)} + a_0 y = b(x)$$

together with the initial conditions: $y(0) = y_0, \dots, y^{(n-1)}(0) = y_0^{(n-1)}$. In cases in which the initial point is not the origin, by a variable transformation we can shift the origin.

Laplace Transform

For one-sided Laplace transforms the following formulas hold:

$$\begin{aligned}\mathcal{L}\left(\frac{df(x)}{dx}\right) &= s\mathcal{L}[f(x)] - f(0) \\ \mathcal{L}\left(\frac{d^n f(x)}{dx^n}\right) &= s^n \mathcal{L}[f(x)] - s^{n-1} f(0) - \dots \\ &\quad - f^{(n-1)}(0)\end{aligned}$$

Suppose that a function $y = y(x)$ satisfies the previous linear equation with constant coefficients and that it admits a Laplace transform. Apply one-sided Laplace transform to both sides of the equation. If $Y(s) = \mathcal{L}[y(x)]$, the following relationships hold:

$$\begin{aligned}L(a_n y^{(n)} + a_{n-1} y^{(n-1)} + \dots + a_1 y^{(1)} + a_0 y) \\ &= L[b(x)] \\ a_n [s^n Y(s) - s^{n-1} y^{(1)}(0) - \dots - y^{(n-1)}(0)] \\ &\quad + a_{n-1} [s^{n-1} Y(s) - s^{n-2} y^{(1)}(0) - \dots - y^{(n-2)}(0)] \\ &\quad + \dots + a_0 Y(s) = B(s)\end{aligned}$$

Solving this equation for $Y(s)$, that is, $Y(s) = g[s, y^{(1)}(0), \dots, y^{(n-1)}(0)]$ the inverse Laplace transform $y(t) = \mathcal{L}^{-1}[Y(s)]$ uniquely determines the solution of the equation.

Because inverse Laplace transforms are integrals, with this method, when applicable, the solution of a differential equation is reduced to the determination of integrals. Laplace transforms and inverse Laplace transforms are known for large classes of functions. Because of the important role that Laplace transforms play in solving ordinary differential equations in engineering problems, there are published reference tables. Laplace transform methods also yield closed-form solutions of many ordinary differential equations of interest in economics and finance.

NUMERICAL SOLUTIONS OF ORDINARY DIFFERENTIAL EQUATIONS

Closed-form solutions are solutions that can be expressed in terms of known functions such as polynomials or exponential functions. Before the advent of fast digital computers, the search for closed-form solutions of differential equations was an important task. Today, thanks to the availability of high-performance computing, most problems are solved numerically. This section looks at methods for solving ordinary differential equations numerically.

The Finite Difference Method

Among the methods used to numerically solve ordinary differential equations subject to initial conditions, the most common is the finite difference method. The finite difference method is based on replacing derivatives with difference equations; differential equations are thereby transformed into recursive difference equations.

Key to this method of numerical solution is the fact that ODEs subject to initial conditions describe phenomena that evolve from some starting point. In this case, the differential equation can be approximated with a system of difference equations that compute the next point based on previous points. This would not be possible should we impose boundary conditions instead of initial conditions. In this latter case, we have to solve a system of linear equations.

To illustrate the finite difference method, consider the following simple ordinary differential equation and its solution in a finite interval:

$$\begin{aligned}f'(x) &= f(x) \\ \frac{df}{f} &= dx \\ \log f(x) &= x + C \\ f(x) &= \exp(x + C)\end{aligned}$$

As shown, the closed-form solution of the equation is obtained by separation of variables, that

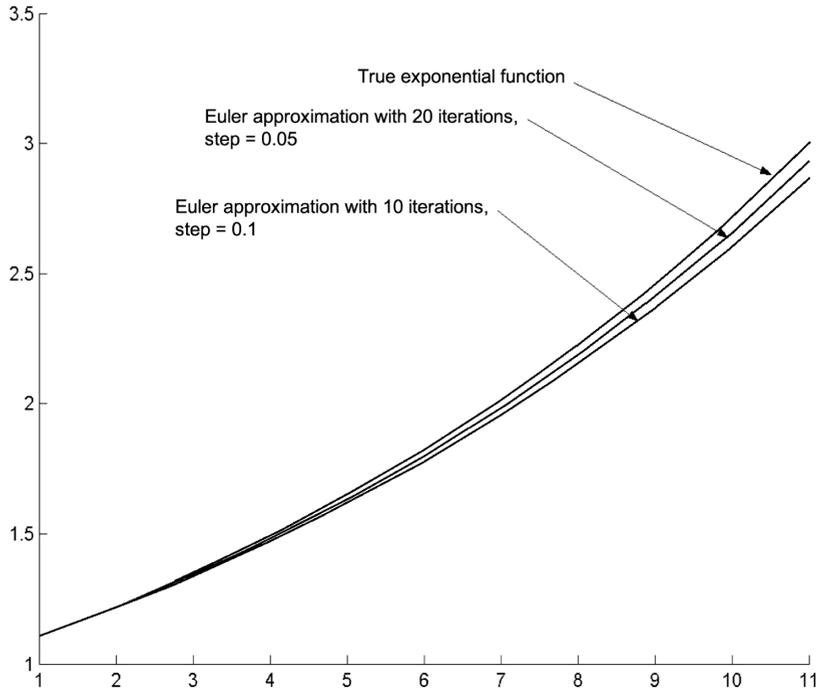


Figure 1 Numerical Solutions of the Equation $f' = f$ with the Euler Approximation for Different Step Sizes

is, by transforming the original equation into another equation where the function f appears only on the left side and the variable x only on the right side.

Suppose that we replace the derivative with its forward finite difference approximation and solve

$$\frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = f(x_i)$$

$$f(x_{i+1}) = [1 + (x_{i+1} - x_i)]f(x_i)$$

If we assume that the step size is constant for all i :

$$f(x_i) = [1 + \Delta x]^i f(x_0)$$

The replacement of derivatives with finite differences is often called the Euler approximation. The differential equation is replaced by a recursive formula based on approximating the derivative with a finite difference. The i -th value of the solution is computed from the $i-1$ -th value. Given the initial value of the func-

tion f , the solution of the differential equation can be arbitrarily approximated by choosing a sufficiently small interval. Figure 1 illustrates this computation for different values of Δx .

In the previous example of a first-order linear equation, only one initial condition was involved. Let's now consider a second-order equation:

$$f''(x) = kf(x) = 0$$

This equation describes oscillatory motion, such as the elongation of a pendulum or the displacement of a spring.

To approximate this equation we must approximate the second derivative. This could be done, for example, by combining difference quotients as follows:

$$f'(x) \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

$$f'(x + \Delta x) \approx \frac{f(x + 2\Delta x) - f(x + \Delta x)}{\Delta x}$$

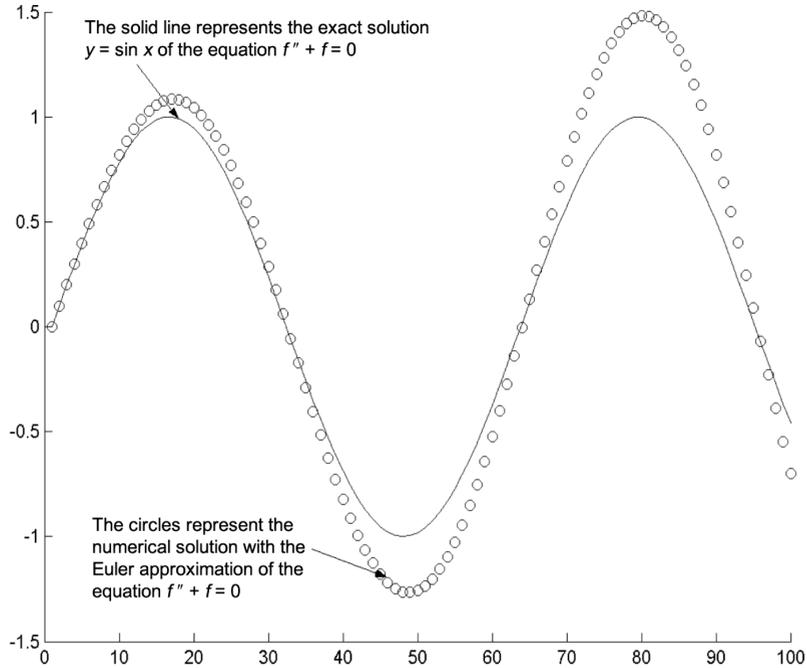


Figure 2 Numerical Solution of the Equation $f'' + f = 0$ with the Euler Approximation

$$\begin{aligned} f''(x) &\approx \frac{f'(x + \Delta x) - f'(x)}{\Delta x} \\ &= \frac{\frac{f(x + 2\Delta x) - f(x + \Delta x)}{\Delta x} - \frac{f(x + \Delta x) - f(x)}{\Delta x}}{\Delta x} \\ &= \frac{f(x + 2\Delta x) - 2f(x + \Delta x) + f(x)}{(\Delta x)^2} \end{aligned}$$

With this approximation, the original equation becomes

$$\begin{aligned} f''(x) + kf(x) &\approx \frac{f(x + 2\Delta x) - 2f(x + \Delta x) + f(x)}{(\Delta x)^2} + kf(x) = 0 \\ f(x + 2\Delta x) - 2f(x + \Delta x) + (1 + k(\Delta x)^2)f(x) &= 0 \end{aligned}$$

We can thus write the approximation scheme:

$$\begin{aligned} f(x + \Delta x) &= f(x) + \Delta x f'(x) \\ f(x + 2\Delta x) &= 2f(x + \Delta x) - (1 + k(\Delta x)^2)f(x) \end{aligned}$$

Given the increment Δx and the initial values $f(0), f'(0)$, using the above formulas we can recursively compute $f(0 + \Delta x), f(0 + 2\Delta x)$, and so on. Figure 2 illustrates this computation.

In practice, the Euler approximation scheme is often not sufficiently precise and more sophisticated approximation schemes are used. For example, a widely used approximation scheme is the *Runge-Kutta method*. We give an example of the Runge-Kutta method in the case of the equation $f'' + f = 0$ which is equivalent to the linear system:

$$\begin{aligned} x' &= y \\ y' &= -x \end{aligned}$$

In this case the Runge-Kutta approximation scheme is the following:

$$\begin{aligned} k_1 &= hy(i) \\ h_1 &= -hx(i) \\ k_2 &= h \left[y(i) + \frac{1}{2}h_1 \right] \\ h_2 &= -h \left[x(i) + \frac{1}{2}k_1 \right] \\ k_3 &= h \left[y(i) + \frac{1}{2}h_2 \right] \\ h_3 &= -h \left[x(i) + \frac{1}{2}k_2 \right] \end{aligned}$$

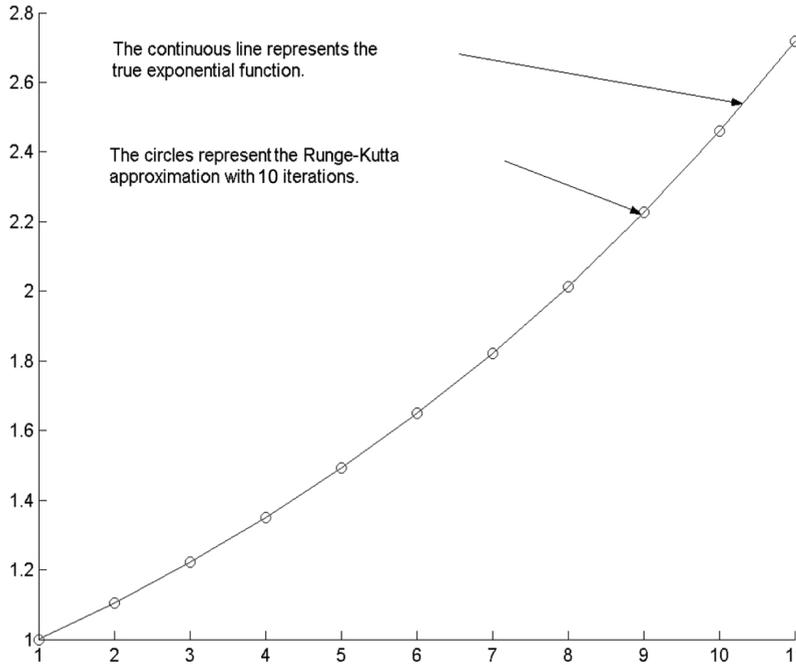


Figure 3 Numerical Solution of the Equation $f' = f$ with the Runge-Kutta Method After 10 Steps

$$\begin{aligned}
 k_4 &= h[y(i) + h_3] \\
 h_4 &= -h[x(i) + k_3] \\
 x(i + 1) &= x(i) + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\
 y(i + 1) &= y(i) + \frac{1}{6}(h_1 + 2h_2 + 2h_3 + h_4)
 \end{aligned}$$

Figures 3 and 4 illustrate the results of this method in the two cases $f' = f$ and $f'' + f = 0$.

As mentioned above, this numerical method depends critically on our having as givens (1) the initial values of the solution, and (2) its first derivative. Suppose that instead of initial values two boundary values were given, for instance the initial value of the solution and its value 1,000 steps ahead, that is, $f(0) = f_0$, $f(0 + 1,000\Delta x) = f_{1000}$. Conditions like these are rarely used in the study of *dynamical systems* as they imply foresight, that is, knowledge of the future position of a system. However, they often appear in static systems and when trying to determine what initial conditions should be imposed to reach a given goal at a given date.

In the case of boundary conditions, one cannot write a direct recursive scheme; it's neces-

sary to solve a system of equations. For instance, we could introduce the derivative $f'(x) = \delta$ as an unknown quantity. The difference quotient that approximates the derivative becomes an unknown. We can now write a system of linear equations in the following way:

$$\begin{cases}
 f(\Delta x) = f_0 + \delta \Delta x \\
 f(2\Delta x) = 2f(\Delta x) - (1 + k(\Delta x)^2)f_0 \\
 f(3\Delta x) = 2f(2\Delta x) - (1 + k(\Delta x)^2)f(\Delta x) \\
 \vdots \\
 \vdots \\
 f_{1000} = 2f(999\Delta x) - (1 + k(\Delta x)^2)f(998\Delta x)
 \end{cases}$$

This is a system of 1,000 equations in 1,000 unknowns. Solving the system we compute the entire solution. In this system two equations, the first and the last, are linked to boundary values; all other equations are transfer equations that express the dynamics (or the law) of the system. This is a general feature of boundary value problems. We will encounter it again when discussing numerical solutions of partial differential equations.

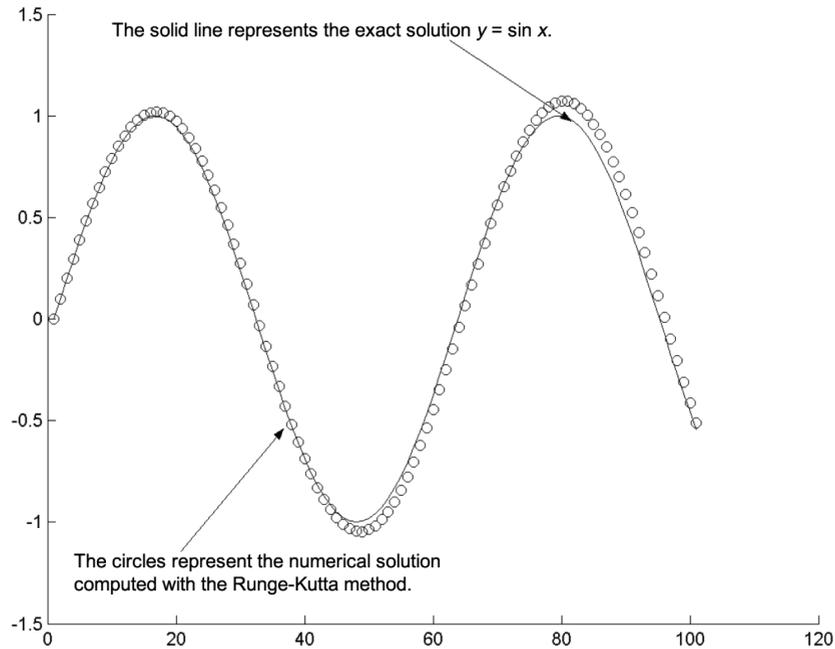


Figure 4 Numerical Solution of the Equation $f'' + f = 0$ with the Runge-Kutta Method

In the above example, we chose a forward scheme where the derivative is approximated with the forward difference quotient. One might use a different approximation scheme, computing the derivative in intervals centered around the point x . When derivatives of higher orders are involved, the choice of the approximation scheme becomes critical. Recall that when we approximated first and second derivatives using forward differences, we were required to evaluate the function at two points $(i, i + 1)$ and three points $(i, i + 1, i + 2)$ ahead respectively. If purely forward schemes are employed, computing higher-order derivatives requires many steps ahead. This fact might affect the precision and stability of numerical computations.

We saw in the examples that the accuracy of a finite difference scheme depends on the discretization interval. In general, a finite difference scheme works, that is, it is consistent and stable, if the numerical solution converges uniformly to the exact solution when the length of the discretization interval tends to zero. Suppose that the precision of an approximation

scheme depends on the length of the discretization interval Δx . Consider the difference $\delta f = \hat{f}(x) - f(x)$ between the approximate and the exact solutions. We say that $\delta f \rightarrow 0$ uniformly in the interval $[a, b]$ when $\Delta x \rightarrow 0$ if, given any ε arbitrarily small, it is possible to find a Δx such that $|\delta f| < \varepsilon$, $\forall x \in [a, b]$.

NONLINEAR DYNAMICS AND CHAOS

Systems of differential equations describe dynamical systems that evolve starting from initial conditions. A fundamental concept in the theory of dynamical systems is that of the *stability of solutions*. This topic has become of paramount importance with the development of nonlinear dynamics and with the discovery of chaotic phenomena. We can only give a brief introductory account of this subject whose role in economics is still the subject of debate.

Intuitively, a dynamical system is considered stable if its solutions do not change much when the system is only slightly perturbed.

There are different ways to perturb a system: changing parameters in its equations, changing the known functions of the system by a small amount, or changing the initial conditions.

Consider an equilibrium solution of a dynamical system, that is, a solution that is time invariant. If a stable system is perturbed when it is in a position of equilibrium, it tends to return to the equilibrium position or, in any case, not to diverge indefinitely from its equilibrium position. For example, a damped pendulum—if perturbed from a position of equilibrium—will tend to go back to an equilibrium position. If the pendulum is not damped it will continue to oscillate forever.

Consider a system of n equations of first order. (As noted above, systems of higher orders can always be reduced to first-order systems by enlarging the set of variables.) Suppose that we can write the system explicitly in the first derivatives as follows:

$$\left\{ \begin{array}{l} \frac{dy_1}{dx} = f_1(x, y_1, \dots, y_n) \\ \frac{dy_2}{dx} = f_2(x, y_1, \dots, y_n) \\ \cdot \\ \cdot \\ \cdot \\ \frac{dy_n}{dx} = f_n(x, y_1, \dots, y_n) \end{array} \right.$$

If the equations are all linear, a complete theory of stability has been developed. Essentially, linear dynamical systems are stable except possibly at singular points where solutions might diverge. In particular, a characteristic of linear systems is that they incur only small changes in the solution as a result of small changes in the initial conditions.

However, during the 1970s, it was discovered that nonlinear systems have a different behavior. Suppose that a nonlinear system has at least three degrees of freedom (that is, it has three independent nonlinear equations). The dynamics of such a system can then become chaotic in the sense that arbitrarily small changes in initial conditions might diverge. This sensitivity

to initial conditions is one of the signatures of *chaos*. Note that while discrete systems such as discrete maps can exhibit chaos in one dimension, continuous systems require at least three degrees of freedom (that is, three equations).

Sensitive dependence from initial conditions was first observed in 1960 by the meteorologist Edward Lorenz of the Massachusetts Institute of Technology. Lorenz remarked that computer simulations of weather forecasts starting, apparently, from the same meteorological data could yield very different results. He argued that the numerical solutions of extremely sensitive differential equations such as those he was using produced diverging results due to rounding-off errors made by the computer system. His discovery was published in a meteorological journal where it remained unnoticed for many years.

Fractals

While in principle deterministic chaotic systems are unpredictable because of their sensitivity to initial conditions, the statistics of their behavior can be studied. Consider, for example, the chaos laws that describe the evolution of weather: While the weather is basically unpredictable over long periods of time, long-run simulations are used to predict the statistics of weather.

It was discovered that probability distributions originating from chaotic systems exhibit *fat tails* in the sense that very large, extreme events have nonnegligible probabilities. (See Brock, Hsieh, and LeBaron [1991] and Hsieh [1991].) It was also discovered that chaotic systems exhibit complex unexpected behavior. The motion of chaotic systems is often associated with self-similarity and fractal shapes.

Fractals were introduced in the 1960s by Benoit Mandelbrot, a mathematician working at the IBM research center in Yorktown Heights, New York. Starting from the empirical observation that cotton price time-series are similar at different time scales, Mandelbrot developed a powerful theory of fractal geometrical objects.

Fractals are geometrical objects that are geometrically similar to part of themselves. Stock prices exhibit this property insofar as price time-series look the same at different time scales.

Chaotic systems are also sensitive to changes in their parameters. In a chaotic system, only some regions of the parameter space exhibit chaotic behavior. The change in behavior is abrupt and, in general, it cannot be predicted analytically. In addition, chaotic behavior appears in systems that are apparently very simple.

While the intuition that chaotic systems might exist is not new, the systematic exploration of chaotic systems started only in the 1970s. The discovery of the existence of nonlinear chaotic systems marked a conceptual crisis in the physical sciences: It challenges the very notion of the applicability of mathematics to the description of reality. Chaos laws are not testable on a large scale; their applicability cannot be predicted analytically. Nevertheless, the statistics of chaos theory might still prove to be meaningful.

The economy being a complex system, the expectation was that its apparently random behavior could be explained as a deterministic chaotic system of low dimensionality. Despite the fact that tests to detect low-dimensional chaos in the economy have produced a substantially negative response, it is easy to make macroeconomic and financial econometric models exhibit chaos. (See Brock, Dechert, Scheinkman, and LeBaron [1996] and Brock and Hommes [1997].) As a matter of fact, most macroeconomic models are nonlinear. Though chaos has not been detected in economic time-series, most economic dynamic models are nonlinear in more than three dimensions and thus potentially chaotic. At this stage of the research, we might conclude that if chaos exists in economics it is not of the low-dimensional type.

PARTIAL DIFFERENTIAL EQUATIONS

To illustrate the notion of a partial differential equation (PDE), let's start with equations in two

dimensions. An n -order PDE in two dimensions x, y is an equation of the form

$$F\left(x, y, \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \dots, \frac{\partial^{(i)} f}{\partial^{(k)} x \partial^{(i-k)} y}\right) = 0, 0 \leq k \leq i, 0 \leq i \leq n$$

A solution of the previous equation will be any function that satisfies the equation.

In the case of PDEs, the notion of initial conditions must be replaced with the notion of boundary conditions or initial plus boundary conditions. Solutions will be defined in a multidimensional domain. To identify a solution uniquely, the value of the solution on some subdomain must be specified. In general, this subdomain will coincide with the boundary (or some portion of the boundary) of the domain.

Diffusion Equation

Different equations will require and admit different types of boundary and initial conditions. The question of the existence and uniqueness of solutions of PDEs is a delicate mathematical problem. We can only give a brief account by way of an example.

Let's consider the *diffusion equation*. This equation describes the propagation of the probability density of stock prices under the random-walk hypothesis:

$$\frac{\partial f}{\partial t} = a^2 \frac{\partial^2 f}{\partial x^2}$$

The *Black-Scholes equation*, which describes the evolution of option prices, can be reduced to the diffusion equation.

The diffusion equation describes propagating phenomena. Call $f(t, x)$ the probability density that prices have value x at time t . In finance theory, the diffusion equation describes the time-evolution of the probability density function $f(t, x)$ of stock prices that follow a random walk.² It is therefore natural to impose initial and boundary conditions on the distribution of prices.

In general, we distinguish two different problems related to the diffusion equation: the first

boundary value problem and the Cauchy initial value problem, named after the French mathematician Augustin Cauchy who first formulated it. The two problems refer to the same diffusion equation but consider different domains and different initial and boundary conditions. It can be demonstrated that both problems admit a unique solution.

The first boundary value problem seeks to find in the rectangle $0 \leq x \leq l$, $0 \leq t \leq T$ a continuous function $f(t, x)$ that satisfies the diffusion equation in the interior Q of the rectangle plus the following initial condition,

$$f(0, x) = \phi(x), \quad 0 \leq x \leq l$$

and boundary conditions,

$$f(t, 0) = f_1(t), \quad f(t, l) = f_2(t), \quad 0 \leq t \leq T$$

The functions f_1, f_2 are assumed to be continuous and $f_1(0) = \phi(0), f_2(0) = \phi(l)$.

The Cauchy problem is related to an infinite half plane instead of a finite rectangle. It is formulated as follows. The objective is to find for any x and for $t \geq 0$ a continuous and bounded function $f(t, x)$ that satisfies the diffusion equation and which, for $t = 0$, is equal to a continuous and bounded function $f(0, x) = \phi(x), \forall x$.

Solution of the Diffusion Equation

The first boundary value problem of the diffusion equation can be solved exactly. We illustrate here a widely used method based on the separation of variables, which is applicable if the boundary conditions on the vertical sides vanish (that is, if $f_1(t) = f_2(t) = 0$). The method involves looking for a tentative solution in the form of a product of two functions, one that depends only on t and the other that depends only on x : $f(t, x) = h(t)g(x)$.

If we substitute the previous tentative solution in the diffusion equation

$$\frac{\partial f}{\partial t} = a^2 \frac{\partial^2 f}{\partial x^2}$$

we obtain an equation where the left side depends only on t while the right side depends

only on x :

$$\begin{aligned} \frac{dh(t)}{dt} g(x) &= a^2 h(t) \frac{d^2 g(x)}{dx^2} \\ \frac{dh(t)}{dt} \frac{1}{h(t)} &= a^2 \frac{d^2 g(x)}{dx^2} \frac{1}{g(x)} \end{aligned}$$

This condition can be satisfied only if the two sides are equal to a constant. The original diffusion equation is therefore transformed into two ordinary differential equations:

$$\begin{aligned} \frac{1}{a^2} \frac{dh(t)}{dt} &= bh(t) \\ \frac{d^2 g(x)}{dx^2} &= bg(x) \end{aligned}$$

with boundary conditions $g(0) = g(l) = 0$. From the above equations and boundary conditions, it can be seen that b can assume only the negative values,

$$b = -\frac{k^2 \pi^2}{l^2}, \quad k = 1, 2, \dots$$

while the functions g can only be of the form

$$g(x) = B_k \sin \frac{k\pi}{l} x$$

Substituting for h , we obtain

$$h(t) = B'_k \exp\left(-\frac{a^2 k^2 \pi^2}{l^2} t\right)$$

Therefore, we can see that there are denumerably infinite solutions of the diffusion equation of the form

$$f_x(t, x) = C_k \exp\left(-\frac{a^2 k^2 \pi^2}{l^2} t\right) \sin \frac{k\pi}{l} x$$

All these solutions satisfy the boundary conditions $f(t, 0) = f(t, l) = 0$. By linearity, we know that the infinite sum

$$\begin{aligned} f(t, x) &= \sum_{k=1}^{\infty} f_k(t, x) \\ &= \sum_{k=1}^{\infty} C_k \exp\left(-\frac{a^2 k^2 \pi^2}{l^2} t\right) \sin \frac{k\pi}{l} x \end{aligned}$$

will satisfy the diffusion equation. Clearly $f(t, x)$ satisfies the boundary conditions $f(t, 0) = f(t, l) = 0$. In order to satisfy the initial condition, given that $\phi(x)$ is bounded and continuous and

that $\phi(0) = \phi(l) = 0$, it can be demonstrated that the coefficients C_k can be uniquely determined through the following integrals, which are called the Fourier integrals:

$$C_k = \frac{2}{L} \int_0^L \phi(\xi) \sin\left(\frac{\pi k}{L} \xi\right) d\xi$$

The previous method applies to the first boundary value problem but cannot be applied to the Cauchy problem, which admits only an initial condition. It can be demonstrated that the solution of the Cauchy problem can be expressed in terms of a convolution with a Green's function. In particular, it can be demonstrated that the solution of the Cauchy problem can be written in closed form as follows:

$$f(t, x) = \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{\infty} \frac{\phi(\xi)}{\sqrt{t}} \exp\left\{-\frac{(x-\xi)^2}{4t}\right\} d\xi$$

for $t > 0$ and $f(0, x) = \phi(x)$. It can be demonstrated that the Black-Scholes equation, which

is an equation of the form

$$\frac{\partial f}{\partial t} + \frac{1}{2}\sigma^2 x^2 \frac{\partial^2 f}{\partial x^2} + rx \frac{\partial f}{\partial x} - rf = 0$$

can be reduced through transformation of variables to the standard diffusion equation to be solved with the Green's function approach.

Numerical Solution of PDEs

There are different methods for the numerical solution of PDEs. We illustrate the finite difference methods, which are based on approximating derivatives with finite differences. Other discretization schemes such as finite elements and spectral methods are possible but, being more complex, they go beyond the scope of this book.

Finite difference methods result in a set of recursive equations when applied to initial conditions. When finite difference methods are applied to boundary problems, they require the solution of systems of simultaneous linear equations. PDEs might exhibit boundary conditions, initial conditions, or a mix of the two.

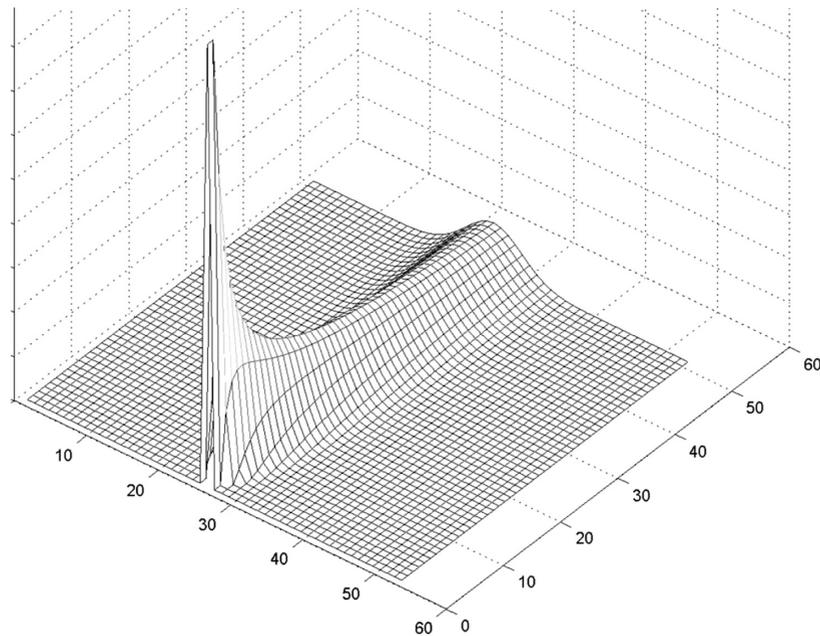


Figure 5 Solution of the Cauchy Problem by the Finite Difference Method

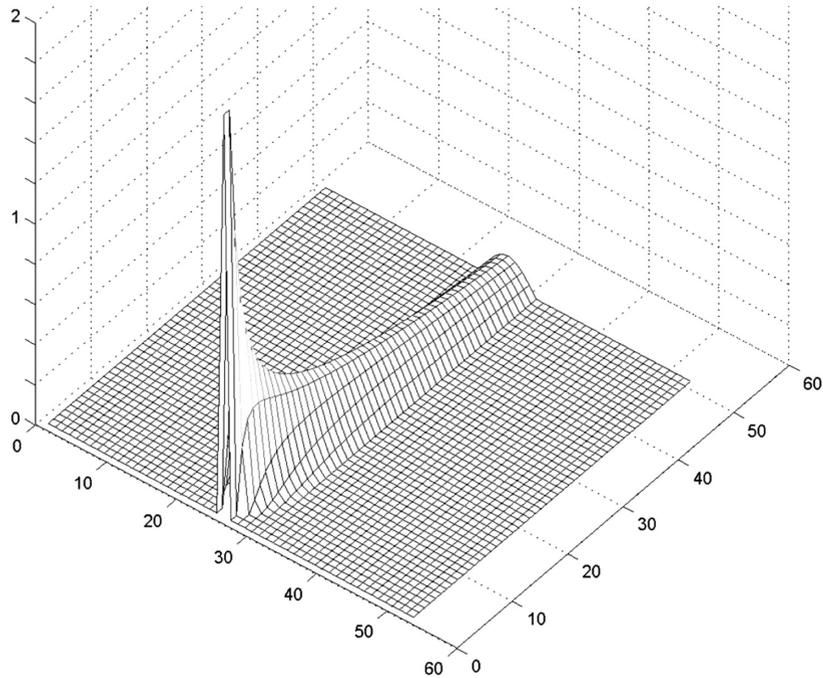


Figure 6 Solution of the First Boundary Problem by the Finite Difference Method

The Cauchy problem of the diffusion equation is an example of initial conditions. The simplest discretization scheme for the diffusion equation replaces derivatives with their difference quotients. As for ordinary differential equations, the discretization scheme can be written as follows:

$$\frac{\partial f}{\partial t} \approx \frac{f(t + \Delta t, x) - f(t, x)}{\Delta t}$$

$$\frac{\partial^2 f}{\partial x^2} \approx \frac{f(t, x + \Delta x) - 2f(t, x) + f(t, x - \Delta x)}{(\Delta x)^2}$$

In the case of the Cauchy problem, this approximation scheme defines the forward recursive algorithm. It can be proved that the algorithm is stable only if the Courant-Friedrichs-Lewy (CFL) conditions

$$\Delta t < \frac{(\Delta x)^2}{2a^2}$$

are satisfied.

Different approximation schemes can be used. In particular, the forward approximation to the derivative used above could be replaced by centered approximations. Figure 5 illustrates the solution of a Cauchy problem for initial conditions that vanish outside of a finite interval. The simulation shows that solutions diffuse in the entire half space.

Applying the same discretization to a first boundary problem would require the solution of a system of linear equations at every step. Figure 6 illustrates this case.

KEY POINTS

- Basically, differential equations are equations that express a relationship between a function and one or more derivatives (or differentials) of that function.
- The two classifications of differential equations are ordinary differential equations and partial differential equations. The classification depends on the type of derivatives

included in the differential equation: ordinary differential equation when there is only an ordinary derivative and partial differential equation where there are partial derivatives.

- Typically in differential equations, one of the independent variables is time.
- The term stochastic differential equation refers to a differential equation in which a derivative of one or more of the independent variables is a random variable or a stochastic process.
- Differential equations are conditions that must be satisfied by their solutions. Differential equations generally admit infinite solutions. Initial or boundary conditions are needed to identify solutions uniquely.
- Differential equations are the key mathematical tools for the development of modern science; in finance they are used in arbitrage pricing, to define stochastic processes, and to compute the time evolution of averages.
- Differential equations can be solved in closed form or with numerical methods. Finite difference methods approximate derivatives with difference quotients. Initial conditions yield recursive algorithms.
- Boundary conditions require the solution of linear equations.

NOTES

1. The condition of existence and continuity of derivatives is stronger than necessary. The Lipschitz condition, which requires that the incremental ratio be uniformly bounded in a given interval, would suffice.
2. In physics, the diffusion equation describes phenomena such as the diffusion of particles suspended in some fluid. In this case, the diffusion equation describes the density of particles at a given moment at a given point.

REFERENCES

- Brock, W., Dechert, W. D., Scheinkman, J. A., and LeBaron, B. (1996). A test for independence based on the correlation dimension. *Econometric Reviews* 15, 3: 197–235.
- Brock, W., and Hommes, C. (1997). A rational route to randomness. *Econometrica* 65, 5: 1059–1095.
- Brock, W., Hsieh, D., and LeBaron, B. (1991). *Nonlinear Dynamics, Chaos, and Instability*. Cambridge, MA: MIT Press.
- Hsieh, D. (1991). Chaos and nonlinear dynamics: Application to financial markets. *Journal of Finance* 46, 5: 1839–1877.
- King, A. C., Billingham, J., and Otto, S. R. (2003). *Differential Equations: Linear, Nonlinear, Ordinary, Partial*. New York: Cambridge University Press.

Partial Differential Equations in Finance

YVES ACHDOU, PhD

Professor, Lab. Jacques-Louis Lions, University Paris-Diderot, Paris, France

OLIVIER BOKANOWSKI, PhD

Associate Professor, Lab. Jacques-Louis Lions, University Paris-Diderot, Paris, France

TONY LELIÈVRE, PhD

Professor, CERMICS, Ecole des Ponts ParisTech, Marne-la-Vallée, France

Abstract: Partial differential equations are useful in finance in various contexts, in particular for the pricing of European and American options, for stochastic portfolio optimization, and for calibration. They can be used for simple options as well as for more exotic ones, such as Asian or lookback options. They are particularly useful for nonlinear models. They allow for the numerical computations of several spot prices at the same time. Numerical aspects, discretization methods, algorithms, and analysis of the numerical schemes have been under constant development during the last three decades. Finite difference methods are the simplest and most basic approaches. Finite element methods allow the use of nonuniform meshes and refinement procedures can then be applied and improve accuracy near a region of interest. Deterministic approaches based on partial differential equation formulations can also be used for calibration of various volatility models (such as local, stochastic, or Levy-driven volatility models) and by making use of Dupire's formula. Current research directions include the development of discretization methods for high-dimensional problems.

Numerical methods based on *partial differential equations (PDEs) in finance* are not very popular. Indeed, the models are usually derived from probabilistic arguments and Monte Carlo methods are therefore much more natural. Stochastic methods are also often simpler to implement than the algorithms used for solving the related PDEs. However, when it is possible to efficiently discretize the PDE (which

is not always the case, the typical counterexample being high-dimensional problems), deterministic methods are usually more efficient than stochastic ones. Moreover, the solution to the partial differential equation gives more information. In the context of option pricing, one obtains, for example, the price of the option for all values of the maturity and for all spot prices, while the probabilistic formulation

typically gives the value of the option for a fixed maturity and a fixed spot price. In particular, this is useful for computing derivatives of the option's price with respect to some parameters of the model (the so-called "Greeks").

The PDEs obtained in finance have several characteristics. First, they are posed on a bounded domain in time $(0, T)$, with typically a singular final condition at the maturity $t = T$, and very often in an unbounded domain in the spot variable, which requires to impose suitable "boundary conditions" at infinity to get well-posed problems and to use appropriate numerical approximations (truncation to a bounded domain and artificial boundary conditions). These PDEs are usually of parabolic type, but often with degenerate diffusions. Because of operational constraints, the numerical methods used for the discretization of the PDE must be sufficiently fast and accurate to be useful in practice. These peculiarities of PDEs in finance explain the need for up-to-date and sometimes involved numerical methods.

In this entry we focus on numerical issues and try to review the main numerical methods used for solving PDEs in finance. This presentation heavily relies on Achdou and Pironneau (2005), as well as Lamberton and Lapeyre (1997), Karatzas and Shreve (1991), and Wilmott, Dewynne, and Howison (1993).

PARTIAL DIFFERENTIAL EQUATIONS FOR OPTION PRICING

In this section, we present the main arguments to derive a PDE for the price of various European and American options.

A Primer: The Black and Scholes Model for European Options

The aim of this section is to recall the basic tools needed to derive a PDE in the context of option pricing, without providing all the detailed assumptions required on the data to perform

this derivation. Karatzas and Shreve (1991) and Lamberton and Lapeyre (1997), for example, provide more details on the mathematical aspects. We adopt the standard *Black and Scholes* model (Black and Scholes, 1973; Merton, 1973) with a risky asset whose price at time t is S_t and a risk-free asset whose price at time t is S_t^0 , such that:

$$dS_t = S_t(\mu dt + \sigma dB_t), \quad dS_t^0 = rS_t^0 dt$$

The process B_t is a standard Brownian motion defined on a probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{Q})$, and μ (the mean rate of return), r (the interest rate), and $\sigma > 0$ (the volatility) are three constants. However, the following can be generalized to the case where μ , r , and $\sigma > 0$ are functions of t and S (under suitable smoothness assumptions).

We introduce the stochastic process $W_t = B_t + \frac{\mu-r}{\sigma}t$. Under the so-called risk-neutral probability \mathbb{P} defined by its Radon-Nikodym derivative with respect to \mathbb{Q} by

$$\frac{d\mathbb{P}}{d\mathbb{Q}} \Big|_{\mathcal{F}_t} = \exp \left(\int_0^t \frac{r-\mu}{\sigma} dB_s - \frac{1}{2} \int_0^t \left(\frac{r-\mu}{\sigma} \right)^2 ds \right)$$

W_t is a Brownian motion and S_t/S_t^0 is a martingale. This is one of the fundamental properties of the stochastic process needed in the following. The process S_t satisfies the following stochastic differential equation (SDE) under \mathbb{P} :

$$dS_t = S_t(r dt + \sigma dW_t) \quad (1)$$

Let us now consider a portfolio with H_t risky assets and H_t^0 no-risk assets. Its value at time t is:

$$P_t = H_t S_t + H_t^0 S_t^0 \quad (2)$$

We suppose that this portfolio is self-financing (any manipulation on this portfolio, i.e., any change of the values of H_t or H_t^0 , is done without any inflows or outflows of money), which translates into

$$dP_t = H_t dS_t + H_t^0 dS_t^0 \quad (3)$$

The value of a self-financing portfolio changes if and only if the price of the risky asset changes.

Using (3), it is possible to show that P_t/S_t^0 is also a martingale.

We consider the following problem: For a given function ϕ (the payoff function) and a given time $T > 0$ (the maturity), is it possible to build a self-financing portfolio such that $P_T = \phi(S_T)$? Classical examples of function ϕ are $\phi(S) = (S - K)_+$ (vanilla call) or $\phi(S) = (S - K)_-$ (vanilla put), where, for any real x , $x_+ = \max(x, 0)$ and $x_- = \max(-x, 0)$. The answer is positive (this is typically based on a martingale representation theorem, the fact that P_t/S_t^0 is a martingale, and the fact that the payoff $\phi(S_T)$ is \mathcal{F}_T -measurable), and it is then possible to show that such a portfolio has the following value at time t :

$$P_t = \mathbb{E} \left(\exp \left(- \int_t^T r \, ds \right) \phi(S_T) \middle| \mathcal{F}_t \right) \quad (4)$$

where here and in the following, \mathbb{E} denotes an expectation with respect to the risk-neutral probability \mathbb{P} . By the so-called arbitrage-free principle, P_t is actually the “fair price” at time t of the option, which enables its owner to get the payoff $\phi(S_T)$ at time T . In the particular context of vanilla options, the solution is analytically known, at least if r and σ are constant: This is the celebrated Black and Scholes formula. However, in the case when r and σ are functions of t and S , (4) cannot be estimated without a numerical method. We are interested in deterministic numerical methods, based on a PDE related to (4).

The second fundamental property of the stochastic process S_t required to obtain a PDE formulation of this problem is a Markov property. Roughly speaking, it states that the expectation of any function of $(S_t)_{0 \leq t \leq T}$ conditionally to \mathcal{F}_t is actually a function of the price S_t of the risky asset at time t . In our context, this property shows that P_t writes

$$P_t = p(t, S_t) \quad (5)$$

where p is a function of $t \in [0, T]$ and $S \in [0, \infty)$, called the pricing function of the option. Notice that even if (5) only involves the value of p at

point (t, S_t) , the pricing function p is a deterministic function defined for all values of $t \geq 0$ and $S \geq 0$. By the Markov property of S_t , we also have the following representation formula for p :

$$p(t, x) = \mathbb{E} \left(\exp \left(- \int_t^T r \, ds \right) \phi(S_T^{t,x}) \right) \quad (6)$$

where $(S_\theta^{t,x})_{t \leq \theta \leq T}$ denotes the process solution to (1) starting from x at time t

$$\begin{cases} dS_\theta^{t,x} = S_\theta^{t,x} (r \, d\theta + \sigma \, dW_\theta), & \theta \geq t, \\ S_t^{t,x} = x \end{cases} \quad (7)$$

By using Ito’s calculus and the fact that P_t/S_t^0 is a martingale, we then obtain that p should satisfy the following backward-in-time PDE:

$$\begin{cases} \frac{\partial p}{\partial t} + rS \frac{\partial p}{\partial S} + \frac{\sigma^2 S^2}{2} \frac{\partial^2 p}{\partial S^2} - rp = 0, \\ p(T, S) = \phi(S) \end{cases} \quad (8)$$

Conversely, it is possible (using again a martingale representation theorem) to show that if p satisfies (8), then $p(t, S_t)$ is the value of a self-financing portfolio with value $\phi(S_T)$ at time T . Moreover, one can check that $\frac{\partial p}{\partial S}(t, S_t) = H_t$, which shows that obtaining an accurate approximation of $\frac{\partial p}{\partial S}$ is important in order to estimate the quantity of risky asset H_t needed at time t to build the portfolio with value P_t (this is the hedging strategy). Collectively, equations (4)–(5) and (8) provide an example of so-called Feynman-Kac formulas, which are used in many other contexts (quantum chemistry or transport equations, for example) either to give a probabilistic interpretation to a PDE, or to recast the computation of an expectation into a PDE problem.

For problem (8) to be well posed (i.e., for one and only one solution to exist), one needs to supply the system with “boundary conditions” when $S = 0$ or $S \rightarrow \infty$. More precisely, one needs to make precise in which functional space the function p is looked for. This will be explained in the next section.

From the PDE (8) and the so-called maximum principle, it is possible to derive many qualitative properties and a priori bounds on the price p (like the call-put parity, for example; see Achdou and Pironneau, 2005). Roughly speaking, the maximum principle states that if the data (initial condition, boundary conditions, right-hand side) for the PDE (8) are positive, then the solution is positive. This property is definitely necessary to hold for a price. It is also an important property to check on the numerical schemes (which is then called a discrete maximum principle as discussed below).

It is also possible to obtain the PDE without introducing the risk-neutral probability (see Wilmott, Dewynne, and Howison, 1993) by considering a portfolio containing some options and some risky assets and by using an arbitrage-free argument.

It is important to recall that the Black and Scholes model for the evolution of the risky asset (1) badly compares with experimental data. We discuss later in this entry some possible refinements that have been introduced in order to better fit the observations (see the discussion on calibration below). However, this model remains very important in practice because it is used as a prototypical description of the evolution of the asset. Moreover, for a given observed price of a derivative, there exists a constant volatility σ (called the implied volatility; see the section on calibration below) for which the Black-Scholes price is the observed price. The implied volatility is a major quantity used in practice to compare derivatives.

Other Options

The argument presented for the Black-Scholes model is prototypical. In particular, the derivation of a PDE satisfied by the pricing function of an option always relies on the two fundamental properties stressed above: the martingale and the Markov properties of a suitable stochastic process. In this section, we present PDEs for the prices of various options without providing all the details of the derivation.

Basket Options

In many cases, the payoff of the option depends on the values of more than one asset, which typically do not evolve independently. Let us, for example, consider the case of two assets, which evolve following the following SDE under the neutral risk probability

$$\begin{cases} dS_t^1 = S_t^1 (r dt + \sigma_1 dW_t^1) \\ dS_t^2 = S_t^2 (r dt + \sigma_2 dW_t^2) \end{cases}$$

where W_t^1 and W_t^2 are possibly correlated standard Brownian motions. We call ρ the correlation of W_t^1 and W_t^2 : $d\langle W_1, W_2 \rangle_t = \rho dt$. We suppose that the maturity is $T > 0$ and the payoff is $\phi(S_T^1, S_T^2)$, where ϕ is a given function. It is then possible to show that the price of the option at time t is $p(t, S_t^1, S_t^2)$ where p satisfies

$$\begin{cases} \frac{\partial p}{\partial t} + rS_1 \frac{\partial p}{\partial S_1} + rS_2 \frac{\partial p}{\partial S_2} + \frac{\sigma_1^2 S_1^2}{2} \frac{\partial^2 p}{\partial S_1^2} \\ \quad + \frac{\sigma_2^2 S_2^2}{2} \frac{\partial^2 p}{\partial S_2^2} + \rho\sigma_1\sigma_2 S_1 S_2 \frac{\partial^2 p}{\partial S_1 \partial S_2} - rp = 0, \\ p(T, S_1, S_2) = \phi(S_1, S_2) \end{cases} \quad (9)$$

Here again, r , σ_1 , and σ_2 may be functions of t and (S_1, S_2) . It is possible to solve such PDEs by standard numerical methods up to dimension 3 or 4. As discussed later, to derive appropriate discretization for higher dimensions is not an easy task and is still the subject of current research.

Barrier Options

Again, let us consider an option on a single asset. For some options, the payoff becomes 0 if there exists a time $t \in [0, T]$ such that S_t goes below a or above b , where a and b are two given values, $0 < a < b$ (the case $a = 0$ or $b = \infty$ can be treated similarly). Mathematically, the payoff is $1_{\forall t \in [0, T], S_t \in [a, b]} \phi(S_T)$ where, for any event $A \subset \Omega$, 1_A denotes the characteristic function of A , and S_t satisfies (1). In this case, the relevant stochastic process for deriving the PDE is $S_{t \wedge \tau}$, where $\tau = \inf\{t \in [0, T], S_t \geq b \text{ or } S_t \leq a\}$ is a stopping time, and, for any real x and y ,

$x \wedge y = \inf(x, y)$. It can be checked that $S_{t \wedge \tau}$ is a Markov process, and that $S_{t \wedge \tau} / S_{t \wedge \tau}^0$ is a martingale. It is then possible to show that the price of the option at time t is $p(t \wedge \tau, S_{t \wedge \tau})$ where p is defined for $t \in [0, T]$ and $S \in [a, b]$ and satisfies:

$$\begin{cases} \frac{\partial p}{\partial t} + rS \frac{\partial p}{\partial S} + \frac{\sigma^2 S^2}{2} \frac{\partial^2 p}{\partial S^2} - rp = 0, \\ p(T, S) = \phi(S), \\ p(t, a) = p(t, b) = 0 \end{cases} \quad (10)$$

Here again, r and σ may be functions of t and S . Moreover, the generalization to basket options is straightforward, as explained above. In this case, it is possible to consider more general *barriers*, namely a payoff of the form $1_{\forall t \in [0, T], (S_t^1, S_t^2, \dots, S_t^d) \in \mathcal{D}} \phi(S_T)$, where d denotes the number of underlying assets and \mathcal{D} is any simple connected domain of \mathbb{R}^d . The appropriate discretization for general domains \mathcal{D} is the finite element method that will be discussed later on.

Options on the Maximum

For some options (the so-called *lookback options*), the payoff involves the maximum of the risky asset. For example, it writes $\phi(S_T, M_T)$ where $M_t = \max_{0 \leq r \leq t} S_r$ and S_t satisfies (1). One can check that (S_t, M_t) is a Markov process. It is then possible to show that the price of the option at time t is $p(t, S_t, M_t)$ where p is defined for $t \in [0, T]$ and $(S, M) \in \{(S, M) \in \mathbb{R}^2, 0 \leq S \leq M\}$ and satisfies:

$$\begin{cases} \frac{\partial p}{\partial t} + \frac{\sigma^2}{2} S^2 \frac{\partial^2 p}{\partial S^2} + rS \frac{\partial p}{\partial S} - rp = 0, \\ p(T, S, M) = \phi(S, M), \\ \frac{\partial p}{\partial M}(t, S, S) = 0 \end{cases} \quad (11)$$

If the payoff is of the form $\phi(S, M) = M\tilde{\phi}(S/M)$, it is possible to reduce the problem to a two-dimensional one (including the time variable). Indeed, one can check by straightforward computations that $p(t, S, M) = Mw(t, S/M)$ where w is a function of $t \in [0, T]$ and $\xi \in [0, 1]$, which

satisfies:

$$\begin{cases} \frac{\partial w}{\partial t} + \frac{\sigma^2}{2} \xi^2 \frac{\partial^2 w}{\partial \xi^2} + r\xi \frac{\partial w}{\partial \xi} - rw = 0, \\ w(T, \xi) = \tilde{\phi}(\xi), \\ \frac{\partial w}{\partial \xi}(t, 1) = w(t, 1) \end{cases} \quad (12)$$

Notice that this reduction is not generally possible for (t, S, M) -dependent interest rate and volatility (except for very peculiar dependencies).

Options on the Average

Some options (the so-called *Asian options*) involve the average of the risky asset. More precisely, the payoff writes $\phi(S_T, A_T)$ where $A_t = \frac{1}{t} \int_0^t S_r dr$ and S_t satisfies (1). One can check that (S_t, A_t) is a Markov process. Using this property, it is possible to show that the price of the option at time t is $p(t, S_t, A_t)$ where p is defined for $t \in [0, T]$ and $(S, A) \in [0, \infty)^2$, and p satisfies:

$$\begin{cases} \frac{\partial p}{\partial t} + \frac{\sigma^2 S^2}{2} \frac{\partial^2 p}{\partial S^2} + rS \frac{\partial p}{\partial S} \\ + \frac{1}{t}(S - A) \frac{\partial p}{\partial A} - rV = 0, \\ p(T, S, A) = \phi(S, A) \end{cases} \quad (13)$$

In some cases (see Rogers and Shi, 1995), it is possible to reduce this problem to a one-dimensional PDE. More precisely, for fixed strike call ($\phi(S, A) = (A - K)_+$) or fixed strike put ($\phi(S, A) = (K - A)_+$), we have $p(t, S, A) = Sf(t, \frac{K-tA/T}{S})$ where f satisfies

$$\begin{cases} \frac{\partial f}{\partial t} + \frac{\sigma^2 \xi^2}{2} \frac{\partial^2 f}{\partial \xi^2} - \left(\frac{1}{T} + r\xi\right) \frac{\partial f}{\partial \xi} = 0, \\ f(T, \xi) = \tilde{\phi}(\xi) \end{cases} \quad (14)$$

and $\tilde{\phi}(\xi) = \xi_-$ (resp. $\tilde{\phi}(\xi) = \xi_+$). This reduction of (13) to (14) is also possible for floating strike call ($\phi(S, A) = (S - A)_+$) (resp. for floating strike put ($\phi(S, A) = (A - S)_+$)) by setting $p(t, S, A) = Sf(t, -\frac{tA}{TS})$ and $\tilde{\phi}(\xi) = (1 + \xi)_+$ (resp. $\tilde{\phi}(\xi) = (1 + \xi)_-$). However, this reduction

is generally not possible for general payoff function or (t, S, A) -dependent interest rate and volatility (except for very peculiar dependencies).

Bermudean Options

As a transition between European and American options, we would like to mention that it is very easy to price Bermudean options with the PDE approach. For such options, the contract can be exercised only at certain days between the present time and the maturity. Mathematically, for an option on a single asset (the spot price is called S) and if ϕ denotes the payoff, the pricing function satisfies $p(t_i^+, S) = \max(p(t_i^-, S), \phi(S))$, at each exercising time t_i , and (8) between the exercising times; see Duffie (1992, p. 211).

The Case of American Options

We have so far presented so-called *European options*, that is, some options that enable their owners to get $\phi(S_T)$ at a fixed time T . On the other hand, *American options* can be exercised at any time up to the maturity. Hence the price of an American option of payoff ϕ and maturity T will be the maximum of all possible expectations such as (6) for stopping times τ between t and T , that is, for $t \in [0, T]$ and $x \geq 0$,

$$p(t, x) = \sup_{\tau \in \mathcal{T}_{[t, T]}} \mathbb{E} \left(e^{-\int_t^\tau r ds} \phi(S_\tau^{t, x}) \right) \quad (15)$$

where $\mathcal{T}_{[t, T]}$ denotes the set of stopping times τ of the filtration \mathcal{F}_t , with values in $[t, T]$.

The PDE for American Options

We now present the main arguments to derive a PDE on p defined by (15) (or more precisely a system of partial differential inequalities).

Notice first that taking $\tau = t$ in (15) yields the inequality

$$p(t, x) \geq \phi(x) \quad (16)$$

Moreover, we clearly have from (15) $p(T, x) = \phi(x)$.

Let t and δt be such that $0 \leq t \leq t + \delta t \leq T$. From (15) we have:

$$\begin{aligned} & e^{-\int_0^{t+\delta t} r ds} p(t + \delta t, S_{t+\delta t}^{t, x}) \\ &= \sup_{\tau \in \mathcal{T}_{[t+\delta t, T]}} \mathbb{E} \left(e^{-\int_0^\tau r ds} \phi \left(S_\tau^{t+\delta t, S_{t+\delta t}^{t, x}} \right) \right), \\ &\leq \sup_{\tau \in \mathcal{T}_{[t, T]}} \mathbb{E} \left(e^{-\int_0^\tau r ds} \phi \left(S_\tau^{t, x} \right) \right), \\ &\leq e^{-\int_0^t r ds} p(t, x) \end{aligned}$$

where we have used the fact that: $S_\tau^{t+\delta t, S_{t+\delta t}^{t, x}} = S_\tau^{t, x}$. By Ito's calculus (taking the limit $\delta t \rightarrow 0$), we thus obtain

$$-\frac{\partial p}{\partial t} + \mathcal{A}p \geq 0 \quad (17)$$

where we have introduced the linear PDE operator

$$\mathcal{A}p = -rS \frac{\partial p}{\partial S} - \frac{\sigma^2 S^2}{2} \frac{\partial^2 p}{\partial S^2} + rp \quad (18)$$

Combined with (16), we then obtain

$$\min \left(-\frac{\partial p}{\partial t} + \mathcal{A}p, p - \phi \right) \geq 0 \quad (19)$$

Our aim is now to show that the inequality in (19) is actually an equality. This is done in several steps, and requires us to identify an optimal stopping time τ^* for which the supremum in (15) is obtained. For a fixed (t, x) , let us introduce the stopping time $\tau^* \in \mathcal{T}_{[t, T]}$ defined by

$$\tau^* = \inf \{ \theta \geq t, p(\theta, S_\theta^{t, x}) = \phi(S_\theta^{t, x}) \}, \text{ a.s.} \quad (20)$$

(notice that $\tau^* \leq T$ since $p(T, x) = \phi(x)$). It can be shown (see Appendix) that

$$\begin{aligned} p(t, x) &= \mathbb{E} \left(e^{-\int_t^{\tau^*} r ds} \phi \left(S_{\tau^*}^{t, x} \right) \right) \\ &= \mathbb{E} \left(e^{-\int_t^{\tau^*} r ds} p(\tau^*, S_{\tau^*}^{t, x}) \right) \end{aligned} \quad (21)$$

Using a decreasing property (65) proved in the Appendix, one then obtains that for any $\delta t > 0$,

$$\begin{aligned} p(t, x) &= \mathbb{E} \left(e^{-\int_t^{\tau_{\delta t}^*} r ds} p \left(\tau_{\delta t}^*, S_{\tau_{\delta t}^*}^{t, x} \right) \right), \\ &\text{where } \tau_{\delta t}^* = (t + \delta t) \wedge \tau^* \end{aligned} \quad (22)$$

This can be seen as a dynamic programming principle (or Bellman’s principle). For a European option we would have more simply

$$p(t, x) = \mathbb{E} \left(e^{-\int_t^{t+\delta t} r ds} p(t + \delta t, S_{t+\delta t}^{t,x}) \right)$$

Now if we suppose that $p(t, x) > \phi(x)$, then for any $\delta t > 0$ we have $\mathbb{P}(\tau_{\delta t}^* > t) = 1$. Considering Ito’s formula in (22), and by (17), we obtain $(-\frac{\partial p}{\partial t} + \mathcal{A}p)(\theta, S_{\theta}^{t,x}) = 0$ for $t \leq \theta \leq \tau_{\delta t}^*$, thus leading to $(-\frac{\partial p}{\partial t} + \mathcal{A}p)(t, x) = 0$. This shows that the inequality in (19) is actually an equality.

Hence the PDE for the American option is

$$\begin{cases} \min \left(-\frac{\partial p}{\partial t} + \mathcal{A}p, p - \phi \right) = 0, \\ t \in [0, T], x \geq 0, \\ p(T, x) = \phi(x), \quad x \geq 0 \end{cases} \quad (23)$$

where \mathcal{A} is defined by (18). The major difference between the PDE (23) for American options and the PDE (8) for European options is that (23) is a nonlinear equation. This makes the theory of existence and uniqueness as well as the numerical approximation more difficult than for European options.

In the presentation above, we have used Ito’s formula, which requires that p is C^1 in time and C^2 in the spot variable. This is not true in general. It is however possible, following the same lines, to prove that p is a weak solution to (23) in the viscosity sense. For a historical derivation of this PDE, see Bensoussan and Lions (1978) or El Karoui (1981) where a variational formulation of (23) is derived (see (52) below). We also refer to Oksendal and Rekvam (1998) for an infinite horizon-related problem, Crandall, Ishii, and Lions (1992) for general results, Pham (1998) for an approach of optimal stopping including jump diffusion processes, and to Barles (1994) for the case of a discontinuous payoff ϕ .

PRICING EUROPEAN OPTIONS WITH PDEs

The aim of this section is to present two classes of methods for solving partial differential equa-

tions with some applications to the PDEs derived previously. We first introduce the *finite difference method*, which is based on approximation of the differential operators by Taylor expansions, and then the *finite element methods*, which belong to the wider class of Galerkin methods and are based on a variational formulation of the PDE. We try to stress the most important aspects of the numerical methods and refer, for example, to Achdou and Pironneau (2005 and 2009) for a more comprehensive presentation.

The Finite Difference Method for European Options

We first present the simplest approach to discretize a PDE: the finite difference method.

Basic Schemes

Let us introduce the finite difference method on the simple PDE (8). Let us first concentrate on the discretization of (8) with respect to the variable S . The principle is to divide the interval $[0, S_{\max}]$ into I intervals of length $\delta S = S_{\max}/I$ (where S_{\max} has to be chosen large enough, see below), and to approximate the derivatives by finite differences. A possible semidiscretization of (8) is: for $i \in \{0, 1, \dots, I\}$,

$$\begin{cases} \frac{\partial P_i}{\partial t} + rS_i \frac{P_{i+1} - P_{i-1}}{2\delta S} \\ + \frac{\sigma^2 S_i^2}{2} \frac{P_{i+1} - 2P_i + P_{i-1}}{\delta S^2} - rP_i = 0, \\ P_i(T) = \phi(S_i) \end{cases} \quad (24)$$

where $S_i = i\delta S$ denotes the i -th discretization point, and $P_i(t)$ is intended to be an approximation of $p(t, S_i)$. Now, (24) is a system of coupled ordinary differential equations (ODEs). The generalization to the case of a time and spot dependent r or σ is straightforward.

Notice that for $S = 0$, P_0 can be solved independently (since $S_0 = 0$): $P_0(t) = \phi(0) \exp(-\int_t^T r ds)$. In order to obtain a solution of the whole system of ODEs, one needs to define an appropriate boundary condition at

$S = S_{\max}$. Indeed, (24) taken at $i = I$ involves P_{I+1} which is a priori not defined. There are basically two methods to deal with this issue. The first one consists of using some a priori knowledge on the values of $p(t, S)$ when S is large and making some approximations of $p(t, S_{\max})$. In this case, the value of P_I is given as a data (this is a so-called Dirichlet boundary condition), and the unknowns are $(P_i)_{0 \leq i \leq I-1}$. For example, in the case of a put ($\phi(S) = (S - K)_-$) (resp. a call ($\phi(S) = (S - K)_+$)), it is known that $\lim_{S \rightarrow \infty} p(t, S) = 0$ (resp., in the limit $S \rightarrow \infty$, $p(t, S) \sim S - K \exp(-\int_t^T r ds)$), so that one can set $P_I(t) = 0$ (resp. $P_I(t) = S_{\max} - K \exp(-\int_t^T r ds)$). The error introduced by these artificial boundary conditions can be estimated. Another method is based on some knowledge on the asymptotic behavior of the derivatives of p . For example, in the case of the put, one can use the so-called homogeneous Neumann boundary condition, which writes $\partial p / \partial S(t, S_{\max}) = 0$ at the continuous level and $\frac{P_{I+1}(t) - P_I(t)}{\delta S} = 0$ at the discrete level. In this case, the unknowns are $(P_i)_{0 \leq i \leq I}$. For both methods, S_{\max} should be chosen sufficiently large. In practice, the quality of the method may be assessed by measuring how sensitive the result is to the value of S_{\max} .

Let us now consider the time discretization. Here again, the idea is to divide the time interval $[0, T]$ into N intervals of length $\delta t = T/N$ and to replace the time derivative by a finite difference. Three numerical methods are classically used:

$$\left\{ \begin{array}{l} \frac{P_i^{n+1} - P_i^n}{\delta t} + rS_i \frac{P_{i+1}^{n+1} - P_{i-1}^{n+1}}{2\delta S} \\ + \frac{\sigma^2 S_i^2}{2} \frac{P_{i+1}^{n+1} - 2P_i^{n+1} + P_{i-1}^{n+1}}{\delta S^2} - rP_i^{n+1} = 0, \\ P_i^N = \phi(S_i) \end{array} \right. \quad (25)$$

$$\left\{ \begin{array}{l} \frac{P_i^{n+1} - P_i^n}{\delta t} + rS_i \frac{P_{i+1}^n - P_{i-1}^n}{2\delta S} \\ + \frac{\sigma^2 S_i^2}{2} \frac{P_{i+1}^n - 2P_i^n + P_{i-1}^n}{\delta S^2} - rP_i^n = 0, \\ P_i^N = \phi(S_i) \end{array} \right. \quad (26)$$

or

$$\left\{ \begin{array}{l} \frac{P_i^{n+1} - P_i^n}{\delta t} + \frac{1}{2} \left(rS_i \frac{P_{i+1}^{n+1} - P_{i-1}^{n+1}}{2\delta S} \right. \\ + \frac{\sigma^2 S_i^2}{2} \frac{P_{i+1}^{n+1} - 2P_i^{n+1} + P_{i-1}^{n+1}}{\delta S^2} - rP_i^{n+1} \\ + rS_i \frac{P_{i+1}^n - P_{i-1}^n}{2\delta S} + \frac{\sigma^2 S_i^2}{2} \frac{P_{i+1}^n - 2P_i^n + P_{i-1}^n}{\delta S^2} \\ \left. - rP_i^n \right) = 0, \\ P_i^N = \phi(S_i) \end{array} \right. \quad (27)$$

where P_i^n is intended to be an approximation of $p(t_n, S_i)$, with $t_n = n\delta t$. Notice that using the discretization scheme (25) (the so-called explicit Euler scheme), the values of $(P_i^n)_{0 \leq i \leq I}$ are explicitly obtained from the values of $(P_i^{n+1})_{0 \leq i \leq I}$. On the contrary, in the two other schemes (26) (implicit Euler scheme) or (27) (Crank-Nicolson scheme), the values of $(P_i^n)_{0 \leq i \leq I}$ are obtained from the values of $(P_i^{n+1})_{0 \leq i \leq I}$ through the resolution of a linear system, which is more demanding from the computational viewpoint. Various numerical methods can be used for solving this linear system; here, we cannot describe them in detail. Let us simply mention that basically, there exist two classes of methods: the direct methods, which are based on Gaussian elimination, and the iterative methods, which consist of computing the solution as the limit of a sequence of approximations and which only require matrix-vector multiplications. The method of choice depends on the characteristics of the problem.

Notions of Stability and Consistency

In order to analyze the convergence of the three discretization schemes (25), (26), and (27), and to understand the differences between these schemes, we need to introduce two important notions. The first notion is the consistency. A numerical method is said to be consistent if, when the exact solution is plugged into the numerical scheme, the error tends to zero when the discretization parameters tend to zero. In our context, it consists of replacing P_i^n in (25),

(26), or (27) by $p(t_n, S_i)$, where p satisfies (8), and to check that the remaining terms tend to zero when δt and δS tend to zero. By using Taylor expansions, one can check that for (25) and (26) (resp. for (27)), the remaining terms are bounded from above by $C(\delta t + \delta S^2)$ (resp. by $C(\delta t^2 + \delta S^2)$), where C denotes a constant, which depends on some norms of the derivatives of p . Therefore (25) and (26) (resp. (27)) are consistent discretization schemes of order 2 in the spot variable, and of order 1 (resp. 2) in time. The second important notion is the stability. A numerical method is said to be stable if the norm of the solution to the numerical scheme is bounded from above by a constant (independent of the discretization parameters) multiplied by the norm of the data (initial condition, boundary conditions, right-hand side). This property is clearly satisfied if the numerical method is convergent, that is, if the numerical approximation converges to the solution of the PDE when the discretization parameters tend to zero. A general result states that, conversely, a consistent and stable discretization scheme is indeed convergent. The estimate of convergence is given by the estimate of consistency error. For example, if p is smooth enough, the error for the EI scheme is bounded from above by $C(\delta t + \delta S^2)$. Notice that the constant C in these estimates depends on the solution p . Higher order schemes will lead to better error estimates as soon as the solution of the continuous problem is smooth enough: The higher the order, the more regular p must be in order to take full advantage of the scheme. For example, for some parameters, it may happen that the results obtained with the CN scheme around $t = T$ are not better than those obtained with an order one scheme (IE or EE) since the solution is not sufficiently regular in time around $t = T$.

To give a precise meaning to all these results would require us to specify the norms used to measure the errors and to prove the stability. Let us simply mention that two norms are used in practice: The stability in L^∞ -norm (the supremum of the absolute values of the components) is related to a discrete maximum prin-

ciple (see below); and the stability in L^2 -norm (the Euclidean norm of the vector) is related to an energy estimate on the variational formulation. We refer, for example, to Achdou and Pironneau (2005) for more details.

The discrete maximum principle is the counterpart at the discrete level of the maximum principle at the continuous level mentioned above. It states that if the data for the numerical schemes are positive, then the solution is positive. Such schemes are by construction stable in L^∞ -norm. There exist deterministic numerical methods based on a probabilistic representation of the stock evolution on a binomial or a trinomial tree. Such methods can be interpreted as explicit finite difference methods to solve the PDE (8) and naturally satisfy a discrete maximum principle.

Convergence Analysis

Let us now discuss the properties of the three discretization schemes. We already mentioned that they are all consistent. On the other hand, it can be shown that the explicit scheme (25) is stable under an additional assumption (a so-called CFL condition; see Courant, Friedrichs, and Lewy, 1967) of the form $\delta t \leq C\delta S^2$, where C denotes a positive constant. The other two schemes (26) and (27) are unconditionally stable (in L^2 -norm). In conclusion, with the explicit scheme, the values of $(P_i^n)_{0 \leq i \leq I}$ can be very rapidly obtained from the values of $(P_i^{n+1})_{0 \leq i \leq I}$, but the time step must be sufficiently small with respect to the spot step to guarantee stability and hence convergence. On the other hand, the implicit schemes (26) and (27) require the resolution of a linear system at each time-step, but converge without any restriction on the time-step. This situation is very general for the parabolic PDEs obtained in finance. In terms of computational costs, the balance is generally in favor of the implicit schemes, since the CFL condition appears to be very stringent in practice. Concerning the stability in L^∞ -norm, let us just mention that the implicit schemes above do not satisfy the discrete maximum

Table 1 Error on the Value of a Call in Function of the Number of Intervals I in the Variable S , for the Implicit Euler (IE) Scheme

$N = 1000$	$I = 150$	$I = 300$	$I = 600$	$I = 1200$
IE	0.165	0.0356	0.00103	0.000452

principle and are not L^∞ -stable as such. These properties are, however, satisfied after a small modification of the discretization of the advection term $rS \frac{\partial p}{\partial S}$ (this is a so-called upwinding technique), which amounts to adding a diffusion term of order δS , which implies that this modified scheme becomes only of order 1 in the spot variable. Thus, the price to pay to get L^∞ -stability is a loss of one order of convergence.

In Tables 1 and 2, we illustrate this analysis by computing the error on the price of a call with $r = 0.1, \sigma = 0.01, K = 100, T = 1, S_0 = 100$, and $S_{\max} = 300$ for the three discretization schemes (25), (26), and (27), and various values of the numerical parameters I and N . The reference value ($P = 9.51625$) is obtained by the analytic Black and Scholes formula. In particular, one can check that the rates of convergence with respect to δt and δS are indeed those predicted by the analysis.

Before presenting an extension of this discretization method to Asian options, we mention the interest of a classical change of variable for the spot variable. It is indeed well known that by a change of variable $x = \ln S$, it is possible to get rid of the dependency in S of the advection and diffusion terms in (8). It is not better to discretize the PDE after this change of

variable, since it corresponds to taking a grid refined near $S = 0$, which is useless in this case. As we will see below, what actually matters is to refine the grid around the singularity of p (i.e., around $S = K$). A finite element approach is better suited in order to implement these refinements.

Application to Asian Options

We now present a less easy implementation of a finite difference method for pricing Asian options (see Dubois and Lelièvre, 2005). More precisely, we focus on computing numerical solutions to (14) for a fixed strike call:

$$\tilde{\phi}(\xi) = \xi_- \quad (28)$$

We have seen in the previous section that a simple finite difference scheme leads to very satisfactory results when computing the solution of the classical Black-Scholes equation (8). On the other hand, when one uses a simple finite difference scheme on (14), very bad results are obtained, especially when the volatility σ is small (see Table 1 in Dubois and Lelièvre, 2005). These bad results are due to the fact that when ξ is close to zero, the advection term ($\frac{1}{T} + r\xi$) is much larger than the diffusion term $\sigma^2 \xi^2 / 2$ in (14). This is known to deteriorate the stability of the numerical scheme, particularly with respect to the L^∞ -norm. In practice, the numerical solution exhibits some oscillations and does not satisfy the discrete maximum principle. Moreover, the finite difference method introduces numerical diffusion, which leads to unsatisfactory results for purely advective equations.

Table 2 Error on the Value of a Call in Function of the Number of Time-Steps N

$I = 500$	$N = 5$	$N = 10$	$N = 20$	$N = 40$	$N = 80$	$N = 160$
EE	28.53	0.386	0.398	0.0739	0.0162	0.00714
IE	0.0892	0.0449	0.0225	0.0113	0.00554	0.00226
CN	0.0299	0.00758	0.00103	0.00169	0.00169	0.00168

Note: We observe that the Euler explicit (EE) scheme is unstable for $N = 5$. The convergence in time of the Crank-Nicolson (CN) scheme is much faster than for the implicit Euler (IE) scheme. The remaining error when N is large is due to the discretization with respect to the variable S .

One way to handle this problem is to use a characteristic method (based on the solution of $d\xi/dt = -1/T$) in order to get rid of the term $1/T$. This means that the following change of variable is introduced:

$$g(t, x) = f(t, x - t/T) \tag{29}$$

One can easily show that g is solution of:¹

$$\begin{cases} \frac{\partial g}{\partial t} + \frac{\sigma^2(x - t/T)^2}{2} \frac{\partial^2 g}{\partial x^2} - r(x - t/T) \frac{\partial g}{\partial x} = 0, \\ g(T, x) = \tilde{\phi}(x - 1) = (1 - x)_+ \end{cases} \tag{30}$$

The PDE (30) satisfied by g is such that when the advection term $r(x - t/T)$ is small, the diffusion term $\frac{\sigma^2(x - t/T)^2}{2}$ is also small. As shown below, a finite difference scheme applied to (30) will indeed lead to satisfactory results.

An important property of the solution to (30) for $\tilde{\phi}(\xi) = \xi_-$ is that (see Rogers and Shi, 1995) $\forall \xi \leq 0$,

$$f(t, \xi) = \frac{1}{rT}(1 - e^{-r(T-t)}) - \xi e^{-r(T-t)} \tag{31}$$

and therefore, $\forall x \leq t/T$,

$$g(t, x) = \frac{1}{rT}(1 - e^{-r(T-t)}) - (x - t/T)e^{-r(T-t)} \tag{32}$$

To prove (31), one can notice that f given by (31) is the solution to (14) with $\phi(\xi) = -\xi$, and that, due to the fact that the diffusion term is null for $\xi = 0$ and that the advection term is negative, the solution to (14) for $\phi(\xi) = \xi_-$ on $\xi \leq 0$ is the same as the solution to (14) for $\phi(\xi) = -\xi$ on $\xi \leq 0$.

To discretize (30), a Crank-Nicolson time scheme is used, with a uniform time step $\delta t = T/N$. In order to use the fact that g is analytically known on $x \leq t/T$ (see (32)), a mesh that properly discretizes the boundary $x = t/T$ is used. Therefore, the space interval $(0, 1)$ is also discretized with N space steps of length $\delta x = 1/N$ (see Figure 1). The mesh is completed by adding J intervals on the right-hand side of $x = 1$, so that $x \in (0, x_{max})$ with $x_{max} = (N + J)\delta x$. The value $J = N/2$ has been found to be sufficient

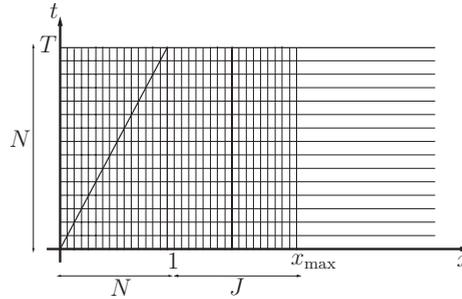


Figure 1 The Mesh and the Computational Domain for the Finite Difference Scheme Used to Discretize (30)

to guarantee the independence of the results on the position of x_{max} .

Notice that at time $t_n = n\delta t$, the number of unknowns is $(N + J - n)$. This means that the dimension of the linear system to solve depends on the time-step.

As far as boundary conditions are concerned, we use a Dirichlet boundary condition on $x = t/T$ (using (32)) and an artificial zero Neumann boundary condition on $x = x_{max}$.

Let us now give some numerical results. In Table 3, a few comparisons of the results obtained with the characteristic method and other methods are given. The characteristic method appears to be accurate for both small and large volatilities. For any values of the parameters, at least 5 digits of precision are obtained in less than one second. Notice that the Thompson bounds and the characteristic method are implemented in Premia.²

The Finite Element Method for European Options

We would like now to introduce the finite element method. This technique is more flexible than the finite difference method. In particular, it allows for local refinements of the spot grid (even in dimensions greater than one), and possibly based on local error estimators that are mentioned below. This is particularly important for American options, because the pricing function is singular near the exercise boundary, and

Table 3 Comparisons of the Prices for an Asian Fixed Call Obtained with Various *Finite Difference* Methods: Characteristic Method, Zvan et al. (1998), Večeř (2001), and Thompson (1999)

σ	K	Charact. Method	N:	Zvan et al.	Večeř	Thompson (low)	Thompson (up)
0.05	95	11.09409	(300)	11.094	11.094	11.094094	11.094096
	100	6.7943	(1000)	6.793	6.795	6.794354	6.794465
	105	2.7444	(3000)	2.748	2.744	2.744406	2.744581
0.30	90	16.512	(300)	16.514	16.516	16.512024	16.523720
	100	10.209	(300)	10.210	10.215	10.208724	10.214085
	110	5.730	(1000)	5.729	5.736	5.728161	5.735488

Note: Values of parameters: $T = 1$, $r = 0.15$, $S_0 = 100$, $J = N/2$. For the characteristic method, the number of time-steps $N \geq 300$ needed to obtain at least 5 digits of precision is given.

this curve is not known a priori. Let us emphasize that the use of a refined mesh around the singularities of the solution (for example, for vanilla option pricing problems, around $t = T$ and $S = K$) is very important in practice to rapidly obtain accurate results. The finite element method can also be used in a flexible way when the geometry of the computational domain becomes complex, which may be of interest for barrier options in dimensions greater than one. Finally, finite element methods are interesting since they are naturally stable (in L^2 -norm) and optimal error bounds (in L^2 -norm) can be derived.

In the following, we first present the finite element method on a simple example, namely equation (8). We then show how to treat more complex European options.

Variational Formulation and Finite Element Space

The conforming finite element method is based on two ingredients: a so-called variational formulation of the PDE on a functional space V and the choice of an appropriate sequence of finite dimensional spaces $V_h \subset V$, which tends to V when h (which is the typical diameter of the cells of the space mesh) tends to 0. Let us illustrate this on (8).

To derive a variational formulation of (8), the principle is to multiply the equation by a test function of the spot variable and to integrate by parts. For these computations to be

well defined, the functions need to be sufficiently smooth. We thus introduce the functional spaces $H = L^2(\mathbb{R}_+) = \{q : [0, \infty) \rightarrow \mathbb{R}, \int_0^\infty q^2 < \infty\}$, and $V = \{q \in L^2(\mathbb{R}_+), S(\partial q / \partial S) \in L^2(\mathbb{R}_+)\}$. Assuming that ϕ is square integrable, a variational formulation of (8) is then (for an S -dependent volatility σ): Find $p \in L^2((0, T), V) \cap C^0([0, T], H)$ such that for all $q \in V$,

$$\begin{cases} \frac{d}{dt} \int_0^\infty pq - \int_0^\infty \frac{\sigma^2 S^2}{2} \frac{\partial p}{\partial S} \frac{\partial q}{\partial S} \\ + \int_0^\infty \left(r - \sigma^2 - S\sigma \frac{\partial \sigma}{\partial S} \right) S \frac{\partial p}{\partial S} q \\ - r \int_0^\infty pq = 0, \\ p(T, S) = \phi(S) \end{cases} \quad (33)$$

All the integrals are with respect to $S \in [0, \infty)$. This rewrites: Find $p \in L^2((0, T), V) \cap C^0([0, T], H)$ such that for all $q \in V$,

$$\begin{cases} \frac{d}{dt} \int_0^\infty pq - a(p, q) = 0, \\ p(T, S) = \phi(S) \end{cases} \quad (34)$$

where a is the bilinear form

$$\begin{aligned} a(p, q) &= \int_0^\infty \frac{\sigma^2 S^2}{2} \frac{\partial p}{\partial S} \frac{\partial q}{\partial S} \\ &- \int_0^\infty \left(r - \sigma^2 - S\sigma \frac{\partial \sigma}{\partial S} \right) S \frac{\partial p}{\partial S} q \\ &+ r \int_0^\infty pq \end{aligned} \quad (35)$$

Under suitable assumptions on the data ($r, \sigma,$ and ϕ), it is possible to prove that this variational problem is well posed (see Achdou and Pironneau, 2005).

The second step is to introduce a sequence of meshes in the spot variable indexed by the maximal step h and related finite dimensional functional spaces $V_h \subset V$. In the case of (33), the problem is posed on an infinite domain, and one needs to first localize the PDE in a finite domain $[0, S_{\max}]$ by using artificial boundary condition at $S = S_{\max}$, as already explained for finite difference discretizations. We consider, for example, a zero Neumann boundary condition on $S = S_{\max}$: $\frac{\partial p}{\partial S}(t, S_{\max}) = 0$. Then, a mesh of $[0, S_{\max}]$ consists of a finite number of intervals (S_i, S_{i+1}) with $S_0 = 0$ and $S_I = S_{\max}$. We set $h = \max_{0 \leq i \leq I-1} (S_{i+1} - S_i)$. The intervals (S_i, S_{i+1}) are called elements. We then need to define a functional space V_h associated with the mesh. A classical example is the $P1$ finite element space, which contains continuous and piecewise affine functions, namely, continuous functions, which are affine on each interval (S_i, S_{i+1}) , for $0 \leq i \leq I - 1$. In this case, a basis of the vector space V_h is given by the so-called hat functions $q_i \in V_h$ such that for $0 \leq i, j \leq I$, $q_i(S_j) = \delta_{i,j} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$ ($\delta_{i,j}$ is the Kronecker symbol). Notice that higher order finite element methods may be easily obtained by taking continuous and element-wise polynomial functions of degree $k > 1$.

The discretization in the spot price variable now simply consists in replacing the functional space V by the finite dimensional space V_h in (33) or (34) (this is the principle of Galerkin methods): Find $p_h \in C^0([0, T], V_h)$ such that for all $q_h \in V_h$,

$$\begin{cases} \frac{d}{dt} \int_0^{S_{\max}} p_h q_h - a(p_h, q_h) = 0, \\ p_h(T, S) = \phi_h(S) \end{cases} \quad (36)$$

where ϕ_h is an approximation of ϕ in the space V_h , and where the integrals in the bilinear form

a are here for $S \in [0, S_{\max}]$ (see (35)). One can take, for example, ϕ_h such that $\int_0^{S_{\max}} (\phi - \phi_h) q_h = 0$ for all $q_h \in V_h$ (ϕ_h is then the L^2 projection of ϕ onto V_h). Problem (36) is a finite-dimensional problem in space of the form $M_h dp_h/dt = A_h P_h$, where $P_h(t)$ is a vector of dimension I containing the values of p_h at the nodes of the mesh ($p_h(t, x) = \sum_{j=0}^I P_{h,j}(t) q_j(x)$) and M_h, A_h are $I \times I$ matrices. The matrix M_h (resp. A_h), with (i, j) -th component $\int_0^{S_{\max}} q_i q_j$ (resp. $a(q_j, q_i)$) is classically called the mass (resp. stiffness) matrix, because the finite element method was originally popularized by the mechanical engineering community. When using the nodal basis (hat functions), these matrices are very sparse (tridiagonal for one-dimensional problems). Problem (36) is somewhat similar to (24) obtained by the finite difference method; the two problems (24) and (36) are actually equivalent if a mesh with uniform space steps is used, and if M_h is replaced by a close diagonal matrix (mass-lumping).

A fundamental result (the Cea's lemma) states that the norm of $(p - p_h)$ (the discretization error) is bounded from above by a constant times the infimum of the norm of $(p - q_h)$, over all $q_h \in V_h$ (the best fit error). Using this result, if V_h gets closer to V when h tends to 0, that is, if the best fit error tends to 0 when h tends to zero, so does the discretization error. In particular, the finite element discretization is thus naturally stable in this norm. A precise meaning for this statement requires us to define the norm and study the best fit error. Let us simply mention that the norms used in this context are related to the L^2 -norm introduced for finite difference schemes. We refer to Achdou and Pironneau (2005) or Quarteroni and Valli (1997) for the details. In our specific example, it is possible to prove that, if the payoff function is regular enough, then

$$\|p - p_h\|_{L^\infty([0,T],H)} + \|p - p_h\|_{L^2([0,T],V)} \leq Ch$$

and that

$$\|p - p_h\|_{L^2([0,T],H)} \leq Ch^2$$

For the discretization in time, the situation is exactly the same as for the finite difference method: One can use the explicit Euler scheme, implicit Euler scheme, or Crank-Nicolson scheme, and the rate of convergence is $O(\delta t)$ for the Euler schemes and $O(\delta t^2)$ for the Crank-Nicolson scheme.

Finite Element Methods for Other Options

We have introduced the finite element method in a very simple case. The aim of this section is to explain how it applies for other options.

Let us first consider basket options, or basket options with barriers, in dimension 2 and 3. The derivation of a variational formulation for (9) is very similar to the one-dimensional case. However, the construction of the mesh is much more complicated in dimension 2 and 3, than in dimension 1. It consists of partitioning the domain into non-overlapping cells (elements) whose shapes are simple and fixed (for example, triangles or quadrilaterals in dimension 2, or tetrahedra or hexahedra in dimension 3). The functional spaces V_h can then be constructed as in dimension 1, for example, by considering continuous piecewise affine functions. One interest of the finite element method in this context is that it is possible to mesh any domain \mathcal{D} for barrier options. In the finite difference method, to mesh nonquadrilateral (or nonhexahedral) domains is complicated.

Let us now consider the case of lookback options whose prices satisfy (11). This is a natural variational formulation of (11) (written here for a constant volatility σ): Find $p : \mathcal{D} \rightarrow \mathbb{R}$ such that, for all $q : \mathcal{D} \rightarrow \mathbb{R}$,

$$\begin{aligned} & \frac{d}{dt} \int_{\mathcal{D}} pq - \int_{\mathcal{D}} \frac{\sigma^2 S^2}{2} \frac{\partial p}{\partial S} \frac{\partial q}{\partial S} - \int_{\mathcal{D}} \frac{\sigma^2 S^2}{2} \frac{\partial p}{\partial S} \frac{\partial q}{\partial M} \\ & + \int_{\mathcal{D}} \frac{\sigma^2 S^2}{2} \frac{\partial p}{\partial M} \frac{\partial q}{\partial S} + \int_{\mathcal{D}} \sigma^2 S \frac{\partial p}{\partial M} q \\ & + \int_{\mathcal{D}} (r - \sigma^2) S \frac{\partial p}{\partial S} q - r \int_{\mathcal{D}} pq = 0, \\ & p(T, S, M) = \phi(S, M) \end{aligned} \tag{37}$$

where $\mathcal{D} = \{(S, M) \in \mathbb{R}^2, 0 \leq S \leq M\}$. The boundary condition $\partial p / \partial M(t, S, S) = 0$ is naturally contained in this variational formulation since, by integration by parts over \mathcal{D} :

$$\begin{aligned} & - \int_{\mathcal{D}} \frac{\sigma^2 S^2}{2} \frac{\partial p}{\partial S} \frac{\partial q}{\partial S} - \int_{\mathcal{D}} \frac{\sigma^2 S^2}{2} \frac{\partial p}{\partial S} \frac{\partial q}{\partial M} \\ & + \int_{\mathcal{D}} \frac{\sigma^2 S^2}{2} \frac{\partial p}{\partial M} \frac{\partial q}{\partial S} + \int_{\mathcal{D}} \sigma^2 S \frac{\partial p}{\partial M} q \\ & - \int_{\mathcal{D}} \sigma^2 S \frac{\partial p}{\partial S} q = \int_{\mathcal{D}} \frac{\sigma^2 S^2}{2} \frac{\partial^2 p}{\partial S^2} q \\ & + \frac{1}{\sqrt{2}} \int_{\{S=M\}} \frac{\sigma^2 S^2}{2} \frac{\partial p}{\partial M} q \end{aligned}$$

The first term corresponds to the diffusion term in (11). The second term is an integral over the boundary $\{S = M\}$ of \mathcal{D} and naturally enforces the boundary condition $\partial p / \partial M(t, S, S) = 0$. In Figure 2, we represent the price of a fixed strike call obtained using the formulation (11), an implicit Euler scheme, and $P1$ finite elements. The computations are made with FreeFem++.³

A Posteriori Error Estimates

A frequently mentioned advantage of the Monte Carlo methods is that they naturally provide a posteriori error bounds through a confidence interval, typically built upon the central limit theorem. It is also possible to obtain such a posteriori error estimates in the framework of the finite element method (this is one additional advantage of this method compared to finite difference methods). Moreover, these a posteriori estimates have two very important features:

- They depend on local error indicators.
- They can be proved to be reliable and efficient, that is, the actual error is bounded above and below by some fixed constants times the a posteriori error, and these estimates can be made local.

Therefore, in the finite element method, the a posteriori error estimates enable us to refine the mesh in space and time adaptively. We will

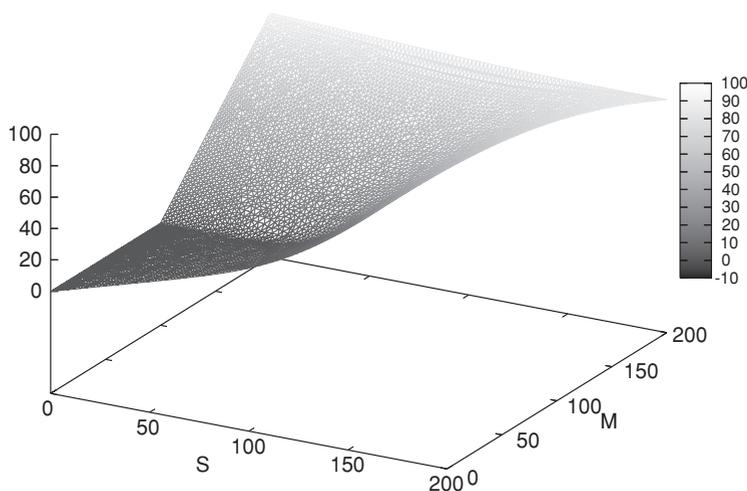


Figure 2 Price of a Lookback Option for a Fixed Strike Call: $\phi(S, M) = (M - K)_+$
Note: The parameters are: $\sigma = 0.3$, $r = 0.1$, $K = 100$, $T = 1$.

give a numerical illustration for American options and refer to Ern, Villeneuve, and Zanette (2004), Achdou and Pironneau (2005 and 2009), or Achdou, Hecht, and Pommier (2008) for more details.

High-Dimensional Problems

In practical problems, options often involve more than three assets. In this case, the PDE is posed in a space of dimension larger than 4, and the finite element or difference methods cannot be used, since the number of unknowns typically grows exponentially with respect to the problem's dimension. This is the so-called curse of dimensionality. Let us mention that such high-dimensional problems also appear in other scientific fields, quantum chemistry, for example, and that it is still a subject of current research to build appropriate discretizations for high-dimensional PDEs. Roughly speaking, the problem is to find an appropriate sequence of functional spaces V_h (whose basis is called a Galerkin basis), such that their dimensions do not grow too rapidly with the dimension of the problem. One approach is the sparse tensor product (see Bungartz and Griebel, 2004; Petersdoff and Schwab, 2004). The main difficulty when using this approach is actually

to project the initial condition on V_h . Another approach used in other contexts for solving high-dimensional problems by deterministic methods is the low separation rank method (see Beylkin and Mohlenkamp, 2002) and the related greedy algorithms (see Ammar et al., 2002; Temlyakov, 2008; Le Bris, Lelièvre, and Maday, 2009; and Nouy, 2009). Let us finally mention that another possible approach for building an appropriate Galerkin basis would be the reduced basis method, where some solutions for a given set of parameters are used to approximate the solution for other values of the parameters. Such methods are currently actively investigated (see, for example, Boyaval, Le Bris, Lelièvre, Maday, Nguyen, and Patera, 2010).

The Uncertain Volatility Model: An Example of a Nonlinear PDE

One major interest of the PDE approach is that it can be applied for nonlinear models. This will be the case for American options, see below, but we would like to give here another example of such a situation. The principle of the uncertain volatility model introduced by Avellaneda, Levy, and Paras (1995) is to give a price for a European option, when the volatility is only supposed to be in an interval $[\sigma_{\min}, \sigma_{\max}]$. The

principle is the following. For a European option with convex payoff, it is easy to check that the price should be the Black-Scholes price obtained with the maximum volatility σ_{\max} . In this case, the profit and loss for the hedging strategy is indeed zero if the realized volatility is constant equal to σ_{\max} . A similar reasoning holds for concave payoffs: In this case, one should consider the Black-Scholes price with the minimum volatility σ_{\min} . For a general payoff, it is thus natural (and it can be checked that this is indeed an approach that leads to a very good hedging strategy, with small profit and loss, and thus cheap price) to consider the solution p to the PDE:

$$\begin{cases} \frac{\partial p}{\partial t} + rS \frac{\partial p}{\partial S} + \\ \frac{1}{2} \left(\sigma_{\max}^2 \mathbf{1}_{\frac{\partial^2 p}{\partial S^2} \geq 0} + \sigma_{\min}^2 \mathbf{1}_{\frac{\partial^2 p}{\partial S^2} < 0} \right) S^2 \frac{\partial^2 p}{\partial S^2} - rp = 0, \\ p(T, S) = \phi(S) \end{cases} \quad (38)$$

In other words, σ_{\max} (resp. σ_{\min}) is used where the price is convex (resp. concave), as a function of the spot. This PDE can be solved using extensions of the discretization techniques presented above; see, for instance, section 2.4 in van der Pijl and Oosterlee (2011).

PRICING AMERICAN OPTIONS WITH PDES

This section is devoted to the discretization of the system (23) for the price of an American option. Notice that no closed formulas such as the Black-Scholes formula are available for American put, or for American call with a dividend rate, so that efficient discretization of this system is needed even for these simple payoffs.

The Finite Difference Approach for American Options

We first present the extension of the finite difference approach presented above for European options to American options.

Some Finite Difference Schemes

We consider a regular mesh discretization $S_i = i\delta S$ and a time discretization $t_n = n\delta t$ with $\delta t = \frac{T}{N}$. As in the European case, it is natural to consider the following three iterative numerical schemes for P_i^n , an approximation of $p(t_n, S_i)$. In all cases, the scheme is initialized by $P_i^N = \phi(S_i)$. Let A be the matrix such that

$$(AP^{n+1})_i = -rS_i \frac{P_{i+1}^{n+1} - P_{i-1}^{n+1}}{2\delta S} - \frac{\sigma^2 S_i^2}{2} \frac{P_{i+1}^{n+1} - 2P_i^{n+1} + P_{i-1}^{n+1}}{\delta S^2} + rP_i^{n+1} \quad (39)$$

The explicit Euler (EE) scheme for (23) is, for $n = N - 1, N - 2, \dots, 0$,

$$\min \left(-\frac{P_i^{n+1} - P_i^n}{\delta t} + (AP^{n+1})_i, P_i^n - \phi(S_i) \right) = 0 \quad (40)$$

The scheme computes $P^n = (P_i^n)_{i=0, \dots, I-1}$ from the knowledge of $P^{n+1} = (P_i^{n+1})_{i=0, \dots, I-1}$. Similarly, we can propose an implicit Euler (IE) scheme:

$$\min \left(-\frac{P_i^{n+1} - P_i^n}{\delta t} + (AP^n)_i, P_i^n - \phi(S_i) \right) = 0 \quad (41)$$

and an (implicit) Crank-Nicolson (CN) scheme

$$\min \left(-\frac{P_i^{n+1} - P_i^n}{\delta t} + \frac{1}{2}((AP^n)_i + (AP^{n+1})_i), P_i^n - \phi(S_i) \right) = 0 \quad (42)$$

In the case of the EE scheme, it is easy to see that we have the equivalent formulation

$$P_i^n = \max \left(((I_d - \delta t A)P^{n+1})_i, \phi(S_i) \right) \quad (43)$$

where I_d denotes the identity matrix.

We now have two new difficulties compared to the European case: First, the well-posedness of the schemes (41) or (42) is not immediate (for European options, we obtained a linear system, but this is no longer true for American options), and second, studying the convergence is more difficult.

One way to circumvent the first difficulty is to introduce a splitting method (see Barles and Souganidis, 1991; Barles, Daher, and Romano, 1995; and Lions and Mercier, 1979). For (23), it writes (a similar modification of (42) could also be considered, yielding a Crank Nicolson-splitting (CN-S) scheme):

$$\text{compute } P^{n,1} \text{ s.t. } -\frac{P_i^{n+1} - P_i^{n,1}}{\delta t} + (AP^{n,1})_i = 0 \tag{44a}$$

$$\text{and then compute } P_i^n = \max(P_i^{n,1}, \phi(S_i)) \tag{44b}$$

Hereafter, (44) will be referred to as the implicit Euler-splitting (IE-S) scheme. The first step (44a) consists of solving a linear system, as in the European case. The second step is a projection on the set $\{v = (v_i), v_i \geq \phi(S_i), \forall i\}$, as for the EE scheme (43).

Notice that as for European options, we set the equation on a truncated domain $(0, S_{\max})$ and use artificial boundary conditions on $S = S_{\max}$. We refer to Barles, Daher, and Romano (1995) for error estimates between the truncated problem on $(0, S_{\max})$ and the exact problem.

An Abstract Convergence Result

Assuming for the moment that the schemes are well posed, it is possible to study the convergence in the general framework of finite different schemes for Hamilton-Jacobi equations. Possibly under some restrictions on the mesh sizes δt and δS , we can obtain convergence to the viscosity solution of the PDE (23). We refer to Barles (1994) or Barles, Daher, and Romano (1995) for a short introduction, and Crandall, Ishii, and Lions (1992) for a more detailed overview. To give a rough idea of the convergence results for such schemes, we consider a general Hamilton-Jacobi equation of the form

$$H\left(t, S, p, \frac{\partial p}{\partial t}, \frac{\partial p}{\partial S}, \frac{\partial^2 p}{\partial S^2}\right) = 0 \tag{45}$$

with a terminal condition on $p(T, \cdot)$, where H is assumed to be Lipschitz continuous and

“backward parabolic” in the sense that

$$\begin{aligned} \text{if } \psi_1 \leq \psi_2 \text{ then } H(t, S, p, u, v, \psi_1) \\ \geq H(t, S, p, u, v, \psi_2) \end{aligned} \tag{46a}$$

$$\begin{aligned} \text{and if } u_1 \leq u_2 \text{ then } H(t, S, p, u_1, v, \psi) \\ \geq H(t, S, p, u_2, v, \psi) \end{aligned} \tag{46b}$$

Equation (23) is indeed of the form of (45) with, for $(t, S) \in (0, T) \times (0, S_{\max})$, $H(t, S, p, u, v, \psi) = \min(-u - rSv - \frac{1}{2}\sigma^2 S^2 \psi + rp, p - \phi(S))$, which obviously satisfies (46).

First convergence results were given in the fundamental work of Crandall and Lions (1984) for Lipschitz continuous final condition ϕ (and without $\frac{\partial^2 p}{\partial S^2}$ dependence in (45)).

An abstract and general convergence result is given by Barles and Souganidis (1991), and we now give a simplified presentation of this result.

We first assume that H satisfies a comparison principle, which can be seen as an extension of the maximum principle to some nonlinear equations. The comparison principle is roughly the following (see Crandall, Ishii, and Lions, 1992; Barles, 1994; or Pham, 1998): Assume that u (resp. v) is a subsolution (resp. supersolution) of (45), that is,

$$\begin{aligned} H\left(t, x, u, \frac{\partial u}{\partial t}, \frac{\partial u}{\partial x}, \frac{\partial^2 u}{\partial x^2}\right) \leq 0 \\ \left(\text{resp. } H\left(t, x, v, \frac{\partial v}{\partial t}, \frac{\partial v}{\partial x}, \frac{\partial^2 v}{\partial x^2}\right) \geq 0\right) \end{aligned}$$

for $(t, S) \in (0, T) \times (0, S_{\max})$, and that $u \leq v$ on the boundaries $S = S_{\max}$ and $t = T$, then $u \leq v$ everywhere.

Now, suppose that we can write the scheme in the abstract form: $\forall i \in \{0, \dots, I\}, \forall n \in \{0, \dots, N\}$,

$$S_\rho(t_n, S_i, P_i^n, [P]) = 0 \tag{47}$$

where $\rho = (\delta t, \delta S)$, and $[P]$ stands for a continuous function that takes values $(P_j^k)_{0 \leq k \leq N, 0 \leq j \leq I}$ on the corresponding grid points (t_k, S_j) .⁴ We suppose that (47) admits at least one solution denoted P_ρ . Then, in the limit when ρ goes to

zero, P_ρ converges to p solution to (45) if the following conditions are satisfied:

- (i) A stability condition, which reads $\max_{0 \leq n \leq N, 0 \leq i \leq I} |P_i^n| \leq C$, for some constant C independent of N and I (i.e., independent of ρ).
- (ii) A consistency condition: for any regular function ψ ,

$$\begin{aligned} & \lim_{\xi \rightarrow 0, \rho \rightarrow 0, t_n \rightarrow t, S_i \rightarrow S} \mathcal{S}_\rho(t_n, S_i, \psi(t_n, S_i) + \xi, \psi + \xi) \\ &= H\left(t, S, \psi, \frac{\partial \psi}{\partial t}, \frac{\partial \psi}{\partial S}, \frac{\partial^2 \psi}{\partial S^2}\right)(t, S) \end{aligned}$$

For a weaker statement see Barles and Souganidis (1991).

- (iii) A monotonicity condition, which reads

$$\varphi \leq \psi \Rightarrow \mathcal{S}_\rho((t, S), P, \varphi) \geq \mathcal{S}_\rho((t, S), P, \psi)$$

For most standard financial options, a comparison principle holds. The stability and consistency conditions are close to the stability and consistency conditions already introduced in the case of the schemes for European options. Hence the new condition to check is the monotonicity assumption (which is related to the property (46a) satisfied by H). It is actually related to a discrete maximum principle.

Error estimates can also be obtained for the finite difference schemes (40)–(41)–(42). For example, for the EE scheme, an error estimate of order $\delta S^{1/2}$ in L^∞ -norm can be proved under a CFL condition and for Lipschitz initial data (see Jackobsen, 2003). In the context of the finite element method (see below) an error estimate of order δS^2 can be proved, but in the weaker L^2 -norm.

Implementation and Convergence of the Finite Difference Schemes

It is easy to see, in view of (43), that the EE scheme is stable and monotone if the components of the matrix $(I_d - \delta t A)$ are nonnegative. This is exactly what is needed to prove a discrete maximum principle in the European case.

This property holds under a CFL condition of the form $\delta t \leq C \delta S^2$, C constant, and with an appropriate discretization of the advection term. The CN scheme is also stable and monotone under a CFL-like condition. On the other hand, it can be shown that the IE-S scheme as well as the IE scheme are stable and monotone without any CFL condition.

Now let us explain how to solve the implicit schemes (41) or (42) in practice. Let us consider the IE scheme (41). At each time step, setting $b = P^{n+1}$, $B = I_d + \delta t A$ and $g = (\phi(S_i))_i$, the problem is equivalent to finding $x = P^n$ such that

$$\min((Bx - b)_i, (x - g)_i) = 0, \quad \forall i \quad (48)$$

The Howard algorithm (see Howard, 1960; also called the policy iteration algorithm) is the method of choice to solve (48). To present this algorithm, we rewrite (48) in the following form: Find x such that,

$$\min_{\alpha \in \{0,1\}^I} ((B(\alpha)x - b(\alpha))_i) = 0, \quad \forall i \quad (49)$$

where $B_{i,j}(\alpha) = \begin{cases} B_{i,j} & \text{if } \alpha_i = 0 \\ \delta_{i,j} & \text{if } \alpha_i = 1 \end{cases}$ (where $\delta_{i,j}$ is again the Kronecker symbol, i.e., the (i, j) -th component of I_d) and $b_i(\alpha) = \begin{cases} b_i & \text{if } \alpha_i = 0 \\ g_i & \text{if } \alpha_i = 1 \end{cases}$. The i -th component of $B(\alpha)x - b(\alpha)$ only depends on the i -th component of α , so that the minimum for the i -th component in (49) is indeed taken with respect to the i -th component of α . Thus, for a given x and α realizing the minimum in (49), the component α_i is equal to 0 (resp. to 1) if, at the i -th node, the minimum in (48) is $(Bx - b)_i$ (resp. $(x - g)_i$). For an initial value $\alpha^0 \in \{0, 1\}^I$, the algorithm is written as follows: Iterate for $k \geq 0$,

- (i) Compute x^k such that $B(\alpha^k)x^k = b(\alpha^k)$
- (ii) $\alpha_i^{k+1} = \arg \min_{\alpha_i \in \{0,1\}} (B(\alpha)x^k - b(\alpha))_i$

Santos and Rust (2004) and Bokanowski, Maroso, and Zidani (2009) provide some convergence results. Under suitable assumptions

on the matrix B (which are satisfied for the schemes considered above, which satisfy the monotonicity condition), it can be proved that this method converges in at most I iterations. In practice, only a few iterations are needed for solving (41).

This algorithm can also be seen as:

- A Newton’s method on the function F defined by $F_i(x) = \min((Bx - b)_i, (x - g)_i)$. More precisely, it is a semismooth Newton’s method applied to the slantly differentiable function⁶ F .
- A primal-dual algorithm to implement the fully implicit Euler scheme (41) as introduced in Hintermüller, Ito, and Kunisch (2002).

Another well-known method for solving (48) is the projected successive over relaxation (PSOR) method, which is a modification of the successive over relaxation (SOR) method to solve iteratively systems of linear equations (see Saad, 2003). In its simplest form, it consists of decomposing $B = L + U$ where L is a lower triangular matrix and U is an upper triangular matrix with zero coefficients on the diagonal. The algorithm consists of choosing an initial guess x^0 and then computing iteratively for $n \geq 1$, for $i = 1, \dots, I$, $x_i^n = \arg \min \{(Lx^{n-1} - (b - Ux^{n-1}))_i, (x^{n-1} - g)_i\}$. This method converges only if B is a diagonal dominant matrix, and the convergence is rather slow in practice. However, it leads to a highly efficient method for the finite element method discussed below, when combined with a suitable splitting scheme.

For the specific case of an American put option on a single asset, a fast algorithm was proposed by Brennan and Schwartz (1977) for solving (41) and proved to converge in Jaillet, Lamberton, and Lapeyre (1990) in the finite element setting (see also Bokanowski, Maroso, and Zidani [2009] in the finite difference setting). Also in this case it can be proved that the region of exercise (namely $\Gamma_t = \{x \in \mathbb{R}_+, p(t, x) > \phi(x)\}$) is of the form $\Gamma_t = [\gamma(t), \infty[$ where γ is continuous under some regularity assumption

Table 4 Error on the Value of an American Put in Function of the Number I of Intervals in the Variable S (and for $N = 1000$)

$(N = 1000)$	$I = 100$	$I = 200$	$I = 400$	$I = 800$	$I = 1600$
IE-S	0.00267	0.0361	0.00180	0.00210	0.00210
IE	0.00379	0.0146	0.00011	0.00024	0.00018

of the data. Then the Howard algorithm takes a simple form, which is known as the front-tracking algorithm (see, for instance, Achdou and Pironneau, 2005). However, these algorithms are very specific to the one-dimensional case and do not apply for general payoff functions.

Numerical Results for the American Put Option

In Table 4, we give numerical results obtained with the IE-S and IE schemes for an American put option with $r = 0.1$, $\sigma = 0.1$, $K = 100$, $T = 1$, $S = 100$, and $S_{\max} = 150$. We have computed all error values by taking the reference value $P = 1.63380$ (obtained with a Cox-Ross-Rubinstein algorithm with $N = 10^5$, CPU-time = 1750 s.; see Cox, Ross, and Rubinstein [1979] and Lamberton and Lapeyre [1997]). In this example, the IE scheme is one digit more accurate than the IE-S scheme. With these numerical parameters, the EE scheme would yield bad results since the CFL condition is not respected. The IE scheme has been implemented using the Howard algorithm. The remaining error when I is large is due to the time discretization.

In Table 5, we compare the EE, IE-S and IE schemes. Since the error is of order of $O(\delta t) + O(\delta S^2)$, we have used parameters N and I such that $\delta t \simeq \delta S^2$ (i.e., $N \simeq I^2$), and such that the CFL condition is satisfied. We remark that the EE scheme gives satisfactory results in less than a few seconds. The IE is more accurate but more costly than IE-S. Hence in view of the CPU-time it is more advantageous here to use simply the EE or the IE-S scheme. This conclusion holds for a finite difference discretization, but may be

Table 5 Error and CPU-Times for the Value of an American Put as a Function of the Number N of Time-Steps N and the Number I of Intervals in the Variable S

	$I = 100$ $N = 100$	$I = 200$ $N = 400$	$I = 400$ $N = 1600$	$I = 800$ $N = 6400$	$I = 1600$ $N = 25000$
EE	0.00593	0.00069	0.00045	0.00003	0.00003
CPU-time (sec.)	0.01	0.10	0.5	2.6	10.7
IE-S	0.01177	0.00616	0.00098	0.00029	0.00007
CPU-time (sec.)	0.05	0.22	1.23	7.06	44.31
IE	0.00201	0.00181	0.00016	0.00004	0.00001
CPU-time (sec.)	0.2	0.9	7.3	75.0	1033.0

different for a finite element discretization, or for another set of parameters.

Markov Chains Approximations

There exist related discretization schemes for American options based on Markov chain approximations. Markov chain schemes (see Kushner and Dupuis, 2001) are based on the approximation of the dynamic programming principle between times t and $t + \delta t$ and on the use of a spatial interpolation over a mesh (S_j). This leads to another class of schemes that are also in finite difference form. Their convergence can be established by showing the convergence to the dynamic programming equation, or by using the Barles-Souganidis theorem (see Barles and Souganidis, 1991). Finite difference schemes enter this framework as well as semi-Lagrangian schemes (Capuzzo-Dolcetta and Falcone, 1989; Falcone and Ferretti, 1994). An inversed CFL condition, typically of the form $\delta S^2 / \delta t \xrightarrow{\delta t, \delta S \rightarrow 0} 0$ can then be needed. Notice that the Cox-Ross-Rubinstein algorithm (Cox, Ross, and Rubinstein, 1979) can also be seen as a discrete Markov chain approximation scheme using a very particular spatial mesh such that no interpolation appears at the end.

Portfolio Optimization

The techniques developed above for pricing American options can be used in the context of portfolio optimization. A portfolio op-

timization problem (or stochastic optimization problem) is typically of the form

$$p(t, x) = \max_{\alpha \in L^\infty([t, T], K)} \mathbb{E} \left(\int_t^T e^{-\int_t^u r(s) ds} \times f(u, X_u^{t, x, \alpha}, \alpha(u)) du + e^{-\int_t^T r(s) ds} \phi(X_T^{t, x, \alpha}) \right) \quad (50)$$

where K is compact, α is a progressively measurable function with values in K , and with

$$\begin{cases} dX_u^{t, x, \alpha} = b(u, X_u^{t, x, \alpha}, \alpha(u)) du \\ \quad + \sigma(u, X_u^{t, x, \alpha}, \alpha(u)) dW_u, \quad u \geq t, \\ X_t^{t, x, \alpha} = x \end{cases}$$

The corresponding PDE can be shown to be

$$\min_{\alpha \in K} \left(-\frac{\partial p}{\partial t} - \frac{1}{2} \sigma^2(t, S, \alpha) \frac{\partial^2 p}{\partial S^2} - b(t, S, \alpha) \frac{\partial p}{\partial S} + rp - f(t, S, \alpha) \right) = 0$$

in the viscosity sense (see Pham, 2006). Finite difference schemes similar to those presented above for American options can be applied. Implicit schemes, if considered, can be solved by the Howard algorithm. This can also be generalized to an optimal stopping time problem, adding in (50) a supremum over stopping times with values in $[t, T]$ (as in (15)). For such general HJB equations, a discretization by a finite element approach is not always possible because an appropriate variational formulation cannot always be obtained; see Bensoussan and Lions (1978).

The Finite Element Approach for American Options

As in the European case, the finite element approach requires a variational formulation of the PDE (23). Let us consider the case of the American put option. Let V be the functional space used for the variational formulation, and

$$K = \{q \in V, q \geq \phi\}$$

We first notice that (23) is equivalent to the set of inequalities⁷ (together with $p(T, S) = \phi(S)$)

$$\begin{cases} p - \phi \geq 0, \\ -\frac{\partial p}{\partial t} + \mathcal{A}p \geq 0, \\ \left(-\frac{\partial p}{\partial t} + \mathcal{A}p\right)(p - \phi) = 0 \end{cases} \quad (51)$$

We can check that this is equivalent (for sufficiently smooth function p) to the following variational formulation for (23): find $p \in L^2([0, T], K) \cap C^0([0, T], L^2(\mathbb{R}_+))$ such that for all $t \in [0, T)$,

$$\forall q \in K, -\int \frac{\partial p}{\partial t}(q - p) + a(p, q - p) \geq 0 \quad (52)$$

where a is the bilinear form (35) defined above (recall that for compactly supported functions p and q , $a(p, q) = \int \mathcal{A}p q$), with the final condition

$$p(T, S) = \phi(S)$$

Indeed, by writing $q - p = (q - \phi) - (p - \phi)$, it is clear that (51) implies (52). Conversely, choosing a sufficiently large $q \in K$ in (52), we obtain that $-\frac{\partial p}{\partial t} + \mathcal{A}p \geq 0$. Taking then $q = \phi$ in (52), we obtain that $\langle -\frac{\partial p}{\partial t} + \mathcal{A}p, \phi - p \rangle \geq 0$, but this inequality is actually an equality since $-\frac{\partial p}{\partial t} + \mathcal{A}p \geq 0$ and $\phi - p \leq 0$.

Notice that if we take $K = V$ in (52), we recover the variational formulation (34) for the European option. Precise existence and uniqueness results for such variational inequalities can be found in Bensoussan and Lions (1978). For results and applications in the finance context,

we refer to Achdou and Pironneau (2005 and 2009).

Now, as in the case of the finite element method for European options, we introduce a sequence of finite dimensional functional spaces $V_h \subset V$, such that the functions in V are better and better approximated by functions in V_h as h goes to 0. One can, for example, consider a P1 finite element space on a mesh $(S_i)_{0 \leq i \leq I}$. Remember that a basis of V_h is given by a set of functions $(q_i)_{0 \leq i \leq I}$. The finite element approximation of (52) is obtained by replacing V by V_h : Find $p_h \in C^0([0, T], K \cap V_h)$ such that for all $t \in [0, T)$,

$$\forall q_h \in K \cap V_h, -\int \frac{\partial p_h}{\partial t}(q_h - p_h) + a(p_h, q_h - p_h) \geq 0 \quad (53)$$

with the final condition $p_h(T) = \phi_h$, where $\phi_h \in V_h$ is an approximation of ϕ .

For time discretization, one can again use the schemes we have introduced in the case of the discretization of European options. For example, the implicit Euler scheme applied to (53) is naturally defined as follows: Find $p_h^N, p_h^{N-1}, \dots, p_h^0$ in $V_h \cap K$ such that $p_h^N = \phi_h$ (initialization) and, for $n = N - 1, \dots, 0$:

$$\forall q_h \in V_h \cap K, -\int \frac{p_h^{n+1} - p_h^n}{\delta t}(q_h - p_h^n) + a(p_h^n, q_h - p_h^n) \geq 0 \quad (54)$$

One can easily check that

$$q_h \in V_h \cap K \Leftrightarrow q_h \in V_h \text{ and } q_h(S_i) \geq \phi(S_i), \forall i$$

Denoting A_h and M_h the mass and stiffness matrices as in the case of the finite element method for European options, and reasoning as for the equivalence between (23), (51), and (52), it can be checked that (54) is equivalent to solve in \mathbb{R}^I :

$$\min \left(\left(-M_h \frac{P^{n+1} - P^n}{\delta t} + A_h P^n\right)_i, (P^n - g)_i \right) = 0, \quad \forall i$$

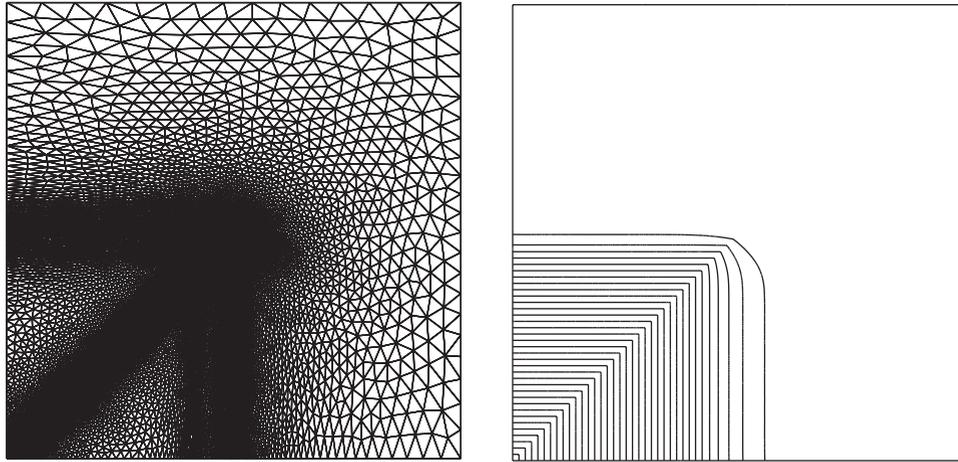


Figure 3 The Adapted Mesh and the Contours of P One Year to Maturity. $\sigma_1 = 0.2, \sigma_2 = 0.1, \rho = -0.6$.

where $g_i = \phi(S_i)$ and $P_i^n = p_h^n(S_i)$. Equivalently, the problem is to find P^n such that

$$\min \left((M_h + \delta t A_h) P^n - M_h P^{n+1} \right)_i, (P^n - g)_i = 0, \quad \forall i$$

This is a similar problem as for the IE finite difference scheme (see (48)) where the matrix $(I_d + \delta t A)$ is now replaced by $(M_h + \delta t A_h)$. It can be solved by the Howard algorithm previously presented. For the particular American put problem under some assumptions on the mesh steps δt and h , it can also be solved by the Brennan and Schwartz algorithm or the front-tracking algorithm mentioned above.

Notice that a Crank-Nicolson scheme can be derived in a similar way. The expected error (in L^2 -norm) is (as in the European case) $O(h^2) + O(\delta t)$ for the IE scheme and $O(h^2) + O(\delta t^2)$ for the CN scheme.

We conclude this section by a numerical illustration of the mesh refinement procedure (that can be implemented by using a posteriori error estimates) applied to the pricing of an American option on two assets. Such an automatic refinement procedure is particularly useful for American options because the pricing function is not smooth at any given time $t \in [0, T]$. Figures 3 and 4 illustrate such a mesh refinement for a typical two-assets American option with

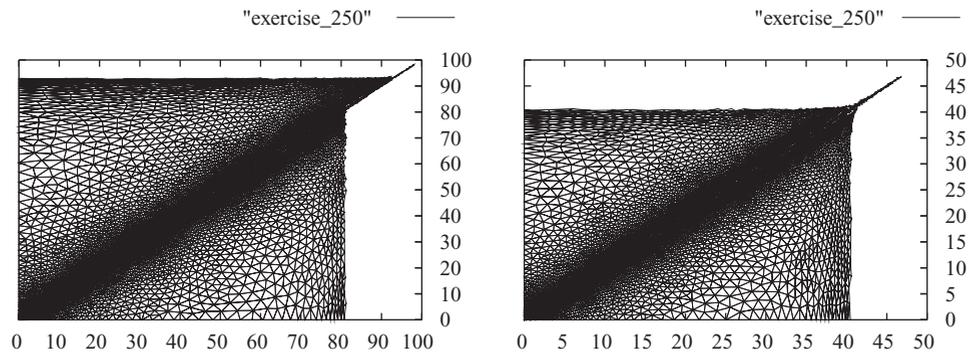


Figure 4 The Exercise Region One Year to Maturity. Left: $K = 100, \sigma_1 = 0.2, \sigma_2 = 0.1, \rho = -0.6$. Right: $K = 50, \sigma_1 = \sigma_2 = 0.2, \rho = 0$.

payoff $\phi(S_1, S_2) = (K - \max(S_1, S_2))_+$. The artificial boundary Γ_0 is $\{\max(S_1, S_2) = \bar{S} = 200\}$. Homogeneous Dirichlet conditions are imposed on Γ_0 . We have chosen two examples. In the first example, the parameters are $\sigma_1 = 0.2$, $\sigma_2 = 0.1$, $r = 0.05$, $\rho = -0.6$, and $K = 100$. In the second example, the parameters are $\sigma_1 = \sigma_2 = 0.2$, $r = 0.05$, $\rho = 0$, and $K = 50$. The implicit Euler scheme has been used with a uniform time step of $1/250$ year. For the variables S_1 and S_2 , we have used adaptive finite elements. For solving the linear complementarity problems, we have used the Howard algorithm. Mesh adaption in the (S_1, S_2) variable has been performed every $1/10$ year. In Figure 3, we have plotted the adapted mesh (left) and the contours of the pricing function (right) one year to maturity for the first example. Note that the contours exhibit right angles in the exercise region. In Figure 4, we plot the exercise region one year to maturity for the first example (left) and for the second example (right). One sees that the exercise boundary has singularities. It is also visible that the mesh has been adapted near the exercise boundary.

CALIBRATION

Let us now discuss the question of the determination of the parameters that appear in the models we introduced, with an emphasis on the *calibration* of the *local volatility*.

Limitation of the Black-Scholes Model: The Need for Calibration

Consider a European-style option on a given stock with a maturity T and a payoff function ϕ , and assume that this option is on the market. Call p its present price. Also, assume the risk-free interest rate is the constant r . One may associate with p the so-called implied volatility, that is, the volatility σ_{imp} such that the price given by formula (4) at time $t = 0$ with $\sigma = \sigma_{imp}$ coincides with p . If the Black-Scholes model was sharp, then the implied volatility would not depend on the payoff function ϕ . Unfortunately,

for vanilla European puts or calls, for example, it is often observed that the implied volatility is far from constant. Rather, it is often a convex function of the strike price. This phenomenon is known as the volatility smile. A possible explanation for the volatility smile is that the deeply out-of-the-money options are less liquid, thus relatively more expensive than the options in-the-money.

This shows that the critical parameter in the Black-Scholes model is the volatility σ . Assuming σ constant and using (8) often leads to poor predictions of the options' prices. The volatility smile is the price paid for the too great simplicity of Black-Scholes' assumptions.

Let us now discuss some of the possible enrichments of the Black-Scholes model:

- Local volatility models: The volatility is a function of time and of the spot price, that is, $\sigma_t = \sigma(t, S_t)$. With suitable assumptions on the regularity and the behavior at infinity of the function σ , (4) holds, and $P_t = p(t, S_t)$, where p satisfies the final value problem (8), in which σ varies with t and S . Calibrating the model consists of tuning the function σ in such a way that the prices computed, for example, with the PDE coincide with the observed prices. This will be discussed in detail below.
- Stochastic volatility models: One assumes that $\sigma_t = f(y_t)$, where y_t is a continuous time stochastic process, correlated or not to the process driving S_t ; see Fouque, Papanicolaou, and Sircar (2000) for a nice presentation. Several models have been proposed, among which are the following:

1. Hull-White model (see Hull and White, 1987): $f(y) = \sqrt{y}$ and y_t is a lognormal process.
2. Scott model: $f(y) = \sqrt{y}$ and y_t is a mean-reverting Ornstein-Uhlenbeck process:

$$dy_t = \alpha(m - y_t)dt + \beta dZ_t \quad (55)$$

where α and β are positive constants, Z_t is a Brownian motion.

3. Heston model (see Heston, 1993): $f(y) = \sqrt{y}$ and y_t is a Cox-Ingersoll-Ross process,

$$dy_t = \kappa(m - y_t)dt + \lambda\sqrt{y_t}dZ_t \quad (56)$$

where κ , m , and λ are positive constants.

4. Stein-Stein model (see Stein and Stein, 1991): $f(y) = \sqrt{y}$ and y_t is a mean-reverting Ornstein-Uhlenbeck process.

There are two risk factors, one for the stock price and the other for the volatility. If the two driving processes are not completely correlated, it is not possible to construct a hedged portfolio containing simply one option and shares of the underlying asset. One says that the market is incomplete. Nevertheless, if one fixes the contribution of the second source of randomness dZ_t to the risk premium, that is, the market price of the volatility risk or the risk premium factor as a function of t , S_t and y_t , then it is possible to prove that the option's price is of the form $P_t = p(t, S_t, y_t)$, where the pricing function satisfies a PDE in the variables (t, S, y) . The PDE may be degenerate for the values of y corresponding to volatility cancellation. Calibrating the model consists of tuning the parameters of the process y_t and the function f in order to match the observed prices.

- Lévy-driven spot price: One may generalize the Black-Scholes model by assuming that the spot price is driven by a more general stochastic process, for example, a Lévy process (see Cont and Tankov, 2003; Merton, 1976; and Carr, Geman, and Yor, 2002). Lévy processes are processes with stationary and independent increments which are continuous in probability. For a Lévy process X_t on a filtered probability space with probability \mathbb{P}^* , the Lévy-Khintchine formula says that there exists a function $\chi : \mathbb{R} \rightarrow \mathbb{C}$ such that

$$\begin{aligned} \mathbb{E}^*(e^{iuX_t}) &= e^{-\tau\chi(u)}, \\ \chi(u) &= \frac{\sigma^2 u^2}{2} - i\beta u - \int_{|z|<1} (e^{iuz} - 1 - iuz)v(dz) \\ &\quad - \int_{|z|>1} (e^{iuz} - 1)v(dz) \end{aligned}$$

for $\sigma \geq 0$, $\beta \in \mathbb{R}$ and a positive measure ν on $\mathbb{R} \setminus \{0\}$ such that $\int_{\mathbb{R}} \min(1, z^2)\nu(dz) < +\infty$. The measure ν is called the Lévy measure of X . We focus on the Lévy measure with a density, $\nu(dz) = k(z)dz$. It is assumed that the discounted price of the risky asset is a square integrable martingale under \mathbb{P}^* , and that it is represented as the exponential of a Lévy process:

$$e^{-r\tau} S_\tau = S_0 e^{X_\tau}$$

The martingale property is that $\mathbb{E}^*(e^{X_\tau}) = 1$, i.e.

$$\begin{aligned} \int_{|z|>1} e^z \nu(dz) < \infty, \quad \text{and} \\ \beta &= -\frac{\sigma^2}{2} - \int_{\mathbb{R}} (e^z - 1 - z1_{|z|\leq 1})k(z)(dz) \end{aligned}$$

and the square integrability comes from the condition $\int_{|z|>1} e^{2z}k(z)dz < \infty$.

With such models, the pricing function for a European option is obtained by solving a partial integrodifferential equation (PIDE), with a nonlocal term. Calibrating the model consists of tuning σ and the function k in such a way that the prices computed with the PIDE, for example, match the observed prices (see Cont and Tankov, 2004).

Local Volatility and Dupire's Formula

We consider a local volatility model and call $(t, S) \mapsto C(t, S, \tau, x)$ the pricing function for a vanilla European call with maturity τ and strike x . It satisfies the final value problem: for $t \in [0, \tau)$ and $x \in \mathbb{R}_+$,

$$\begin{aligned} \frac{\partial C}{\partial t} + \frac{\sigma^2(t, S)S^2}{2} \frac{\partial^2 C}{\partial S^2} + (r - q)S \frac{\partial C}{\partial S} - rC &= 0 \\ C(\tau, S) &= (S - x)_+ \end{aligned} \quad (57)$$

where we have supposed that the underlying asset yields a distributed dividend, $qS_t dt$. By reasoning directly on (4) or by using PDE arguments, it can be proved that the function $(\tau, x) \mapsto C(t, S, \tau, x)$ (now t and S are fixed)

satisfies the forward parabolic PDE:

$$\frac{\partial C}{\partial \tau} - \frac{1}{2}\sigma^2(\tau, x)x^2\frac{\partial^2 C}{\partial x^2} + (r - q)x\frac{\partial C}{\partial x} + qC = 0 \tag{58}$$

for $\tau > t$ and $x \in \mathbb{R}_+$. This observation was first made by Dupire (1994), and the proof of (58) by PDE arguments can be found in Achdou and Pironneau (2005) or Pironneau (2007). We also mention that similar partial differential equations can be derived for other options, like binary options, *barrier options*, options on Lévy-driven assets, or basket options (see Pironneau, 2007).

Using (58) is useful for two reasons. First, consider a family of calls on the same stock with different maturities and strikes (τ_i, x_i) , $I \in I$, where I is a finite set. Assume that the spot price is known, that is, $S = S_0$. In order to numerically compute the prices of the calls, that is, $C(0, S_0, \tau_i, x_i)$, $i \in I$, one may solve (58) for $\max_{i \in I} \tau_i > \tau > 0$ and initial data $C(\tau = 0, x) = (S_0 - x)_+$ with, for example, a finite difference or a finite element method. Only one initial value problem is needed. On the contrary, using (8) would necessitate solving $\#I$ initial value problems. We see that (58) may save a lot of work.

Second, (58) may be used for local volatility calibration. Indeed, if all the possible vanilla options were on the market, the local volatility in (57) could be computed:

$$\sigma^2(\tau, x) = 2 \frac{\frac{\partial C}{\partial \tau}(\tau, x) + (r - q)x\frac{\partial C}{\partial x}(\tau, x) + qC(\tau, x)}{x^2\frac{\partial^2 C}{\partial x^2}(\tau, x)} \tag{59}$$

This is known as *Dupire's formula* for the local volatility. In practice, (59) cannot be used directly because only a finite number of options are on the market.

Assume that the observations are the prices $(\tilde{C}_i)_{i \in I}$ of a family of calls with maturity/strike $(\tau_i, x_i)_{i \in I}$. Finding a function $(\tau, x) \mapsto \sigma(\tau, x)$ such that the solution of (58) with $C(0, x) = (S_0 - x)_+$ takes the value \tilde{C}_i at (τ_i, x_i) , $i \in I$ is called an inverse problem.

A natural idea is to somehow interpolate the observed prices by a sufficiently smooth function $\tilde{C} : [0, \max_{i \in I} \tau_i] \rightarrow \mathbb{R}_+$, then use (59) with $C = \tilde{C}$. For example, bicubic splines may be used. This approach has several serious drawbacks:

- It is difficult to design an interpolation process such that $\frac{\partial^2 \tilde{C}}{\partial x^2}$ does not take the value 0, and such that the right-hand side of (59) is nonnegative.
- There is an infinity of possible interpolations of \tilde{C}_i at (τ_i, x_i) , $i \in I$, and for two possible choices, the volatility obtained by (59) may differ considerably.

We see that financially relevant additional information has to be added to the interpolation process.

Least-Square Methods

Here, we show how (58) can be used for calibration. The first idea is to use *least squares*, that is, to minimize a functional $J : \sigma \mapsto \sum_{i \in I} \omega_i |\tilde{C}_i - C(\tau_i, x_i)|^2$ for σ in a suitable function set Σ , where ω_i are positive weights, and the pricing function C is the solution of (58) with $C(0, x) = (S_0 - x)_+$. The evaluation of J requires the solution of an initial value problem. The set Σ where the volatility is to be found must be chosen in order to ensure that from a minimizing sequence one can extract at least a subsequence that converges in Σ , and that its limit is indeed a solution of the least square problem. For example, Σ may be a compact subset of a Hilbert space W (in principle W could be a more general Banach space but it is easier to work in Hilbert spaces if gradients are needed) such that the mapping J is continuous in W . In practice, W has a finite dimension and is compactly embedded in the space of bounded and continuous functions σ such that $x\partial_x\sigma$ is bounded. Thus, the existence of a solution to the minimization problem is most often guaranteed. What is more difficult to guarantee is uniqueness and stability: Is there a unique solution to the least square problem? If yes, is the solution

insensitive to small variations of the data? The answer to these questions is no in general, and we say that the problem is ill-posed.

As a possible cure to ill-posedness, one usually modifies the problem by minimizing the functional $\sigma \mapsto J(\sigma) + J_R(\sigma)$ instead of J , where J_R is a sufficiently large strongly convex functional defined on W and containing some financially relevant information. For example, one may choose $J_R(\sigma) = \omega \|\sigma - \bar{\sigma}\|^2$, where ω is some positive weight, $\|\cdot\|$ is a norm in W , and $\bar{\sigma}$ is a prior local volatility, which may come from historical knowledge. The difficulty is that ω must not be too large not to perturb the inverse problem too much, but not too small to guarantee some stability. The art of the practitioner lies in the choice of J_R .

Once the least square problem is chosen, we are left with proposing a strategy for the construction of minimizing sequences. If J and J_R are C^1 functional, then gradient methods may be used. The drawbacks and advantages of such methods are well known: On the one hand, they do not guarantee convergence to the global minimum if the functional is not convex, because the iterates can be trapped near a local minimum. On the other hand, they are fast and accurate when the initial guess is close enough to the minimum. For these reasons, gradient methods are often combined with techniques that permit us to localize the global minimum but that are slow, like simulated annealing or evolutionary algorithms.

Anyhow, gradient methods require the evaluation of the functional's gradient. Since J_R explicitly depends on σ , its gradient is easily computed. The gradient of J is more difficult to evaluate, because the prices $C(\tau_i, x_i)$ depend on σ in an indirect way: One needs to evaluate the variations of $C(\tau_i, x_i)$ caused by a small variation of σ ; calling $\delta\sigma$ the variation of σ and δC the induced variation of C , one sees by differentiating (58) that $\delta C(\tau = 0, \cdot) = 0$ and

$$\begin{aligned} \partial_\tau \delta C - \frac{\sigma^2(\tau, x)x^2}{2} \partial_{xx}^2 \delta C + (r - q)x \partial_x \delta C + q \delta C \\ = \sigma \delta \sigma x^2 \partial_{xx}^2 C \end{aligned} \quad (60)$$

To express δJ in terms of $\delta\sigma$, an adjoint state function P is introduced as the solution to the adjoint problem: Find the function P such that $P(\bar{\tau}, \cdot) = 0$ and for $\tau < \bar{\tau}$,

$$\begin{aligned} \partial_\tau P + \partial_{xx}^2 \left(\frac{\sigma^2 x^2}{2} P \right) - \partial_x (P(r - q)x) - qP \\ = 2 \sum_{i \in I} \omega_i (C(\tau_i, x_i) - \bar{C}_i) \delta_{\tau_i, x_i} \end{aligned} \quad (61)$$

where $\bar{\tau}$ is an arbitrary time greater than $\max_{i \in I} \tau_i$ and in the right-hand side, the δ_{τ_i, x_i} denote Dirac functions in time and strike at (τ_i, x_i) . The meaning of (61) is the following:

$$\begin{aligned} - \int_Q \left(\partial_\tau v - \frac{\sigma^2 x^2}{2} \partial_{xx}^2 v + (r - q)x \partial_x v + qv \right) P \\ = 2 \sum_{i \in I} \omega_i (C(\tau_i, x_i) - \bar{C}_i) v(\tau_i, x_i) \end{aligned} \quad (62)$$

where $Q = (0, \bar{\tau}) \times \mathbb{R}_+$, and v is any function such that $v \in L^2((0, \bar{\tau}), V)$ with $\partial_\tau v \in L^2(Q)$ and $x^2 \partial_{xx}^2 v \in L^2(Q)$. Taking $v = \delta C$ in (62) and using (60), one finds

$$\begin{aligned} 2 \sum_{i \in I} \omega_i (C(\tau_i, x_i) - \bar{C}_i) \delta C(\tau_i, x_i) \\ = 2 \sum_{i \in I} \omega_i (C(\tau_i, x_i) - c_i) (\delta_{\tau_i, x_i}, \delta C) \\ = - \int_Q \left(\partial_\tau \delta C - \frac{\sigma x^2}{2} \partial_{xx}^2 \delta C + (r - q)x \partial_x \delta C + q \delta C \right) P \\ = - \int_Q \sigma \delta \sigma x^2 P \partial_{xx}^2 C \end{aligned}$$

We have worked in a formal way, but all the integrations above can be justified. This leads to the estimate

$$\left| \delta J + \int_Q \sigma \delta \sigma x^2 P \partial_{xx}^2 C \right| \leq c \|\delta \sigma\|_{L^\infty(Q)}^2$$

which implies that J is differentiable, and that its differential at point σ is given by

$$DJ(\sigma) : \eta \mapsto - \int_Q \sigma \eta x^2 P(\sigma) \partial_{xx}^2 C(\sigma)$$

where $P(\sigma)$ satisfies (61), and $C(\sigma)$ satisfies (58). We see that the gradient of J can be evaluated. When (58) is discretized with, for example, finite elements, all that has been done can be repeated with a discrete adjoint problem, and the gradient of the functional can be evaluated in the same way. Let us stress that the gradient

$DJ(\sigma)$ is computed exactly, which would not be the case with, for example, a finite difference method.

Local volatility can also be calibrated with American options, but it is not possible to find the analogue of Dupire’s equation. Thus, in the context of a least square approach, the evaluation of the cost function requires the solution of #1 variational inequalities, which is computationally expensive (see Achdou and Pironneau, 2005). In this case, it is also possible to find the necessary optimality conditions involving an adjoint state (see Achdou, 2005).

Appendix: Proof of (21)

First, from the definition (15) of p we have, for any stopping time $\rho \in \mathcal{T}_{[t,T]}$,

$$\begin{aligned} & e^{-\int_t^\rho r ds} p(\rho, S_\rho^{t,x}) \\ &= \text{ess sup}_{\tau \in \mathcal{T}_{[\rho,T]}} \mathbb{E} \left(e^{-\int_t^\tau r ds} \phi(S_\tau^{t,x}) \mid \mathcal{F}_\rho \right), \text{ a.s.} \end{aligned} \tag{63}$$

where $\mathcal{T}_{[\rho,T]}$ denotes the set of stopping times τ such that $\rho \leq \tau \leq T$. Then it is possible to show that (see, for instance, Karatzas and Shreve, 2010, Eq. (D.7)), for any stopping time $\rho \in \mathcal{T}_{[t,T]}$,

$$\begin{aligned} & \mathbb{E} \left(e^{-\int_t^\rho r ds} p(\rho, S_\rho^{t,x}) \right) \\ &= \sup_{\tau \in \mathcal{T}_{[\rho,T]}} \mathbb{E} \left(e^{-\int_t^\tau r ds} \phi(S_\tau^{t,x}) \right) \end{aligned} \tag{64}$$

We obtain from (64) the decreasing property: For all stopping times $\rho_1, \rho_2 \in \mathcal{T}_{[t,T]}$, such that $\rho_1 \geq \rho_2$,

$$\begin{aligned} & \mathbb{E} \left(e^{-\int_t^{\rho_1} r ds} p(\rho_1, S_{\rho_1}^{t,x}) \right) \\ & \leq \mathbb{E} \left(e^{-\int_t^{\rho_2} r ds} p(\rho_2, S_{\rho_2}^{t,x}) \right) \end{aligned} \tag{65}$$

We deduce from (63) that, for any $\tau \in \mathcal{T}_{[\tau^*,T]}$,

$$\begin{aligned} & \mathbb{E} \left(e^{-\int_t^\tau r ds} \phi(S_\tau^{t,x}) \mid \mathcal{F}_{\tau^*} \right) \leq e^{-\int_t^{\tau^*} r ds} p(\tau^*, S_{\tau^*}^{t,x}) \\ &= e^{-\int_t^{\tau^*} r ds} \phi(S_{\tau^*}^{t,x}) \end{aligned} \tag{66}$$

where the last identity comes from the definition (20) of τ^* . Then, for any stopping time

$\tau \in \mathcal{T}_{[t,T]}$, we have (by decomposing on the events $\{\tau < \tau^*\}$ and $\{\tau \geq \tau^*\}$), and using (66) for $\tau \geq \tau^*$):

$$\mathbb{E} \left(e^{-\int_t^\tau r ds} \phi(S_\tau^{t,x}) \right) \leq \mathbb{E} \left(e^{-\int_t^{\tau \wedge \tau^*} r ds} \phi(S_{\tau \wedge \tau^*}^{t,x}) \right)$$

Hence, by taking the supremum over all the stopping times $\tau \in \mathcal{T}_{[t,T]}$,

$$\begin{aligned} p(t, x) & \leq \sup_{\tau \in \mathcal{T}_{[t,T]}} \mathbb{E} \left(e^{-\int_t^{\tau \wedge \tau^*} r ds} \phi(S_{\tau \wedge \tau^*}^{t,x}) \right) \\ &= \sup_{\tau \leq \tau^*, \tau \in \mathcal{T}_{[t,T]}} \mathbb{E} \left(e^{-\int_t^\tau r ds} \phi(S_\tau^{t,x}) \right) \end{aligned} \tag{67}$$

By (15), the right-hand side of (67) is bounded from above by $p(t,x)$, and thus we obtain the equality

$$p(t, x) = \sup_{\tau \leq \tau^*, \tau \in \mathcal{T}_{[t,T]}} \mathbb{E} \left(e^{-\int_t^\tau r ds} \phi(S_\tau^{t,x}) \right) \tag{68}$$

In fact the supremum in (68) is reached only for $\tau = \tau^*$ a.s.. Indeed, for $\tau \in \mathcal{T}_{[t,T]}$, if $\tau \leq \tau^*$ and $\mathbb{P}(\tau < \tau^*) > 0$, we have, by the definition of τ^* , $\mathbb{E}(e^{-\int_t^\tau r ds} \phi(S_\tau^{t,x})) < \mathbb{E}(e^{-\int_t^{\tau^*} r ds} p(\tau, S_\tau^{t,x})) \leq p(t, x)$. This concludes the proof of (21).

KEY POINTS

- When a deterministic method is available to price an option, it is generally more efficient than a brute force Monte Carlo algorithm.
- Deterministic techniques are usually more involved to implement than stochastic approaches and typically require specific developments for each targeted pricing problem.
- Deterministic approaches are particularly useful for nonlinear problems (including the pricing of American options and portfolio optimization) and for calibration.
- Future research subjects for such approaches include the development of efficient discretization methods for high-dimensional problems, and the combination of deterministic and stochastic approaches to take advantage of both techniques (using variance-reduction techniques or predictor-corrector methods, for example).

NOTES

1. Notice that the same equation has been considered by Vecer (2001) using some financial arguments.
2. <http://www-rocq.inria.fr/mathfi/Premia/index.html>
3. <http://www.freefem.org/>
4. More precisely, the interpolating operator should also satisfy $[P] \leq [Q]$ everywhere as soon as $P_j^k \leq Q_j^k$ for all k, j .
5. A good initial guess is indeed the vector α obtained at the previous time iteration.
6. F is slantly differentiable if there exist $C > 0$ and a matrix $G(x)$ such that $\forall x, \|G(x)^{-1}\|_\infty < C$ and $F(x+h) = F(x) + G(x)h + o(h)$ as $h \rightarrow 0$. Here $G(x)$ can be defined by $G(x)_{ij} = B_{ij}$ if $(Bx - b)_i \leq (x - g)_i$, and $G(x)_{ij} = \delta_{ij}$ otherwise.
7. Such a problem is called a linear complementarity problem.

REFERENCES

- Achdou, Y. (2005). An inverse problem for a parabolic variational inequality arising in volatility calibration with American options. *SIAM J. Control Optim.* 43, 5: 1583–1615.
- Achdou, Y., and Pironneau, O. (2005). Computational methods for option pricing. *Frontiers in applied mathematics. Society for Industrial & Applied Mathematics* 30.
- Achdou, Y., Hecht, F., and Pommier, H. (2008). Space-time a posteriori error estimates for variational inequalities. *Journal of Scientific Computing* 37: 336–366.
- Achdou, Y., and Pironneau, O. (2009). Partial differential equations for option pricing. In P. G. Ciarlet, ed., *Handbook of Numerical Analysis*, Vol. XV, Special Volume: *Mathematical Modeling and Numerical Methods in Finance*, Guest Eds., Alain Bensoussan and Qiang Zhang. Netherlands: North-Holland, 369–495.
- Ammar, A., Mokdad, B., Chinesta, F., and Keunings, R. (2002). A new family of solvers for some classes of multidimensional partial differential equations encountered in kinetic theory modeling of complex fluids. *Journal of Non-Newtonian Fluid Mechanics* 139: 153–176.
- Avellaneda, M., Levy, A., and Paras, A. (1995). Pricing and hedging derivative securities in markets with uncertain volatilities. *Applied Mathematical Finance* 2: 73–88.
- Barles, G. (1994). *Solutions de viscosité des équations de Hamilton-Jacobi*, vol. 17 of *Mathematics & Applications*. Paris: Springer-Verlag.
- Barles, G., Daher, C., and Romano, M. (1995). Convergence of numerical schemes for parabolic equations arising in finance theory. *Mathematical Models and Methods in Applied Sciences* 5, 1: 125–143.
- Barles, G., and Souganidis, P. E. (1991). Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Analysis* 4, 3: 271–283.
- Bensoussan, A., and Lions, J.-L. (1978). *Applications des inéquations variationnelles en contrôle stochastique. Méthodes Mathématiques de l'Informatique*, No. 6. Paris: Dunod.
- Beylkin, G., and Mohlenkamp, M. J. (2002). Numerical operator calculus in higher dimensions. *Proceedings of the National Academy of Sciences* 99, 16: 10246–10251.
- Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.
- Bokanowski, O., Maroso, S., and Zidani, H. (2009). Some convergence results for Howard's algorithm. *SIAM Journal of Numerical Analysis* 47, 4: 3001–3026.
- Boyaval, S., Le Bris, C., Lelièvre, T., Maday, Y., Nguyen, N. C., and Patera, A. T. (2010). Reduced basis techniques for stochastic problems. *Archives of Computational Methods in Engineering* 17, 4: 435–454.
- Brennan, M. J., and Schwartz, E. S. (1977). The valuation of the American put option. *Journal of Finance* 32: 449–462.
- Bungartz, H. J., and Griebel, M. (2004). Sparse grids. *Acta Numerica* 13: 147–269.
- Capuzzo-Dolcetta, I., and Falcone, M. (1989). Discrete dynamic programming and viscosity solutions of the Bellman equation. *Annales de l'Institut Henri Poincaré, Analyse Non Linéaire* 6: 161–183.
- Carr, P., Geman, H., Madan, D., and Yor, M. (2002). The fine structure of asset returns: An empirical investigation. *Journal of Business* 75, 2: 305–332.
- Cont, R., and Tankov, P. (2003). *Financial Modelling with Jump Processes*. Boca Raton: Chapman and Hall.
- Cont, R., and Tankov, P. Nonparametric calibration of jump-diffusion option pricing models. *Journal of Computational Finance* 7, 3: 1–49.

- Courant, R., Friedrichs, K., and Lewy, H. (1967). On the partial difference equations of mathematical physics. *IBM Journal of Research and Development* 11: 215–234.
- Cox, J., Ross, S., and Rubinstein, M. (1979). Option pricing: A simplified approach. *The Journal of Financial Economics* 7: 44–50.
- Crandall, M. G., Ishii, H., and Lions, P.-L. (1992). User's guide to viscosity solutions of second order partial differential equations. *Bulletin of the American Mathematical Society (N.S.)* 27, 1: 1–67.
- Crandall, M. G., and Lions, P.-L. (1984). Two approximations of solutions of Hamilton-Jacobi equations. *Mathematics of Computation* 43, 167: 1–19.
- Dubois, F., and Lelièvre, T. (2005). Efficient pricing of Asian options by the PDE approach. *Journal of Computational Finance* 8, 2: 55–64.
- Duffie, D. (1992). *Dynamic Asset Pricing Theory*. Princeton: Princeton University Press.
- Dupire, B. (1994). Pricing with a smile. *Risk*, 18–20.
- El Karoui, N. (1981). Les aspects probabilistes du contrôle stochastique. In *Ninth Saint Flour Probability Summer School—1979 (Saint Flour, 1979)*, vol. 876 of *Lecture Notes in Math.*, pp. 73–238. Berlin: Springer.
- Ern, A., Villeneuve, S., and Zanette, A. (2004). Adaptive finite element methods for local volatility European option pricing. *International Journal of Theoretical and Applied Finance* 7, 6: 659–684.
- Falcone, M., and Ferretti, R. (1994). Discrete time high-order schemes for viscosity solutions of Hamilton-Jacobi-Bellman equations. *Numerische Mathematik* 67, 3: 315–344.
- Fouque, J.-P., Papanicolaou, G., and Sircar, R. (2000). *Derivatives in Financial Markets with Stochastic Volatility*. Cambridge: Cambridge University Press.
- Heston, S. (1993). A closed form solution for options with stochastic volatility with application to bond and currency options. *Review with Financial Studies* 6: 327–343.
- Hintermuüller, M., Ito, K., and Kunisch, K. (2002). The primal-dual active set strategy as a semi-smooth Newton method. *SIAM Journal of Optimization* 13, 3: 865–888 (electronic, 2003).
- Howard, R. A. (1960). *Dynamic Programming and Markov Processes*. Cambridge, MA: The Technology Press of M.I.T.
- Hull, J. C., and White, A. (1987). The pricing of options on assets with stochastic volatilities. *Journal of Finance* 42: 281–300.
- Jaillet, P., Lamberton, D., and Lapeyre, B. (1990). Variational inequalities and the pricing of American options. *Acta Applicandae Mathematicae* 21, 3: 263–289.
- Jakobsen, E. R. (2003). On the rate of convergence of approximation schemes for Bellman equations associated with optimal stopping time problems. *Mathematical Models and Methods in Applied Sciences* 13, 5: 613–644.
- Karatzas, I., and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*, 2nd ed. New York: Springer-Verlag.
- Karatzas, I., and Shreve, S. E. (2010). *Methods of Mathematical Finance*. New York: Springer-Verlag.
- Kushner, H. J., and Dupuis, P. (2001). Numerical methods for stochastic control problems in continuous time. *Applications of Mathematics*, Vol. 24. New York: Springer-Verlag.
- Lamberton, D., and Lapeyre, B. (1997). *Introduction au calcul stochastique appliqué à la finance*. Ellipses, Paris.
- Le Bris, C., Lelièvre, T., and Maday, Y. (2009). Results and questions on a nonlinear approximation approach for solving high-dimensional partial differential equations. *Constructive Approximation* 30: 621–651.
- Lions, P.-L., and Mercier, B. (1979). Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis* 16, 6: 964–979.
- Merton, R. C. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science* 4: 141–183.
- Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3: 125–144.
- Nouy, A. (2009). Recent developments in spectral stochastic methods for the numerical solution of stochastic partial differential equations. *Archives of Computational Methods in Engineering* 16: 251–285.
- Øksendal, B., and Reikvam, K. (1998). Viscosity solutions of optimal stopping problems. *Stochastics and Stochastics Reports* 62, 3–4: 285–301.
- Pham, H. (2006). *Optimisation et Contrôle Stochastique Appliqués à la Finance*. Springer Verlag, Berlin.
- Pham, H. (1998). Optimal stopping of controlled jump diffusion processes: A viscosity solution approach. *Journal of Mathematical Systems, Estimation, and Control* 8, 1: 27 pp. (electronic).

- Pironneau, O. (2007). Dupire-like identities for complex options. *Comptes Rendus de l'Académie des sciences, Série I* 344: 127–133.
- Quarteroni, A., and Valli, A. (1997). *Numerical Approximation of Partial Differential Equations*. Springer, Berlin.
- Rogers, L. C. G., and Shi, Z. (1995). The value of an Asian option. *Journal of Applied Probability* 32: 1077–1088.
- Saad, Y. (2003). Iterative methods for sparse linear systems. *Society for Industrial & Applied Mathematics*.
- Santos, M. S., and Rust, J. (2004). Convergence properties of policy iteration. *SIAM Journal on Control and Optimization* 42, 6: 2094–2115.
- Stein, E., and Stein, J. (1991). Stock price distributions with stochastic volatility: An analytic approach. *Review of Financial Studies* 4, 4: 727–752.
- Temlyakov, V. N. (2008). Greedy approximation. *Acta Numerica* 17: 235–409.
- Thompson, G. W. P. (1999). Fast narrow bounds on the value of Asian options. Working paper, Judge Institute U. of Cambridge, 1999.
- van der Pijl, S. P., and Osterlee, C. W. (2011). An ENO-based method for second order equations and application to the control of dike levels. *Journal of Scientific Computing*, in press.
- Věcěr, J. (2001). A new PDE approach for pricing arithmetic average Asian options. *Journal of Computational Finance* 4, 4: 105–113.
- von Petersdoff, T., and Schwab, C. (2004). Numerical solutions of parabolic equations in high dimensions. *Mathematical Modelling and Numerical Analysis* 38: 93–128.
- Wilmott, P., Dewynne, J., and Howison, S. (1993). *Option Pricing: Mathematical Models and Computation*. Oxford: Oxford Financial Press.
- Zvan, R., Forsyth, P. A., and Vetzal, K. (1998). Robust numerical methods for PDE models of Asian options. *Journal of Computational Finance* 1, 2: 39–78.

Model Risk and Selection

Model Risk

KEVIN DOWD, PhD

Partner, Cobden Partners, London

Abstract: Model risk is the risk of error in pricing or risk-forecasting (such as value at risk, or VaR) models. It arises in part because any model involves simplification and calibration, and both of these require subjective judgments that are inevitably open to error. Model risk can also arise where a model is used inappropriately. Model risk is therefore an inescapable consequence of model use, and there is abundant anecdotal and other evidence that it is a major problem, especially for VaR models. However, there are also many ways in which risk managers and financial institutions can manage this problem.

This entry examines the subject of model risk. Loosely speaking, *model risk* is the risk of error in the valuations produced by a pricing model or in the estimated risk measures produced by a risk model. The nature of model risk and its diverse causes and manifestations are examined. The entry also briefly addresses the scale of the problem and the dangers it entails, and then goes on to discuss ways in which model risk can be managed.

MODELS AND MODEL RISK

A *model* can be defined as “a simplified description of reality that is at least potentially useful in decision-making” (Geweke, 2005, p. 7). A model attempts to identify the key features of whatever it is meant to represent and is, by its very nature, a highly simplified structure. We should therefore not expect any model to give a perfect answer: Some degree of error is to be expected, and we can think of this risk of error as a form of model risk.

However, the term *model risk* is more subtle than it looks, and not all output errors are due to model inadequacy. For example, simulation methods generally produce errors due to sampling variation, so even the best simulation-based model will produce results affected by random noise. Conversely, models that are theoretically inappropriate can sometimes provide good results. The most obvious cases in point are the well-known “holes in Black-Scholes”: Simple option pricing models often work well even when some of the assumptions on which they are based are known to be invalid. They work well not because they are accurate, but because those who use them are aware of their limitations and use them discerningly.

In finance, we are concerned with both *pricing* (or valuation) *models* and risk (or VaR) *models*. The former are models that enable us to price a financial instrument, and with these model risk boils down to the risk of mispricing. These models are typically used on a stand-alone basis and it is often very important that

they give precise answers: Mispricing can lead to rapid and large arbitrage losses. Their exposure to this risk depends on such factors as the complexity of the position, the presence or otherwise of unobserved variables (e.g., such as volatilities), interactions between risk factors, the use of numerical approximations, and so on.

Risk models are models that forecast financial risks or probabilities. These models are exposed to many of the same problems as pricing models, but all are often also affected by the difficulties of trying to integrate risks across different positions or business units, and this raises a host of issues (e.g., aggregation problems, potential inconsistencies across constituent positions or models, etc.) that do not (typically) arise in stand-alone pricing models. So risk models are exposed to more sources of model risk than pricing models typically are. However, with risk models there is far less need for accuracy: Errors in risk estimates do not lead directly to arbitrage losses, and the old engineering principle applies that the end output is only as good as the weakest link in the system. With risk models, we therefore want to be approximate and right, and efforts to achieve high levels of precision would be pointless because any reported precision would be spurious.

We are particularly concerned in this entry with how models can go wrong, and to appreciate these problems it helps to understand how our models are constructed in the first place. To get to know our models we should:

- Understand the securities involved and the markets in which they are traded.
- Isolate the most important variables and separate out the causal variables (or exogenous variables) from the caused (or endogenous) variables.
- Decide which exogenous variables are deterministic and which are stochastic or random, decide how the exogenous variables are to be modeled, and decide how the exogenous variables affect the endogenous ones.
- Decide which variables are observable or measurable and which are not; decide how the former are measured, and consider whether and how the unobservable variables can be proxied or implicitly solved from other variables.
- Try to ensure that the model captures all key features of the problem at hand, but also has no unnecessary complexity.
- Consider how the model can be solved and look for the simplest possible solutions. We should also consider the possible benefits and drawbacks of using approximations instead of exact solutions.
- Program the model, taking account of programming considerations, computational time, and so on.
- Calibrate the model using suitable methods: For example, we might estimate parameters using maximum likelihood methods and then adjust them using subjective judgments about factors such as changing market conditions that might not be fully reflected in our data set.
- Test the model using data not used to calibrate the model.
- Implement the model, regularly evaluate its performance, and identify its strengths and weaknesses.
- Keep a log of all these activities and their outcomes.

SOURCES OF MODEL RISK

Incorrect Model Specification

One of the most important sources of model risk is incorrect model specification, and this can manifest itself in many ways:

- **Stochastic processes might be misspecified.** We might assume that a stochastic process follows a geometric Brownian motion when it is in fact heavy-tailed, we might fit a symmetric distribution to skewed data, and so forth. It is very easy to misspecify stochastic processes, because the “true” stochastic

process is very difficult to identify and it is impossible in practice to distinguish between a “true” process and a similar but false alternative. The misspecification of stochastic processes can lead to major errors in estimates of risk measures: The classic example is where we incorrectly assume normality in the presence of heavy tails, an error that can lead to major underestimates of VaR and other risk measures.

- **Incorrect calibration of parameters.** Even if we do manage to identify the “true” model, the model might be calibrated with incorrect parameter values. Parameters might be estimated with error, not kept up to date, estimated over inappropriate sample periods, and so forth. This problem is often referred to as *parameter risk*, and it arises everywhere in risk management because it is practically impossible to determine “true” parameter values. Besides leading to major errors in risk estimates, incorrect calibration can also lead to major losses if the models are used to price traded instruments. A good example was the £90 million loss made by the NatWest Bank from 1995 to 1997, where a trader had fed his own (artificially high) estimates of volatility into a model used to price long-dated over-the-counter (OTC) interest rate options. We can also get problems when correlations unexpectedly polarize in a crisis: In such cases, the portfolio loses much of its effective diversification, and the “true” risks taken can be much greater than estimates based on earlier correlations might suggest.
- **Missing risk factors and misspecified relationships.** We might ignore stochastic volatility or fail to consider enough points across the term structure of interest rates, ignore background risk factors such as macroeconomic ones, or we might misspecify important relationships (e.g., by ignoring correlations).
- **Ignoring of transactions costs, liquidity, and crisis factors.** Many models ignore transactions costs and assume that markets are perfectly liquid. Such assumptions are very

convenient for modelling purposes, but can lead to major errors where transactions costs are significant, where market liquidity is limited, or where a crisis occurs. These sorts of problems were highlighted by the difficulties experienced by portfolio insurance strategies in the October 1987 crash—where strategies predicated on dynamic hedging were unhinged by the inability to unwind positions as the market fell. The failure to allow for illiquidity led to much larger losses than the models anticipated—a classic form of model risk.

There is empirical evidence that model misspecification risk is a major problem. To give a couple of examples: Hendricks (1996) investigated differences between alternative VaR estimation procedures applied to 1,000 randomly selected simple foreign exchange portfolios, finding that these differences were sometimes substantial; more alarmingly, a famous study by Beder 1995 examined eight common VaR methodologies used by a sample of commercial institutions applied to three hypothetical portfolios, and among other worrying results found that alternative VaR estimates for the same portfolio could differ by a factor of up to 14. Some further evidence is provided by Berkowitz and O’Brien (2001) who examined the VaR models used by six leading U.S. financial institutions. Their results indicated that these models can be highly inaccurate: Banks sometimes experienced high losses very much larger than their models predicted, and this suggests that these models are poor at dealing with heavy tails or extreme risks. Their results also suggest that banks’ structural models embody so many approximations and other implementation compromises that they lose any edge over much simpler models such as generalized autoregressive conditional heteroskedasticity (GARCH) ones. The implication is that financial institutions’ risk models are very exposed to model risk—and one suspects many risk managers are not aware of the extent of the problem.

Incorrect Model Application

Model risk can also arise because a good model is incorrectly applied. To quote Emanuel Derman:

There are always implicit assumptions behind a model and its solution method. But human beings have limited foresight and great imagination, so that, inevitably, a model will be used in ways its creator never intended. This is especially true in trading environments, where not enough time can be spent on making interfaces fail-safe, but it's also a matter of principle: you just cannot foresee everything. So, even a "correct" model, "correctly" solved, can lead to problems. The more complex the model, the greater this possibility. (Derman, 1997, p. 86)

One can give very many instances of this problem: We might use the wrong model in a particular context (e.g., we might use a Black-Scholes model for pricing options when we should have used a stochastic volatility model, etc.); we might have initially had the right model, but have fallen behind best market practice and not kept the model up to date, or not replaced it when a superior model became available; we might run Monte Carlo simulations with a poor random number generator or an insufficient number of trials, and so on. We can also get "model creep," where a model is initially designed for one type of problem and performs well on that problem, but is then gradually applied to more diverse situations to which it is less suited or not suited at all. A perfectly good model can then end up as a major liability not because there is anything wrong with it, but because users don't appreciate its limitations.

Implementation Risk

Model risk also arises from the ways in which models are implemented. No model can provide a complete specification of model implementation in every conceivable circumstance because of the very large number of possible instruments and markets, and because of their

varying institutional, statistical, and other properties. However complete the model, implementation decisions still need to be made about such factors as valuation (e.g., mark to market versus mark to model, whether to use the mean bid-ask spread, etc.), whether and how to clean data, how to map instruments, how to deal with legacy systems, and so on.

The possible extent of *implementation risk* is illustrated by the results of a study by Marshall and Siegel (1997). They sought to quantify implementation risk by looking at differences between how various commercial systems applied the RiskMetrics variance-covariance approach to specified positions based on a common set of assumptions (that is, a one-day holding period, a 95% VaR confidence level, delta-valuation of derivatives, RiskMetrics mapping systems, etc.). They found that any two sets of VaR estimates were always different, and that VaR estimates could vary by up to nearly 30% depending on the instrument class; they also found these variations were in general positively related to complexity: The more complex the instrument or portfolio, the greater the range of variation of reported VaRs. These results suggested that:

[A] naive view of risk assessment systems as straightforward implementations of models is incorrect. Although software is deterministic (i.e., given a complete description of all the inputs to the system, it has well-defined outputs), as software and the embedded model become more complex, from the perspective of the only partially knowledgeable user, they behave stochastically. . . . Perhaps the most critical insight of our work is that as models and their implementations become more complex, treating them as entirely deterministic black boxes is unwise, and leads to real implementation and model risks. (Marshall and Siegel, 1997, pp. 105–106)

Endogenous Model Risk

There is also a particularly subtle and invidious form of model risk that arises from the ways in which traders or asset managers respond to the

models themselves: Traders or asset managers will “game” against the model. Traders are likely to have a reasonable idea of the errors in the parameters—particularly volatility or correlation parameters—used to estimate VaR, and such knowledge will give the traders an idea of which positions have under- and overestimated risks. If traders face VaR limits or face risk-adjusted remuneration with risks specified in VaR terms, they will therefore have an incentive to seek out such positions and trade them. To the extent they do, they will take on more risk than suggested by VaR estimates, which will therefore be biased downward. Indeed, VaR estimates are likely to be biased even if traders do not have superior knowledge of underlying parameter values. The reason for this is that if a trader uses an estimated variance-covariance matrix to select trading positions, then he or she will tend to select positions with low estimated risks, and the resulting changes in position sizes mean that the initial variance-covariance matrix will tend to underestimate the resulting portfolio risk. As Shaw nicely puts it:

[M]any factor models fail to pick up the risks of *typical* trading strategies which can be the greatest risks run by an investment bank. According to naïve yield factor models, huge spread positions between on-the-run bonds and off-the-run bonds are riskless! According to naïve volatility factor models, hedging one year (or longer dated) implied volatility with three month implied volatility is riskless, provided it is done in the “right” proportions—i.e., the proportions built into the factor model! It is the *rule*, not the exception, for traders to put on spread trades which defeat factor models *since they use factor type models to identify richness and cheapness!* (Shaw, 1997, p. 215; his emphasis)

Other Sources of Model Risk

There are also other sources of model risk. Programs might have errors or bugs in them, simulation methods might use poor random number generators or suffer from discretization errors, approximation routines might be inaccurate or fail to converge to sensible solutions, rounding

errors might add up, and so on. We can also get problems when programs are revised by people who did not originally write them, when programs are not compatible with user interfaces or other systems (e.g., datafeeds), when programs become complex or hard to read (e.g., when programs are rewritten to make them computationally more efficient but then become less easy to follow). We can also get simple blunders. Derman (1997, p. 87) reported the example of a convertible bond model that was good at pricing many of the options features embedded in convertible bonds, but sometimes miscounted the number of coupon payments left to maturity.

Finally, models can give incorrect answers because poor data are fed into them—“garbage in, garbage out,” as the saying goes. Data problems can arise from many sources: from the way data are constructed (e.g., whether we mark to market or mark to model, whether we use actual trading data or end-of-day data, how we deal with bid-ask spreads, etc.), from the way time is handled (e.g., whether we use calendar time, trading time, how we deal with holidays, etc.), from the way in which data are cleansed or standardized, from data being non-synchronous, and from many other sources.

MANAGING MODEL RISK

Some Guidelines for Risk Managers

Given that risk managers can never eliminate model risk, the only option left is to learn to live with it and, hopefully, manage it. Practitioners can do so in a number of ways:

- **Be aware of model risk.** First and foremost, practitioners should simply be aware of it, and be aware of the limitations of the models they use. They should also be aware of the comparative strengths and weaknesses of different models, be knowledgeable of which models suit which problems, and be on the

lookout for models that are applied inappropriately.

- **Identify, evaluate, and check key assumptions.** Users should explicitly set out the key assumptions on which a model is based, evaluate the extent to which the model's results depend on these assumptions, and check them as much as possible (e.g., using statistical tests).
- **Choose the simplest reasonable model.** Exposure to model risk is reduced if practitioners always choose the simplest reasonable model for the task at hand. Occam's razor applies just as much in model selection as in anything else: Unnecessary complexity is never a virtue. Whenever we choose a more complex model over a simpler one, we should always have a clear reason for doing so.
- **Don't ignore small problems.** Practitioners should resist the temptation to explain away small discrepancies in results and sweep them under the rug. Small discrepancies are often good warning signals of larger problems that will manifest themselves later if they are not sorted out.
- **Test models against known problems.** It is always a good idea to check a model on simple problems to which one already knows the answer, and many problems can be distilled to simple special cases that have known answers. If the model fails to give the correct answer to a problem whose solution is already known, then we immediately know that there must be something wrong with it.
- **Plot results and use nonparametric statistics.** Graphical outputs can be extremely revealing, and simple histograms or plots often show up errors that might otherwise be very hard to detect. For example, a plot might have the wrong slope or shape or have odd features such as kinks that flag an underlying problem. Summary statistics and simple nonparametric tests can also be useful for helping to impart a feel for data and results.

- **Back-test and stress-test the model.** Practitioners should evaluate model adequacy using stress tests and back tests.
- **Estimate model risk quantitatively.** Where feasible, practitioners should seek to estimate model risk quantitatively (e.g., using simulation methods). However, it helps to keep in mind that any quantitative estimate of model risk is almost certainly an underestimate because not all model risk is quantifiable.
- **Reevaluate models periodically.** Models should be re-calibrated and reestimated on a regular basis, and the methods used should be kept up to date.

Some Institutional Guidelines

Financial institutions themselves can also combat model risk through appropriate institutional devices. One defense is a sound system to vet models before they are approved for use and then periodically review them. A good model-vetting procedure is proposed by Crouhy et al. (2001, pp. 607–608) and involves the following four steps:

1. **Documentation.** The risk manager should ask for a complete specification of the model, including its mathematics, components, computer code, and implementation features (e.g., numerical methods and pricing algorithms used). The information should be in sufficient detail to enable the risk manager to reproduce the model from the information provided.
2. **Soundness.** The risk manager should check that the model is a reasonable one for the instrument(s) or portfolio concerned.
3. **Benchmark modeling.** The risk manager should develop a benchmark model and test it against well-understood approximation or simulation methods.
4. **Check results and test the proposed model.** The final stage involves the risk manager using the benchmark model to check the

performance of the proposed model. The model should also be checked for zero-arbitrage properties such as put-call parity, and should then be stress tested to help determine the range of parameter values for which it will give reasonable estimates.

All these stages should be carried out free of undue pressures from the front office, and traders should not be allowed to vet their own pricing models. It is also important to keep good records, so each model should be fully documented in the middle (or risk) office. Risk managers should have full access to the model at all times, as well as access to real trading and other data that might be necessary to check models and validate results. The ideal should be to give the middle office enough information to be able to check any model or model results at any time, and do so using appropriate (that is, up to date) data sets. This information set should include a log of model performance with particular attention to any problems encountered and what (if anything) has been done about them. There should also be a periodic review (as well as occasional spot check) of the models in use, to ensure that model calibration is up to date and that models are upgraded in line with market best practice, and to ensure that obsolete models are identified as such and taken out of use. Such risk audits should also address not just the risk models, but all aspects of the firm's risk management. And, of course, all these measures should take place in the context of a strong and independent risk oversight or middle office function.

KEY POINTS

- A model attempts to identify the key features of whatever it is meant to represent and is, by its very nature, a highly simplified structure.
- In financial modeling, the concern is with both pricing (or valuation) models and risk (or VaR) models. The risk of error in pricing or risk-forecasting models is referred to as model risk.

ing or risk-forecasting models is referred to as model risk.

- Model risk is an inescapable consequence of model use and affects both pricing models and VaR models.
- The main sources of model risk include incorrect specification, incorrect application, implementation risk, and the problem of *endogenous model risk* where traders “game” against the model.
- There are ways in which practitioners can manage model risk. These include (1) recognizing model risk, (2) identifying, evaluating, and checking the model's key assumptions, (3) selecting the simplest reasonable model, (4) resisting the temptation to ignore small discrepancies in results, (5) testing the model against known problems, (6) plotting results and employing nonparametric statistics, (7) back-testing and stress-testing the model, (8) estimating model risk quantitatively, and (9) reevaluating models periodically.

REFERENCES

- Beder, T. (1995). VaR: Seductive but dangerous. *Financial Analysts Journal* 51, 5: 12–24.
- Black, F. (1992). How to use the holes in Black-Scholes. In R. Kolb (ed.), *The Financial Derivatives Reader* (pp. 198–204). Miami: Kolb Publishing.
- Berkowitz, J., and O'Brien, J. (2002). How accurate are value-at-risk models at commercial banks? *Journal of Finance* 57, 3: 1093–1112.
- Crouhy, M., Galai, D., and Mark, R. (2001). *Risk Management*. New York: McGraw-Hill.
- Derman, E. (1997). Model risk. In S. Grayling (ed.), *VaR—Understanding and Applying Value-at-Risk* (pp. 83–88). London: Risk Publications.
- Derman, E. (2004). *My Life as a Quant: Reflections on Physics and Finance*. Hoboken, NJ: John Wiley & Sons.
- Dowd, K. (2005). *Measuring Market Risk*, 2nd ed. Chichester: John Wiley & Sons.
- Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*. Hoboken, NJ: John Wiley & Sons.

- Ju, X., and Pearson, N. D. (1999). Using value-at-risk to control risk taking: How wrong can you be? *Journal of Risk* 1, 2: 5–36.
- Kato, T., and Yoshida, T. (2000). Model risk and its control. *Bank of Japan Institute for Monetary and Economic Studies Discussion Paper No. 2000-E-15*.
- Marshall, C., and Siegel, M. (1997). Value at risk: Implementing a risk measurement standard. *Journal of Derivatives* 4, 1: 91–110.
- Shaw, J. (1997). Beyond VaR and stress testing. In S. Grayling (ed.), *VAR—Understanding and Applying Value at Risk* (pp. 221–224). London: KPMG/Risk Publications.
- Siu, T. K., Tong, H., and Yang, H. (2001). On Bayesian value at risk: From linear to non-linear portfolios. Mimeo, National University of Singapore, University of Hong Kong, and London School of Economics.

Model Selection and Its Pitfalls

SERGIO M. FOCARDI, PhD

Partner, The Intertek Group

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

PETTER N. KOLM, PhD

Director of the Mathematics in Finance Masters Program and Clinical Associate Professor,
Courant Institute of Mathematical Sciences, New York University

Abstract: Financial modelers have to solve the critical problem of selecting or perhaps building the optimal model to represent the phenomena they seek to study. The task calls for a combination of personal creativity, theory, and machine learning.

In this entry we discuss methods for model selection and analyze the many pitfalls of the model selection process.

MODEL SELECTION AND ESTIMATION

In his book *Complexity*, Mitchell Waldrop (1992) describes the 1987 Global Economy Workshop held at The Santa Fe Institute, a research center dedicated to the study of complex phenomena and related issues. Organized by the economist Bryan Arthur and attended by distinguished economists and physicists, the seminar introduced the idea that economic laws might be better understood applying the principles of physics and, in particular, the newly developed theory of complex systems. The seminar proceedings were to become the influential

book *The Economy as an Evolving Complex System* (Anderson, Arrow, and Pines, 1998).

An anecdote from the book is revealing of the issues specific to economics as a scientific endeavor. According to Waldrop, physicists attending the seminar were surprised to learn that economists used highly sophisticated mathematics.

A physicist attending the seminar reportedly asked Kenneth Arrow, the 1972 Nobel Prize winner in economics, why, given the lack of data to support theories, economists use such sophisticated mathematics. Arrow replied, "It is just because we do not have enough data that we use sophisticated mathematics. We have to ensure the logical consistency of our arguments." For physicists, on the other hand, explaining empirical data is the best guarantee of the logical consistency of theories. If theories work empirically, then mathematical details are

not so important and will be amended later; if theories do not work empirically, no logical subtlety will improve them.

This anecdote is revealing of one of the key problems that any modeler of economic phenomena has to confront. On the one side, economics is an empirical science based on empirical facts. However, as data are scarce, many theories and models fit the same data. One is tempted to rely on “clear reasoning” to compensate for the scarcity of data. In economics, there is always a tension between the use of pure reasoning to develop *ex ante* economic theories and the need to conform to generally accepted principles of empirical science. The development of high-performance computing has aggravated the problem, making it possible to discover subtle patterns in data and to build models that fit data samples with arbitrary precision. But patterns and models selected in this way are meaningless and reveal no true economic feature.

Given the importance of model selection, let us discuss this issue before actually discussing estimation issues. It is perhaps useful to compare again the methods of economics and of physics. In physics, the process of model choice is largely based on human creativity. Facts and partial theories are accumulated until scientists make a major leap forward, discovering a new unifying theory. Theories are generally expressed through differential equations and often contain constants (i.e., numerical parameters) to be empirically ascertained. Note that the discovery of laws and the determination of constants are separate moments. Theories are often fully developed before the constants are determined; physical constants often survive major theoretical overhauls in the sense that new theories must include the same constants plus, eventually, additional ones.

Physicists are not concerned with problems of “data snooping,” that is, of fitting the data to the same sample that one wants to predict. In general, data are overabundant and models are not determined through a process of fitting

and adaptation. Once a physical law that accurately fits all available data is discovered, scientists are confident that it will fit similar data in the future. The key point is that physical laws are known with a high level of precision. Centuries of theoretical thinking and empirical research have resulted in mathematical models that exhibit an amazing level of correspondence with reality. Any minor discrepancy from predictions to experiments entails a major scientific reevaluation. Often new laws have completely different forms but produce quite similar results. Experiments are devised to choose the winning theory.

Now consider economics, where the conceptual framework is totally different. First, though apparently many data are available, these data come in vastly different patterns. For example, the details of economic development are very different from year to year and from country to country. Asset prices seem to wander about in random ways. Introducing a concept that plays a fundamental role later in this entry, we can state: From the point of view of statistical estimation, economic data are always scarce given the complexity of their patterns.

Attempts to discover simple deterministic laws that accurately fit empirical economic data have proved futile. Furthermore, as economic data are the product of human artifacts, it is reasonable to believe that they will not follow the same laws for very long periods of time. Simply put, the structure of any economy changes too much over time to believe that economic laws are time-invariant laws of nature. One is, therefore, inclined to believe that only approximate laws can be discovered.

However the above considerations create an additional problem: The precise meaning of approximation must be defined. The usual response is to have recourse to probability theory. Here is the reasoning. Economic data are considered one realization of stochastic (i.e., random) data. In particular, economic time series are considered one realization of a stochastic process. The attention of the modeler has

therefore to switch from discovering deterministic paths to determining the time evolution of probability distributions. In physics, this switch was made at the end of the 19th century, with the introduction of statistical physics. It later became an article of scientific faith that we can arrive at no better than a probabilistic description of nature.

The adoption of probability as a descriptive framework is not without a cost: Discovering probabilistic laws with confidence requires working with very large populations (or samples). In physics, this is not a problem as we have very large populations of particles. (Although this statement needs some qualification because physics has now reached the stage where it is possible to experiment with small numbers of elementary particles, it is sufficient for our discussion here.) In economics, however, populations are too small to allow for a safe estimate of probability laws; small changes in the sample induce changes in the laws. We can, therefore, make the following statement: Economic data are too scarce to allow us to make sure probability estimates.

For example, Gopikrishnan, Meyer, Nunes Amaral, and Stanley (1998) conducted a study to determine the distribution of stock returns at short time horizons, from a few minutes to a few days. They found that returns had a power tail distribution with exponent $\alpha \approx 3$. One would expect that the same measurement repeated several times over would give the same result. But this is not the case. Since the publication of the aforementioned paper, the return distribution has been estimated several times, obtaining vastly different results. Each successive measurement was made in *bona fide*, but a slightly different empirical setting produced different results.

As a result of the scarcity of economic data, many statistical models, even simple ones, can be compatible with the same data with roughly the same level of statistical confidence. For example, if we consider stock price processes, many statistical models—including the ran-

dom walk—compete to describe each process with the same level of significance. Before discussing the many issues surrounding model selection and estimation, we will briefly discuss the subject of *machine learning* and the machine-learning approach to modeling.

THE (MACHINE) LEARNING APPROACH TO MODEL SELECTION

There is a fundamental distinction between (1) estimating parameters in a well-defined model and (2) estimating models through a process of learning. Models, as mentioned, are determined by human modelers using their creativity. For example, a modeler might decide that stock returns in a given market are influenced by a set of economic variables and then write a linear model as follows:

$$r_{i,t} = \sum_{k=1}^K \beta_k f_{k,t}$$

where the f are stochastic processes that represent a set of given economic variables. The modeler must then estimate the β_k and test the validity of his model.

In the machine-learning approach to modeling—ultimately a byproduct of the diffusion of computers—the process is the following:

- There is a set of empirical data to explain.
- Data are explained by a family of models that include an unbounded number of parameters.
- Models fit with arbitrary precision any set of data.

That models can fit any given set of data with arbitrary precision is illustrated by *neural networks*, one of the many machine learning tools used to model data that includes *genetic algorithms*. As first demonstrated by Cybenko (1989), neural networks are universal function approximators. If we allow a sufficient number of layers and nodes, a neural network

can approximate any given function with arbitrary precision. The idea of universal function approximators is well known in calculus. The Taylor and Fourier series are universal approximators for broad classes of functions.

Suppose a modeler wants to model the unknown *data generation process* (DGP) of a time series $X(t)$ using a neural network. A DGP is a possibly nonlinear function of the following type:

$$X(t) = F(X(t-1), \dots, X(t-k))$$

that links the present value of the series to its past. A neural network will try to learn the function F using empirical data from the series. If the number of layers and nodes is not constrained, the network can learn F with unlimited precision.

However, the key concept of the theory of machine learning is that a model that can fit any data set with arbitrary precision has no explanatory power, that is, it does not capture any true feature of the data, neither in a de-

terministic setting nor in a statistical setting. In an economic context, machine learning perfectly explains sample data but has no forecasting power. It is only a mathematical device; it does not correspond to any economic property.

We can illustrate this point in a simplified setting. Let us generate an autoregressive trend stationary process according to the following model:

$$X(i) = X(i-1) + \lambda(Di - X(i-1)) + \sigma\varepsilon(i) \\ \lambda = 0.1, \quad D = 0.1, \quad \sigma = 0.5$$

where $\varepsilon(i)$ are normally distributed zero-mean unit-variance random numbers generated with a random number generator. The initial condition is $X = 1$. This process is asymptotically trend stationary. Using the ordinary least squares (OLS) method, let us fit to the process X two polynomials of degree 2 and 20 respectively on a training window of 200 steps. We continue the polynomials five steps after the training window. Figure 1 represents the process plot and the two polynomials. Observe from the

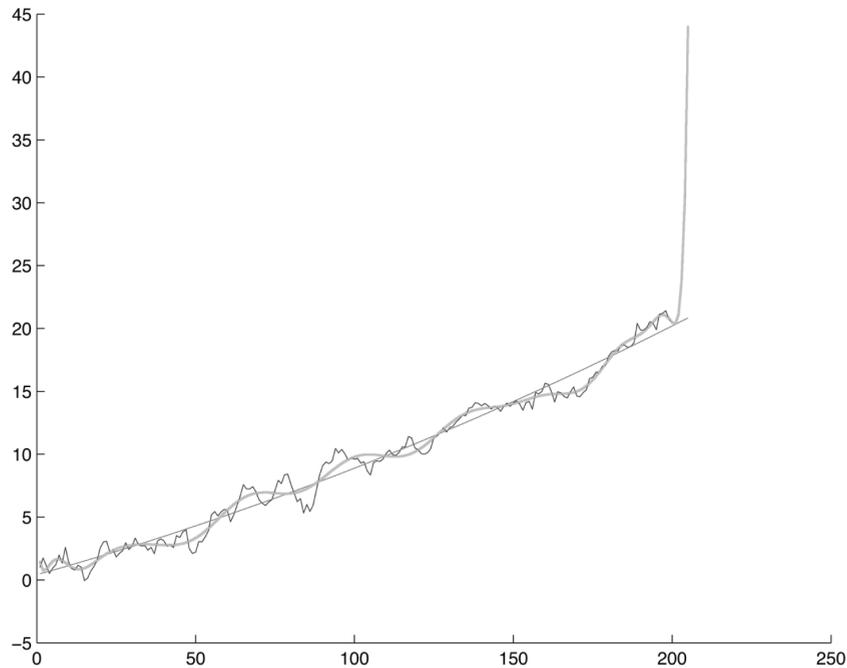


Figure 1 Polynomial Fitting of a Trend Stationary Process Using Two Polynomials of Degree 2 and 20 Respectively on a Training Window of 200 Steps

exhibit the different behavior of the two polynomials. The polynomial of degree 2 essentially repeats the linear trend, while the polynomial of degree 20 follows the random fluctuations of the process quite accurately. Immediately after, however, the training window it diverges.

To address the problem, the theory of machine learning suggests criteria to constrain models so that they fit sample data only partially but, as a trade-off, retain some forecasting power. The intuitive meaning is the following: *The structure of the data and the sample size dictate the complexity of the laws that can be learned by computer algorithms.*

This is a fundamental point. If we have only a small sample data set we can learn only simple patterns, provided that these patterns indeed exist. The theory of machine learning constrains the dimensionality of models to make them adapt to the sample size and structure.

In most practical applications, the theory of machine learning works by introducing a *penalty function* that constrains the models. The penalty function is a function of the size of the sample and of the complexity of the model. One compares models by adding the penalty function to the likelihood function (a definition of the likelihood function is provided later). In this way one can obtain an ideal trade-off between model complexity and forecasting ability.

Several proposals have been made as regards the shape of the penalty function. Three criteria are in general use:

- The Akaike Information Criterion (AIC)
- The Bayesian Information Criterion (BIC) of Schwartz
- The Maximum Description Length principle of Rissanen

More recently, Vapnik and Chervonenkis (1974) have developed a full-fledged quantitative theory of machine learning. While this theory goes well beyond the scope of this book, the practical implication of the theory of learning is important to note: Model complexity must be constrained in function of the sample.

Consider that some “learning” appears in most financial econometric endeavors. For example, determining the number of lags in an autoregressive model is a problem typically solved with methods of learning theory, that is, by selecting the number of lags that minimize the sum of the loss function of the model plus a penalty function. Ultimately, in modern computer-based financial econometrics, there is no clear-cut distinction between a learning approach versus a theory-based a priori approach.

Note, however, that the theory of machine learning offers no guarantee of success. To see this point, let’s generate a random walk and fit two polynomials of degree 3 and 20, respectively. Figure 2 illustrates the random path and the two polynomials. The two polynomials appear to fit the random path quite well. Following the above discussion, the polynomial of order 3 seems to capture some real behavior of the data. But as the data are random, the fit is spurious. This is by no means a special case. In general, it is often possible to fit models to sample data even if the data are basically unpredictable.

Figures 1 and 2 are examples of the simplest cases of model fitting. One might be tempted to object that fitting a curve with a polynomial is not a good modeling strategy for prices or returns. This is true, as one should model a dynamic DGP. However, fitting a DGP implies a multivariate curve fitting. For illustration purposes, we chose the polynomial fitting of a univariate curve: It is easy to visualize and contains all the essential elements of model fitting.

SAMPLE SIZE AND MODEL COMPLEXITY

The four key conclusions reached thus far are

- Economic data are generally scarce for statistical estimation given the complexity of their patterns.
- Economic data are too scarce for sure statistical estimates.

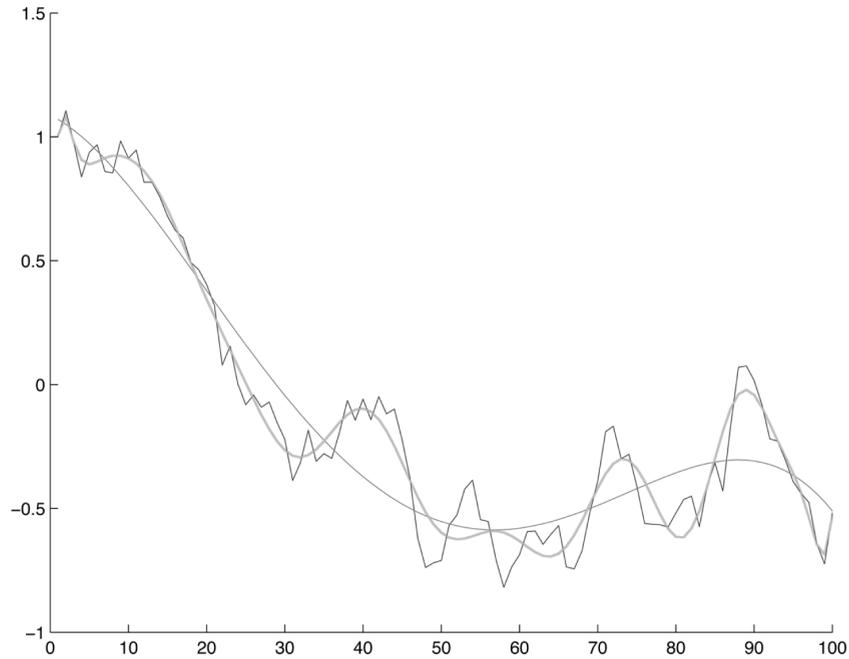


Figure 2 Polynomial Fitting of a Random Walk Using Two Polynomials of Degree 3 and 20 Respectively on a 100-Step Sample

- The scarcity of data means that the data might be compatible with many different models.
- There is a trade-off between model complexity and the size of the data sample.

The last two considerations are critical. To illustrate the quantitative trade-off between the size of a data sample and model complexity, consider an apparently straightforward case: estimating a correlation matrix.

It is well known from the theory of random matrices that the eigenvalues of the correlation matrix of independent random walks are distributed according to the following law:

$$\rho(\lambda) = \frac{Q}{2\pi\sigma^2} \frac{\sqrt{(\lambda_{\max} - \lambda)(\lambda_{\min} - \lambda)}}{\lambda}$$

where Q is the ratio between the number N of sample points and the number M of time series. Figure 3 illustrates the theoretical distribution of eigenvalues for three values of Q : $Q = 1.8$, $Q = 4$, and $Q = 16$.

As can be easily predicted by examining the above formula, the distribution of eigenvalues

is broader when Q is smaller. The corresponding λ_{\max} is larger for the broader distribution. The λ_{\max} are respectively:

$$\lambda_{\max} = 3.0463 \text{ for } Q = 1.8$$

$$\lambda_{\max} = 2.2500 \text{ for } Q = 4$$

$$\lambda_{\max} = 1.5625 \text{ for } Q = 16$$

The eigenvalues of a random matrix do not carry any true correlation information. If we now compute the eigenvalues of an empirical correlation matrix of asset returns with a given Q (i.e., the ratio between number of samples and the number of series), we find that only a few eigenvalues carry information as they are outside the area of pure randomness corresponding to the Q . In fact, with good approximation, λ_{\max} is the cut-off point that separates meaningful correlation information from noise. (The application of random matrices to the estimation of correlation and covariance matrices is developed in Plerou, Gopikrishnan, Rosenow, Nunes Amaral, Guhr, and Stanley [2002].) Therefore, as the ratio of sample points to the number of asset

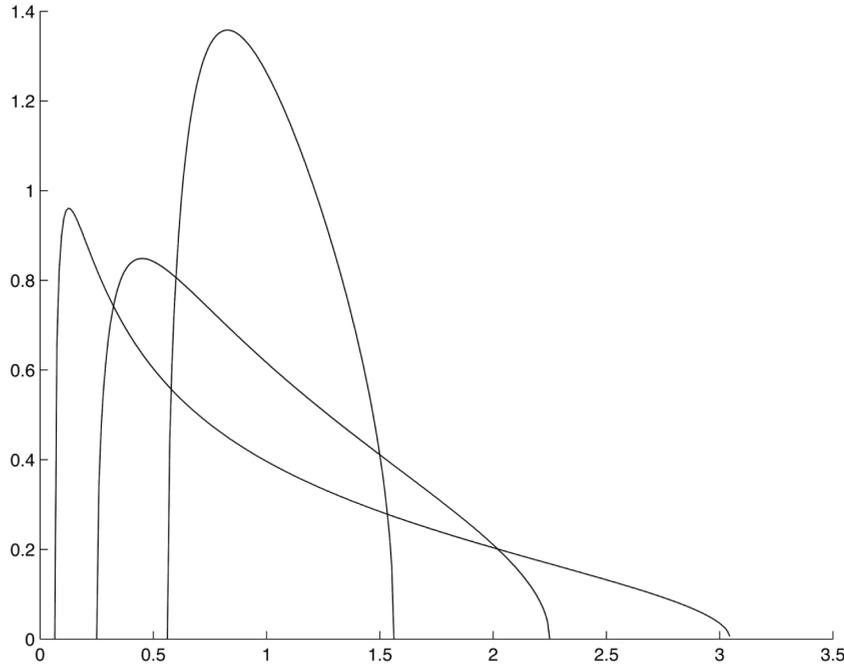


Figure 3 The Theoretical Distribution of Eigenvalues for Three Values of Q : $Q = 1.8$, $Q = 4$, and $Q = 16$

prices grows (i.e., we have more points for each price process) the “noise area” gets smaller.

To show the effects of the ratio Q on the estimation of empirical correlation matrices, let’s compute the correlation matrix for three sets of 900, 400, and 100 stock prices that appeared in the MSCI Europe in a six-year period from December 1998 to February 2005. The return series contain in total 1,611 sample points, each corresponding to a trading day.

First we compute the correlation matrices. The average correlation (excluding the diagonal) is approximately 10% for the three sets of 100, 400, and 900 stocks. Then we compute the eigenvalues. The plot of sorted eigenvalues for the three samples is shown in Figures 4, 5, and 6. One can see from these exhibits that when the ratio Q is equal to 16 (i.e., we have more sample points per stock price process), the plot of eigenvalues decays more slowly.

Now compare the distribution of empirical eigenvalues with the theoretical cut-off point λ_{\max} that we computed above. The parameter Q was chosen to approximately represent the ra-

tios between 1,611 sample points and 100, 400, and 900 stocks. Results are tabulated in Table 1. This exhibit shows that the percentage of meaningful eigenvalues grows as the ratio between the number of sample points and the number of processes increases. If we hold the number of sample points constant (i.e., 1,611) and increase the number of time series from 100 to 900, a larger percentage of eigenvalues becomes essentially noise (i.e., they do not carry information). Obviously the number of meaningful eigenvalues increases with the number of series, but, due to loss of information, it does so more slowly than does the number of series due to loss of information.

Two main conclusions can be drawn from Table 1:

- Meaningful eigenvalues represent a small percentage of the total, even when $Q = 16$.
- The ratio of meaningful eigenvalues to the total grows with Q , but the gain is not linear.

The above considerations apply to estimating a correlation matrix. As we will see, however,

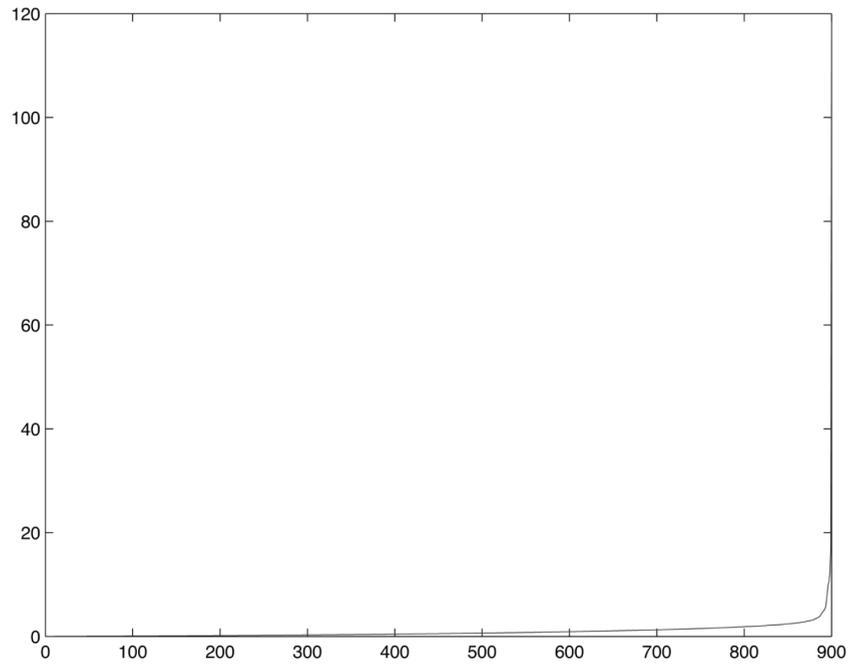


Figure 4 Plot of Eigenvalues for 900 Prices, $Q = 1.8$

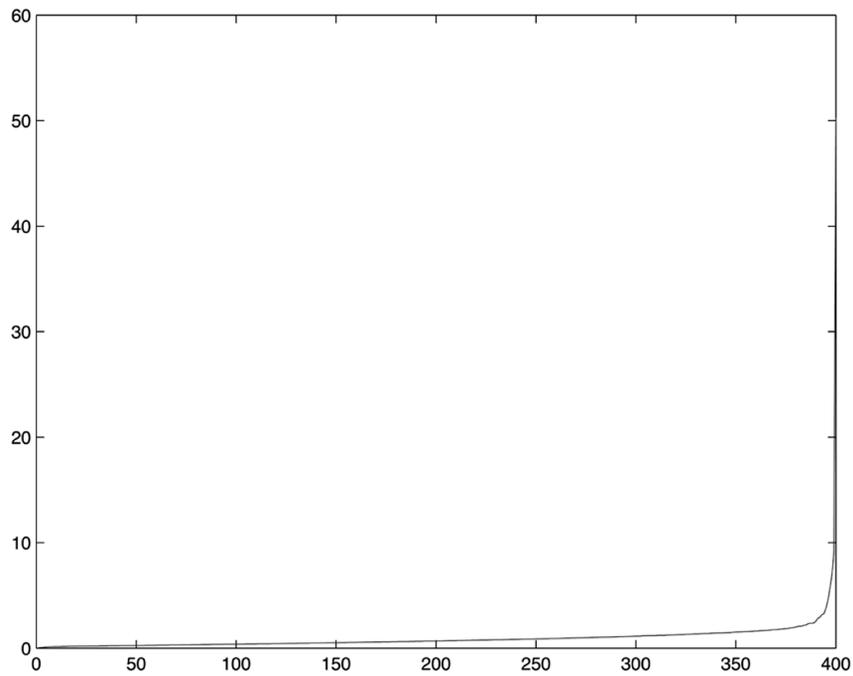


Figure 5 Plot of Eigenvalues for 400 Prices, $Q = 4$

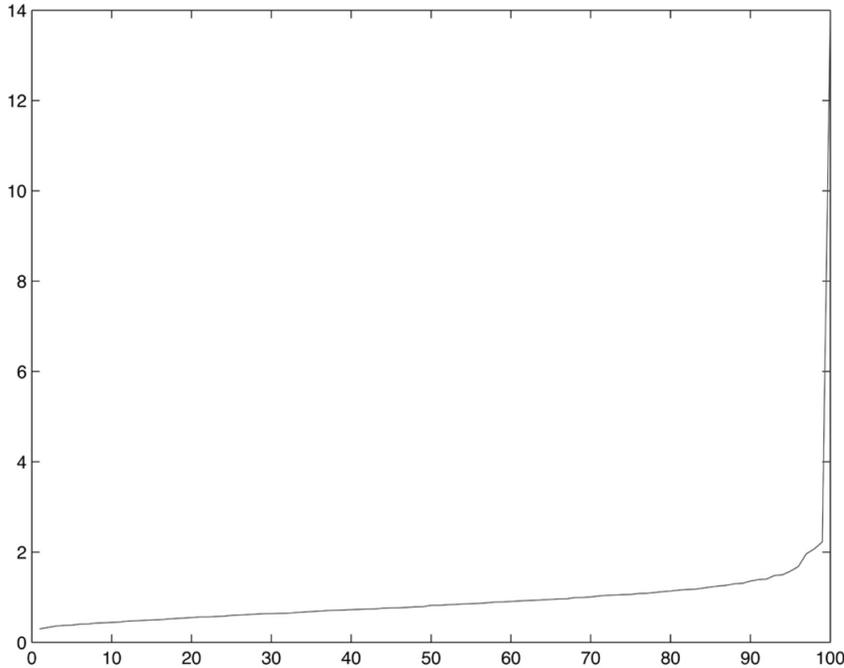


Figure 6 Plot of Eigenvalues for 100 Prices, $Q = 16$

Table 1 Comparison of the Distribution of Empirical Eigenvalues with the Theoretical Cutoff Point for Different Values of Q

Number of processes	Average correlation	Max eigenvalue	Number of meaningful eigenvalues	Percentage of meaningful eigenvalues
900; $Q = 1.8$	10%	118	26	0.029
400; $Q = 4$	9.5%	50	15	0.038
100; $Q = 16$	9.8%	14	6	0.06

they carry over, at least qualitatively, to the estimation of any linear dynamic model. In fact, the estimation of linear dynamic models is based on estimating correlation and covariance matrices.

DANGEROUS PATTERNS OF BEHAVIOR

One of the most serious mistakes that a financial modeler can make is to look for rare or unique patterns that look profitable in-sample but produce losses out-of-sample. This mistake is made easy by the availability of powerful computers that can explore large amounts of data: Any

large data set contains a huge number of patterns, many of which look very profitable. Otherwise expressed, any large set of data, even if randomly generated, can be represented by models that appear to produce large profits. To see the point, perform the following simple experiment. Using a good random number generator, generate a large number of independent random walks with zero drift. In sample, these random walks exhibit large profit opportunities. There are numerous reasons for this. In fact, if we perform a sufficiently large number of simulations, we will generate a number of paths that are arbitrarily close to any path we want. Many paths will look autocorrelated and will

be indistinguishable from trend-stationary processes. In addition, many stochastic trends will be indistinguishable from deterministic drifts.

There is nothing surprising in the above phenomena. A stochastic process or a discrete time series is formed by all possible paths. For example, a trend-stationary process and a random walk are formed by the same paths. What makes the difference between a trend-stationary process and a random walk are not the paths—which are exactly the same—but the probability assignments. Suppose processes are discrete, for example because time is discrete and prices move by only discrete amounts. Any computer simulation is ultimately a discrete process, though the granularity of the process is very small. In this discrete case, we can assign a discrete probability to each path. The difference between processes is the probability assigned to each path. In a large sample, even low probability paths will occur, albeit in small numbers.

In a very large data set, almost any path will be approximated by some path in the sample. If the computer generates a sufficiently large number of random paths, we will come arbitrarily close to any given path, including, for example, to any path that passes the test for trend stationarity. In any large set of price processes, one will therefore always find numerous interesting paths, such as cointegrated pairs and trend-stationary processes.

To avoid looking for ephemeral patterns, we must stick rigorously to the paradigm of machine learning and statistical tests. This sounds conceptually simple, but it is very difficult to do in practice. It means that we have to decide the level of confidence that we find acceptable and then compute probability distributions for the entire sample. This has somewhat counterintuitive consequences. We illustrate this point using as an example the search for cointegrated pairs; the same reasoning applies to any statistical property.

Suppose that we have to decide whether a given pair of time series is cointegrated or not. We can use one of the many cointegration tests.

If the time series are short, no test will be convincing; the longer the time series, the more convincing the test. The problem with economic data is that no test is really convincing as the confidence level is generally in the range of 95% or 99%. Whatever confidence level we choose, given one or a small number of pairs, we decide the cointegration properties of each pair individually. For example, in macroeconomic studies where only a few time series are given, we decide if a given pair of time series is cointegrated or not by looking at the cointegration test for that pair.

Does having a large number of data series, for example 500 price time series, require any change in the testing methodology? The answer, in fact, is that *additional* care is required: In a large data set, for the reasons we outlined above, any pattern can be approximated. One has to look at the probability that a pattern will appear in that data set. In the example of cointegration, if one finds, say, ten cointegrated pairs in 500 time series, the question to ask is: What is the probability that in 500 time series 10 time series are cointegrated? Answering this question is not easy because the properties of pairs are not independent. In fact, given three series a , b , and c we can form three distinct pairs whose cointegration properties are not, however, mutually independent. This makes calculations difficult.

To illustrate the above, let us generate a simulated random walk using the following formula:

$$\begin{aligned} \mathbf{X}(i) &= \mathbf{X}(i-1) + \varepsilon(i) \\ \mathbf{X}(1) &= 1 \end{aligned}$$

where $\mathbf{X}(i)$ is a random vector with 500 elements, and the noise term is generated with a random number generator as 500 independent normally distributed zero-mean unitary-variance numbers. Now run simulations for 500 steps. Next, eliminate linear trends from each realization. (Cointegration tests can handle linear trends. We detrended for clarity of illustrations.) A sample of three typical realizations of

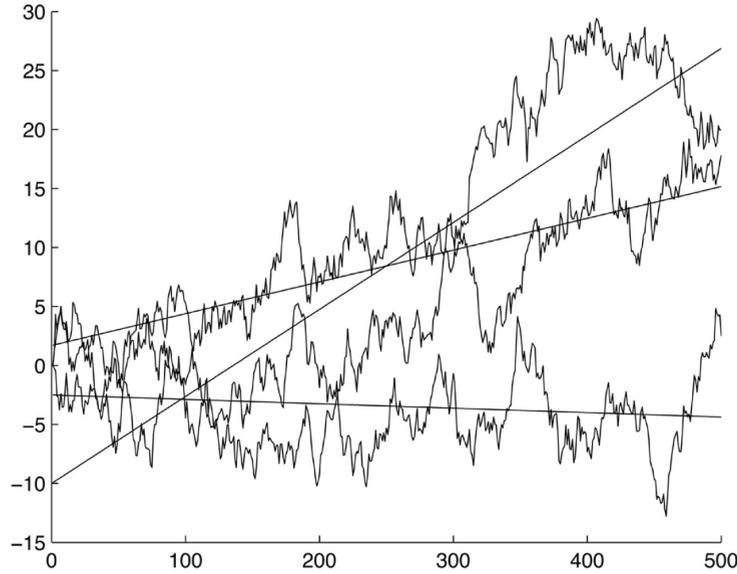


Figure 7 A Sample of Three Typical Realizations of a 500-Step Random Walk with Their Trends

the random walks is illustrated in Figure 7 and the corresponding residuals after detrending in Figure 8.

Now run the cointegration test at a 99% confidence level on each possible pair. In a sample of 10 simulation runs, we obtain the following

number of pairs that pass the cointegration test: 74, 75, 89, 73, 65, 91, 91, 93, 84, 62. There are in total

$$\binom{500}{2} = \frac{500 \times 499}{2} = 124,750 \text{ distinct pairs}$$

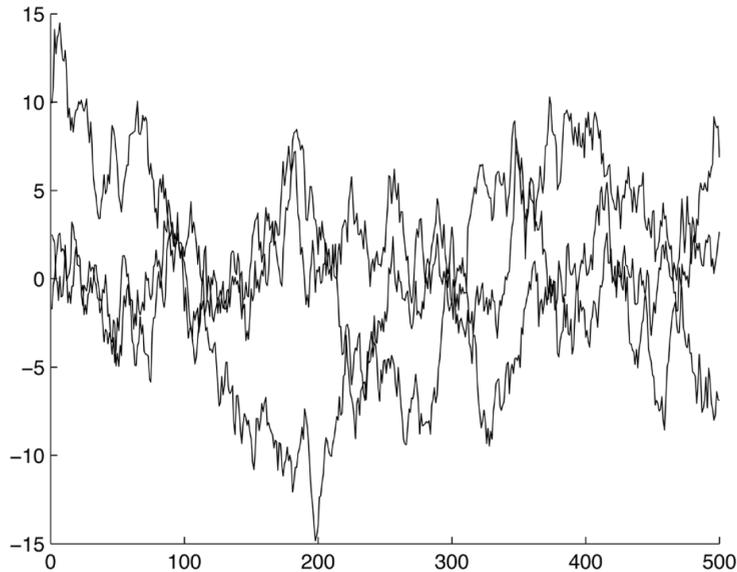


Figure 8 The Residuals of the Same Random Walks after Detrending

If cointegration properties of pairs were independent, given 500 random walks, on the average we should find 124 pairs that pass the cointegration test at the 99% confidence level. However, cointegration properties of pairs are not independent for the reasons mentioned above. This explains why we obtained a smaller number of pairs than expected. This example illustrates the usefulness of running Monte Carlo experiments to determine the number of cointegrated pairs found in random walks.

If, however, the patterns we are looking for are all independent, calculations are relatively straightforward. Suppose we are looking for stationary series applying an appropriate test at a 99% confidence interval. This means that a sample random walk has a 1% probability of passing the test (i.e., to be wrongly accepted as stationary) and a 99% probability of being correctly rejected. In this case, the probability distribution of the number of paths that pass the stationarity test given a sample of 500 generated random walks is a binomial distribution with probabilities $p = 0.01$ and $q = 1 - p = 0.99$ and mean 5.

We apply criteria of this type very often in our professional and private lives. For example, suppose that an inspector has to decide whether to accept or reject a supply of spare parts. The inspector knows that on average one part in 100 is defective. He randomly chooses a part in a lot of 100 parts. If the part is defective, he is likely to ask for additional tests before accepting the lot. Suppose now that he tests 100 parts from 100 different lots of 100 parts and finds only one defective part. He is likely to accept the 100 lots because the incidence of faulty parts is what he expected it to be, that is, one in 100. The point is that we are looking for statistical properties, not real identifiable objects.

A profitable price time series is not a recognizable object. We find what seems to be a profitable time series but we cannot draw any conclusion because the level of the “authenticity test” of each series is low. When looking at very large data sets, we have to make data work for

us and not against us, examining the entire sample. For example, consider a strategy known as “pair trading.” In this strategy, an investor selects pairs from a stock universe and maintains a market neutral (i.e., zero beta) long-short portfolio of several pairs of stocks with a mean-reverting spread. When there are imbalances in the market causing the spread to diverge, the investor seeks to determine the reason for the divergence. If the investor believes that the spread will revert, he or she takes a position in the two stocks to capitalize on the reversion. A modeler who would define a pair trading strategy based on the cointegrated pair in the previous example would be disappointed. Based on extensive Monte Carlo simulations to compare the number of cointegrated pairs among the stocks in the S&P 500 index for the period 2001–2004 and in computer-generated random walks, the number of cointegrated pairs we found was slightly larger in the real series than in the simulated random walks.

We can conclude that it is always good practice is to test any model or pattern recognition method against a surrogated random sample generated with the same statistical characteristics as the empirical ones. For example, it is always good practice to test any model and any strategy intended to find excess returns on a set of computer-generated random walks. If the proposed strategy finds profit in computer-generated random walks, it is highly advisable to rethink the strategy.

DATA SNOOPING

Given the scarcity of data and the basically uncertain nature of any econometric model, it is generally required to calibrate models on some data set, the so-called *training set*, and test them on another data set, the *test set*. In other words, it is necessary to perform an out-of-sample validation on a separate test set. The rationale for this procedure is that any machine-learning process—or even the calibration mechanism itself—is a heuristic methodology, not

a true discovery process. Models determined through a machine-learning process must be checked against the reality of out-of-sample validation. Failure to do so is referred to as *data snooping*, that is, performing training and tests on the same data set.

Out-of-sample validation is typical of machine-learning methods. Learning entails models with unbounded capabilities of approximation constrained by somewhat artificial mechanisms such as a penalty function. This learning mechanism is often effective but there is no guarantee that it will produce a good model. Therefore, the learning process is considered *discovery heuristics*. The true validation test, say the experiments, has to be performed on the test set. Needless to say, the test set must be large and cover all possible patterns, at least in some approximate sense. For example, in order to test a trading strategy one would need to test data in many different market conditions: with high volatility and low volatility, in expansionary and recessionary economic periods, under different correlation situations, and so on.

Data snooping is not always easy to understand or detect. Suppose that a modeler wants to build the DGP of a time series. A DGP is often embodied in a set of difference equations with parameters to be estimated. Suppose that four years of data of a set of time series are available. A modeler might be tempted to use the entire four years to perform a “robust” model calibration and to “test” the model on the last year. This is an example of data snooping that might be difficult to recognize and to avoid. In fact, one might (erroneously) reason as follows. If there is a true DGP, it is more likely that it is “discovered” on a four-year sample than on shorter samples. If there is a true DGP, data snooping is basically innocuous and it is therefore correct to use the entire data set. On the other hand, if there is no stable DGP, then it does not make sense to calibrate models as their coefficients would be basically noise.

This reasoning is wrong. In general, there is no guarantee that, even if a true DGP exists, a learning algorithm will learn it. Among the reasons for learning failure are (1) the slow convergence of algorithms which might require more data than that available, and (2) the possibility of getting stuck in local optima. However, the real danger is the possibility that no true DGP exists. Should this be the case, the learning algorithm might converge to a false solution or not converge at all. We illustrated this fact earlier in this entry where we showed how it is possible to successfully fit a low dimensionality polynomial to a randomly generated path.

There are other forms of data snooping. Suppose that a modeling team works on a sample of stock price data to find a profitable trading strategy. Suppose that they respect all of the above criteria of separation of the training set and the data set. Different strategies are tried and those that do not perform are rejected. Though sound criteria are used, there is still the possibility that by trial and error the team hits a strategy that performs well in sample but poorly when applied in the real world. Another form of hidden data snooping is when a methodology is finely calibrated to sample data. Again, there is the possibility that by trial and error one finds a calibration parameterization that works well in sample and poorly in the real world.

There is no sound theoretical way to avoid this problem *ex ante*. In practice, the answer is to separate the sets of training data and test data, and to decide on the existence of a DGP in function of performance on the test data. However, this type of procedure requires a lot of data. “Resampling” techniques have been proposed to alleviate the problem. Intuitively, the idea behind resampling methods is that a stable DGP calibrated on any portion of the data should work on the remaining data. Widely used resampling techniques include “leave-one-out” and “bootstrapping.” The bootstrap technique creates surrogated data from the initial sample data. (The bootstrap is an important technique but its description goes beyond the scope of this

entry. For a review of bootstrapping, see Davison and Hinkley [1997].)

Data snooping is a defect of training processes which must be controlled but which is very difficult to avoid given the size of data samples currently available. Suppose samples in the range of ten years are available. (Technically much longer data sets on financial markets, up to 50 years of price data, are available. While useful for some applications, these data are useless for most asset management applications given the changes in the structure of the economy.) One can partition these data and perform a single test free from data snooping biases. However, if the test fails, one has to start all over again and design a new strategy. The process of redesigning the modeling strategy might have to be repeated several times over before an acceptable solution is found. Inevitably, repeating the process on the same data includes the risk of data snooping. The real danger in data snooping is the possibility that by trial and error or by optimization, one hits upon a model that casually performs well on the sample data but that will perform poorly in real-world forecasts. Fabozzi, Focardi, and Ma (2005) explore at length different ways in which data snooping and other biases might enter the model discovery process and propose a methodology to minimize the risk of biases, as will be explained in the last section of this entry.

SURVIVORSHIP BIASES AND OTHER SAMPLE DEFECTS

We now examine possible defects of the sample data themselves. In addition to errors and missing data, one of the most common (and dangerous) defects of sample data are the so-called *survivorship biases*. The survivorship bias is a consequence of selecting time series, in particular asset price time series, based on criteria that apply at the end of the period. For example, suppose a sample contains 10 years of price data for all stocks that are in the S&P 500 today and that existed for the last 10 years. This sample, ap-

parently well formed, is, however, biased: The selection, in fact, is made on the stocks of companies that are in the S&P 500 today, that is, those companies that have “survived” in sufficiently good shape to still be in the S&P 500 aggregate. The bias comes from the fact that many of the surviving companies successfully passed through some difficult period. Surviving the difficulty is a form of reversion to the mean that produces trading profits. However, at the moment of the crisis it was impossible to predict which companies in difficulty would indeed have survived.

To gauge the importance of the survivorship bias, consider a strategy that goes short on a fraction of the assets with the highest price and long on the corresponding fraction with the lowest price. This strategy might appear highly profitable in sample. Looking at the behavior of this strategy, however, it becomes clear that profits are very large in the central region of the sample and disappear approaching the present day. This behavior should raise flags. Although any valid trading strategy will have good and bad periods, profit reduction when approaching the present day should command heightened attention.

Avoiding the survivorship bias seems simple in principle: It might seem sufficient to base any sample selection at the moment where the forecast begins, so that no invalid information enters the strategy prior to trading. However, the fact that companies are founded, merged, and closed plays havoc with simple models. In fact, calibrating a simple model requires data of assets that exist over the entire training period. This in itself introduces a potentially substantial training bias.

A simple model cannot handle processes that start or end in the middle of the training period. On the other hand, models that take into account the foundation or closing of firms cannot be simple. Consider, for example, a simple linear autoregressive model. Any addition or deletion of companies introduces a nonlinearity in the model and precludes using standard tools such as the OLS method.

There is no ideal solution. Care is required in estimating possible performance biases consequent to sample biases. Suppose that we make a forecast of return processes based on models trained on the past three or four years of returns data on the same processes that we want to forecast. Clearly there is no data snooping, as we use only information available prior to forecasting. However, it should be understood that we are estimating our models on data that contain biases. If the selection of companies to forecast is subject to strong criteria, for example companies that belong to a major index, it is likely that the model will suffer a loss of performance. This is due to the fact that models will be trained on spurious past performance. If the modeler is constrained to work on a specific stock selection, for example because he has to create an active strategy against a selected benchmark, he might want to consider Bayesian techniques to reduce the biases.

The survivorship bias is not the only possible bias of sample data. More in general, any selection of data contains some bias. Some of these biases are intentional. For example, selecting large caps or small caps introduces special behavioral biases that are intentional. However, other selection biases are more difficult to appreciate. In general, any selection based on belonging to indexes introduces index-specific biases in addition to the survivorship bias. Consider that presently thousands of indexes are in use—the FTSE alone has created some 60,000. Institutional investors and their consultants use these indexes to create asset allocation strategies and then give the indexes to asset managers for active management.

Anyone creating active management strategies based on these indexes should be aware of the biases inherent in the indexes when building their strategies. Data snooping applied to carefully crafted stock selection can result in poor performance because the asset selection process inherent in the index formation process can produce very good results in sample; these results vanish out-of-sample as “snow under the sun.”

MOVING TRAINING WINDOWS

Thus far we assumed that the DGP exists as a time-invariant model. Can we also assume that the DGP varies and that it can be estimated on a moving window? If yes, how can it be tested? These are complex questions that do not admit an easy answer. It is often assumed that the economy undergoes “structural breaks” or “regime shifts” (i.e., that the economy undergoes discrete changes at fixed or random time points).

If the economy is indeed subject to breaks or shifts and the time between breaks is long, models would perform well for a while and then, at the point of the break, performance would degrade until a new model is learned. If regime changes are frequent and the interval between the changes short, one could use a model that includes the changes. The result is typically a nonlinear model such as the Markov-switching models. Estimating models of this type is very onerous given the nonlinearities inherent in the model and the long training period required.

There is, however, another possibility that is common in modeling. Consider a model that has a defined structure, for example a linear VAR model, but whose coefficients are allowed to change in time with the moving of the training window. In practice, most models used work in this way as they are periodically recalibrated. The rationale of this strategy is that models are assumed to be approximate and sufficiently stable for only short periods of time. Clearly there is a trade-off between the advantage of using long training sets and the disadvantage that a long training set includes too much change.

Intuitively, if model coefficients change rapidly, this means that the model coefficients are noisy and do not carry genuine information. We have seen an example above in the simple case of estimating a correlation matrix. Therefore, it is not sufficient to simply reestimate the model: One must determine how to separate the noise from the information in the coefficients.

For example, a large VAR model used to represent prices or returns will generally be unstable. It would not make sense to reestimate the model frequently; one should first reduce model dimensionality with, for example, factor analysis. Once model dimensionality has been reduced, coefficients should change slowly. If they continue to change rapidly, the model structure cannot be considered appropriate. One might, for example, have ignored fat tails or essential nonlinearities.

How can we quantitatively estimate an acceptable rate of change for model coefficients? Are we introducing a special form of data snooping in calibrating the training window? Clearly the answer depends on the nature of the true DGP—assuming that one exists. It is easy to construct artificially DGPs that change slowly in time so that the learning process can progressively adapt to them. It is also easy to construct true DGPs that will play havoc with any method based on a moving training window. For example, if one constructs a linear model where coefficients change systematically at a frequency comparable with a minimum training window, it will not be possible to estimate the process as a linear model estimated on a moving window.

Calibrating a training window is clearly an empirical question. However, it is easy to see that calibration can introduce a subtle form of data snooping. Suppose a rather long set of time series is given, say six to eight years, and that one selects a family of models to capture the DGP of the series and to build an investment strategy. Testing the strategy calls for calibrating a moving window. Different moving windows are tested. Even if training and test data are kept separate so that forecasts are never performed on the training data, clearly the methodology is tested on the same data on which the models are learned.

Other problems with data snooping stem from the psychology of modeling. A key precept that helps to avoid biases is the following: Modeling hunches should be based on theoretical reasoning and not on looking at the data.

This statement might seem inimical to an empirical enterprise, an example of the danger of “clear reasoning” mentioned above. Still, it is true that by looking at data too long one might develop hunches that are sample-specific. There is some tension between looking at empirical data to discover how they behave and avoiding to capture the idiosyncratic behavior of the available data.

In his best-seller *Chaos: Making a New Science*, James Gleick (1987) reports that one of the initiators of chaos theory used to spend long hours flying planes (at his own expense) just to contemplate clouds to develop a feeling for their chaotic movement. Obviously there is no danger of data snooping in this case as there are plenty of clouds on which any modeling idea can be tested. In other cases, important discoveries have been made working on relatively small data samples. The 20th-century English hydrologist Harold Hurst developed his ideas of rescaled range analysis from the yearly behavior of the Nile River, approximately 500 years of sample data, not a huge data sample.

Clearly simplicity (i.e., having only a small number of parameters to calibrate) is a virtue in modeling. A simple model that works well should be favored over a complex model that might produce unpredictable results. Nonlinear models in particular are always subject to the danger of unpredictable chaotic behavior. It was a surprising discovery that even simple maps originate highly complex behavior. The conclusion is that every step of the discovery process has to be checked for empirical, theoretical, and logical consistency.

MODEL RISK

As we have seen above, any model choice and estimation process might result in biases and poor performance. In other words, any model selection process is subject to *model risk*. One might well ask if it is possible to mitigate model risk. In statistics, there is a long tradition,

initiated by the 18th-century English mathematician Thomas Bayes, of considering uncertain not only individual outcomes but the probability distribution itself. It is therefore natural to see if ideas from Bayesian statistics and related concepts could be applied to mitigate model risk.

A simple idea that is widely used in practice is to take the average of different models. This idea can take different forms. Suppose that we have to estimate a variance-covariance matrix. It makes sense to take radically different estimates such as noisy empirical estimates and capital asset pricing model (CAPM) estimates that only consider covariances with the market portfolio and average. Averaging is done with the principle of *shrinkage*, that is, one does not form a pure average but weights the two matrices with weights a and $1 - a$, choosing a according to some optimality principle. This idea can be extended to dynamic models, weighting all coefficients in a model with a probability distribution. Here we want to make some additional qualitative considerations that lead to strategies in model selection.

There are two principal reasons for applying model risk mitigation. First, we might be uncertain as to which model is best, and so mitigate risk by diversification. Second, perhaps more cogent, we might believe that different models will perform differently under different circumstances. By averaging, we hope to reduce the volatility of our forecasts. It should be clear that averaging model results or working to produce an average model (i.e., averaging coefficients) are two different techniques. The level of difficulty involved is also different.

Averaging results is a simple matter. One estimates different models with different techniques, makes forecasts and then averages the forecasts. This simple idea can be extended to different contexts. For example, in rating stocks one might want to do an exponential averaging over past ratings, so that the proposed rating today is an exponential average of the model rating today and model ratings in the past.

Obviously parameters must be set correctly, which again forces a careful analysis of possible data snooping biases. Whatever the averaging process one uses, the methodology should be carefully checked for statistical consistency. For example, one obtains quite different results applying methodologies based on averaging to stationary or nonstationary processes. The key principle is that averaging is used to eliminate noise, not genuine information.

Averaging models is more difficult than averaging results. In this case, the final result is a single model, which is, in a sense, the average of other models. Shrinkage of the covariance matrix is a simple example of averaging models.

MODEL SELECTION IN A NUTSHELL

It is now time to turn all the caveats into some positive approach to model selection. As remarked in Fabozzi, Focardi, and Ma (2005), any process of model selection must start with strong economic intuition. Data mining and machine learning alone are unlikely to produce significant positive results. The possibility that scientific discovery, and any creative process in general, can be “outsourced” to computers is still far from today’s technological reality. A number of experimental artificial intelligence (AI) programs have indeed shown the ability to “discover” scientific laws. For example, the program KAM developed by Yip (1989) is able to analyze nonlinear dynamic patterns and the program TETRAD developed at Carnegie Mellon is able to discover causal relationships in data (see Glymour, Scheines, Spirtes, and Kelly, 1987). However, practical applications of machine intelligence use AI as a tool to help perform specific tasks.

Economic intuition clearly entails an element of human creativity. As in any other scientific and technological endeavor, it is inherently dependent on individual abilities. Is there a body of true, shared science that any modeler can

use? Or do modelers have to content themselves with only partial and uncertain findings reported in the literature? As of the writing of this book, the answer is probably a bit of both.

One would have a hard time identifying economic laws that have the status of true scientific laws. Principles such as the absence of arbitrage are probably what comes closest to a true scientific law but are not, per se, very useful in finding, say, profitable trading strategies. Most economic findings are of an uncertain nature and are conditional on the structure of the economy or the markets.

It is fair to say that economic intuition is based on a number of broad economic principles plus a set of findings of an uncertain and local nature. Economic findings are statistically validated on a limited sample and probably hold only for a finite time span. Consider, for example, findings such as volatility clustering. One might claim that volatility clustering is ubiquitous and that it holds for every market. In a broad sense this is true. However, no volatility clustering model can claim the status of a law of nature as all volatility clustering models fail to explain some essential fact.

It is often argued that profitable investment strategies can be based only on secret proprietary discoveries. This is probably true but its importance should not be exaggerated. Secrecy is typically inimical to knowledge building. Secrets are also difficult to keep. Historically, the largest secret knowledge-building endeavors were related to military efforts. Some of these efforts were truly gigantic, such as the Manhattan Project to develop the first atomic bomb. Industrial projects of a non-military nature are rarely based on a truly scientific breakthrough. They typically exploit existing knowledge.

Financial econometrics is probably no exception. Proprietary techniques are, in most cases, the application of more or less shared knowledge. There is no record of major economic breakthroughs made in secrecy by investment teams. Some firms have advantages in terms of data. Custodian banks, for example, can ex-

plot data on economic flows that are not available to (or in any case are very expensive for) other entities. Until the recent past, availability of computing power was also a major advantage, reserved to only the biggest Wall Street firms; however, computing power is now a commodity.

As a consequence, it is fair to say that economic intuition can be based on a vast amount of shared knowledge plus some proprietary discovery or interpretation. In the last 25 years, a number of computer methodologies were experimented with in the hope of discovering potentially important sources of profits. Among the most fascinating of these were nonlinear dynamics and chaos theory, as well as neural networks and genetic algorithms. None has lived up to initial expectations. With the maturing of techniques, one discovers that many new proposals are only a different language for existing ideas. In other cases, there is a substantial equivalence between theories.

After using intuition to develop an *ex ante* hypothesis, the process of model selection and calibration begins in earnest. This implies selecting a sample free from biases and determining a quality-control methodology. In the production phase, an independent risk control mechanism will be essential. A key point is that the discovery process should be linear. If at any point the development process does not meet the quality standards, one should resist the temptation of adjusting parameters and go back to develop new economic intuition.

This process implies that there is plenty of economic intuition to work on. The modeler must have many ideas to develop. Ideas might range from the intuition that certain market segments have some specific behavior to the discovery that there are specific patterns of behavior with unexploited opportunities. In some cases it will be the application of ideas that are well known but have never been applied on a large scale.

A special feature of the model selection process is the level of uncertainty and noise.

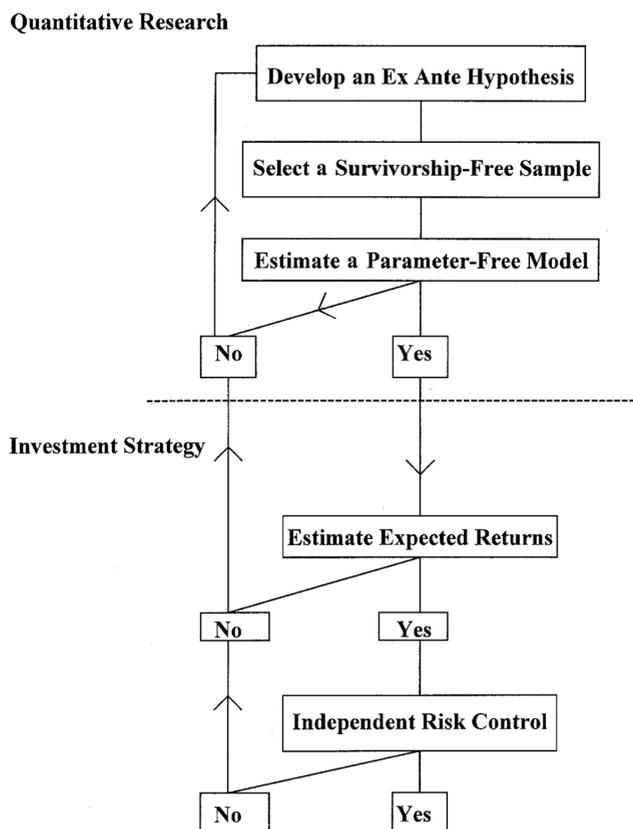


Figure 9 Process of Quantitative Research and Investment Strategy
 Source: Fabozzi, Focardi, and Ma (2005, p. 73)

Models capture small amounts of information in a vast “sea of noise.” Models are always uncertain, and so is their potential longevity. The psychology of discovery plays an important role. These considerations suggest the adoption of a rigorous objective research methodology. Figure 9 illustrates the work flow for a sound process of discovery of profitable strategies. (For a further discussion, see Fabozzi, Focardi, and Ma (2005).)

A modeler working in financial econometrics is always confronted with the risk of finding an artifact that does not, in reality, exist. And, as we have seen, paradoxically one cannot look too hard at the data; this risks introducing biases formed by available but insufficient data sets. Even trying too many possible solutions, one risks falling into the trap of data snooping.

KEY POINTS

- Model selection in financial econometrics requires a blend of theory, creativity, and machine learning.
- The machine-learning approach starts with a set of empirical data that we want to explain. Data are explained by a family of models that include an unbounded number of parameters and are able to fit data with arbitrary precision.
- There is a trade-off between model complexity and the size of the data sample. To implement this trade-off, ensuring that models have forecasting power, the fitting of sample data is constrained to avoid fitting noise. Constraints are embodied in criteria such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC).

- Economic data are generally scarce given the complexity of their patterns. This scarcity introduces uncertainty as regards our statistical estimates. It means that the data might be compatible with many different models with the same level of statistical confidence.
- A serious mistake in model selection is to look for models that fit rare or unique patterns; such patterns are purely random and lack predictive power.
- Another mistake in model selection is data snooping, that is, fitting models to the same data that we want to explain. A sound model selection approach calls for a separation of sample data and test data: Models are fitted to sample data and tested on test data.
- Because data are scarce, techniques have been devised to make optimal use of data; perhaps the most widely used of such techniques is bootstrapping.
- Financial data are also subject to “survivorship bias,” that is, data are selected using criteria known only a posteriori, for example companies that are presently in the S&P 500. Survivorship bias induces biases in models and results in forecasting errors.
- Model risk is the risk that models are subject to forecasting errors in real data. Techniques to mitigate model risk include Bayesian techniques, averaging/shrinkage, and random coefficient models.
- A sound model selection methodology includes strong theoretical considerations, the

rigorous separation of sample and testing data, and discipline to avoid data snooping.

REFERENCES

- Anderson, P. W., Arrow, K. J., and Pines, D. (eds.) (1998). *The Economy as an Evolving Complex System*. New York: Westview Press.
- Cybenko, G. (1989). Approximations by superpositions of a sigmoidal function. *Mathematics of Control Signals & Systems* 2: 303–314.
- Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.
- Fabozzi, F. J., Focardi, S. M., and Ma, K. C. (2005). Implementable quantitative research and investment strategies. *Journal of Alternative Investment* 8: 71–79.
- Gleick, J. (1987). *Chaos: Making a New Science*. New York: Viking Penguin Books.
- Glymour, C., Scheines, R., Spirtes, P., and Kelly, K. (1987). *Discovering Causal Structure*. Orlando, FL: Academic Press.
- Gopikrishnan, P., Meyer, M., Amaral, L. A., and Stanley, H. E. (1998). Inverse cubic law for the distribution of stock price variations. *The European Physical Journal B*, 3: 139–140.
- Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A., Guhr, T., and Stanley, H. E. (2002). Random matrix approach to cross correlations in financial data. *Physical Review E*, 65, 066126.
- Vapnik, V. N., and Chervonenkis, Y. A. (1974). *Theory of Pattern Recognition* (in Russian). Moscow: Nauka.
- Waldrop, M. (1992). *Complexity*. New York: Simon & Schuster.
- Yip, K. M.-K. (1989). KAM: Automatic Planning and Interpretation of Numerical Experiments Using Geometrical Methods, Ph.D. Thesis, MIT.

Managing the Model Risk with the Methods of the Probabilistic Decision Theory

VACSLAV S. GLUKHOV, PhD

Head of Quantitative Strategies and Data Analytics, Liquidnet Europe Ltd, London, United Kingdom

Abstract: Practical applications of financial models require a proper assessment of the model risk due to uncertainty of the model parameters. Methods of the probabilistic decision theory achieve this objective. Probabilistic decision making starts from the Bayesian inference process, which supplies the posterior distribution of parameters. Bayesian incorporation of priors, or opinions, which influence posterior confidence intervals for the model parameters, is indispensable in real-world financial applications. Then, the utility function is used to evaluate practical implications of uncertainty of parameters by comparing the relative expected values of differing decisions. Probabilistic decision making involves computer simulations in all realistic situations. Still, a complete analytical treatment is possible in simple cases.

Practical applications of financial models require their parameters to be given concrete numerical values. These values are typically fitted to empirical data to ensure that the model predictions match historical observations. Parameter values obtained by such fitting procedures never propagate into the future unchanged: Tracing the model's steps back in time, we find that its parameters are always more or less in error. The convention is that predictions made by the model are better if its parameters are known with better precision.

Thus, financial models are always in error—to an extent. Additional variability of actual outcomes due to models themselves, or *model risks*, can be loosely associated with Knightian uncertainty. Methods of *Bayesian inference* estimate

the extent of this uncertainty, whereas the utility theory helps evaluate relative costs of decisions made under this uncertainty. *Probabilistic decision theory*, which combines Bayesian inference with the concept of utility, is the natural and powerful tool for handling intrinsic risks of financial models. The purpose of this chapter is to demonstrate how it works in practice.

AN OUTLINE OF PROBABLISTIC DECISION THEORY

As McKay (2008) cleverly puts it, probabilistic decision theory is trivial—apart from computational details. It has its roots in the Bayesian

inference and in the concept of the utility, or the loss, function. Bayesian inference with its pure probabilistic methods is now gaining its long-deserved position in financial applications.

The *utility function* $U : d \rightarrow V$ that maps the outcomes of possible decisions d onto the value space (or, conversely, the cost space) V is a concept that embodies personal choice and individual risk preferences. In its simplest form, the cost space is one-dimensional. This makes it possible to order decisions by their costs. The decision that has a minimum cost (or a maximum value) is the best decision in the sense of the utility function U .

We will proceed to the formulation of the probabilistic decision-making theory according to Jaynes (2003) and McKay (2008). If $E(\cdot)$ is the expectation, d is the decision, $U(d)$ is the utility function of the decision, θ is the probable future state of the world, and $P(\theta, d)$ is the probability of θ , possibly influenced by the decision, then the optimal decision that maximizes the expectation of the utility function is

$$d = \arg \max \{ E(U(d)) = \int d \theta U(\theta, d) P(\theta, d) \}$$

In exact sciences, the states of the world θ are represented by objective quantities such as temperature, energy density, barometric pressure, acidity, and the like. Measurements of these quantities are subject to errors whose distribution is often fairly well known from the theory of the underlying physical process. For example, in electronics, the probability of an error of a weak signal is closely linked to the ambient temperature, which is an objective and measurable quantity. In engineering the contribution of side factors can often be accounted for and controlled for to a great degree. The existence of the underlying theory capable of quantitative description of the noise and other factors greatly simplifies decision making under uncertainty in engineering and in other exact sciences in comparison with financial applications.

It is customary to employ the same reasoning in finance. When we talk about “more precise

prediction of volatility” or “an accurate correlation coefficient” we implicitly assume that these quantities and parameters in finance are objective. They are not. Not unless we supply an underlying micro-model derived from the first principles, as we routinely do in exact sciences. In contrast, states of the world θ in finance are not inexact measurements of some “true quantities” linked to natural phenomena. Rather, they are mental constructs, which help us reason about financial phenomena—with more or less success. In financial observations, controlling for other factors is not possible, so the concept of *ceteris paribus* does not exist in nontrivial cases of any practical significance. It is better to think about states of the world in financial applications as relatively stable properties of markets and financial instruments. Depending on circumstances, such mental concepts as volatility, correlations, liquidity, expected time to default, and so on can be regarded as states of the world in finance.

States θ are functions of the model employed $\theta = \theta(M)$. Given the set of observations Y and subjective priors I , each state θ is assigned a probability:

$$P(\theta, d) = P(\theta(M), d | Y, I)$$

Being the function of the model, the data, prior beliefs, and, possibly, the decision, the probability of the state θ encapsulates all that is known to be relevant about the phenomenon under consideration.

Probabilistic inference, apart from very special cases, is often tractable only by computer techniques: $P(\theta | Y)$ has no analytical representation and must be ultimately sampled from the data.

The utility function $U(\theta, d)$ introduces the cost (or utility) of each decision in each state of the world. In academic research, one typically chooses a smooth and convex utility function. This should not necessarily be the case in the real world of financial applications where various smooth and nonsmooth constraints must

be satisfied—such as *risk tolerance*, tax considerations, strict and soft budget constraints.

Note that except for the observable data, all other components in the probabilistic decision-making process are user-dependent—the model, the beliefs, and the utility preferences. In the world of subjective views, there is no universal truth, there are no unconditionally good or poor decisions. All decisions are ultimately conditional on personal preferences.

Let's consider how it works in two simple financial applications: risk management of a simple portfolio and valuation of a risky bond.

MODEL RISK OF A SIMPLE PORTFOLIO

A portfolio manager considers creating an investment vehicle based on the instrument Y . The portfolio manager's objective is to extract as much idiosyncratic alpha as possible from Y while reducing the risks associated with the factor X . Instruments highly correlated with X are available for short selling, or instruments highly negatively correlated with X are available for purchase. There are costs associated with these actions. The portfolio manager has an amount of capital equal to C and access to an abundant and relatively low-risk security Z , which can be used to preserve capital. The objective is to meet investment goals $G(T)$, which include return on capital and risk parameters over a definite time horizon T .

The portfolio manager's decisions are based on prior beliefs and the data. The portfolio manager begins splitting capital among X , Y , and Z such that

$$C = C_X + C_Y + C_Z$$

The allocation of capital is determined by the optimization of the utility function given by:

$$C_X, C_Y, C_Z = \arg \max (E(U(C(T) - C)))$$

Expectations of future returns depend on the model parameters. In the Bayesian decision

framework, the distributions of these parameters are important:

1. Distribution of future returns of X .
2. Uncertainty of knowledge about how Y and X are related.
3. Distribution of idiosyncratic risk of Y after Y 's relationship to X is accounted for.
4. Uncertainty of expectations about future alpha.

In the list, the first risk can be understood as the true risk; the last three risks are the model risks or uncertainty.

Consider the model that links contemporaneous data y_t and x_t in a linear fashion:

$$y_t = \beta x_t + \epsilon_t$$

This model is a simplification of the industry-standard factor *risk model* and is akin to that used in the capital asset pricing theory (Sharpe, 1964), where a similar relationship is defined implicitly, or Fama and French (1992), where several factors are used.

The probability of observing the datum y_t given the unknown parameter of the model β is

$$P(y_t | \beta x_t) = P(y_t - \beta x_t) = P(\epsilon_t)$$

It is customary to select a normal model of the idiosyncratic noise $P(\epsilon_t) \propto N(\mu, \sigma)$ as it is a well-behaving distribution that falls off very fast and which for this reason has all its moments well defined. This, in turn, assists in obtaining clear analytical results with helpful illustrative properties.

One needs to remember, however, that real financial noise is neither normal, nor log-normal: It has fat tails, which can be so poorly behaving that the distribution may not even have its first moment well defined. In the probabilistic decision framework, it is almost never possible to obtain a neat analytical expression for the final result. Consequently, the advantages of the normally distributed noise fade in comparison with more realistic models. Another advantage of the probabilistic framework is that

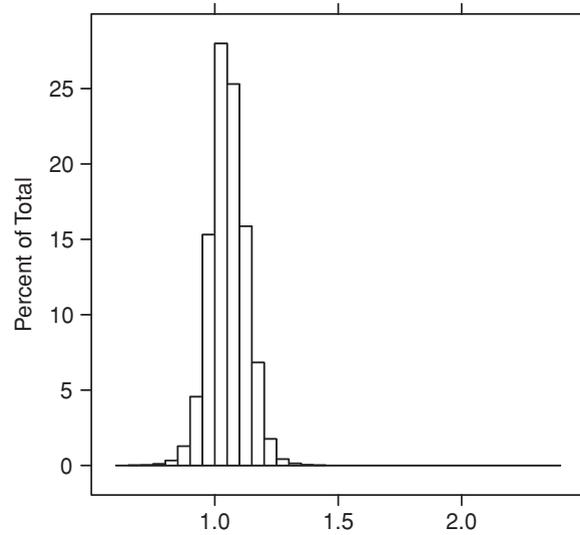


Figure 1 Distribution of β of Daily Price Changes over Three Years for Microsoft Corporation (MSFT) and the S&P 500 EFT (SPY)

one can easily compare evidence in favor of or against any conceivable model. In the presentation here, we retain the normalcy of the residual noise, bearing in mind that it is used for the sole purpose of illustrating the main idea.

Noise values being identically distributed and independent, which again is not a requirement for the probabilistic decision theory, the probability of observing the data set consisting of N points $x_1, y_t, \ell = 1 \dots N$ is

$$\begin{aligned} P(X|\beta X) &= \prod p(\epsilon_1) = \prod P(y_t - \beta x_t) \\ &= \prod P(\epsilon_t) \end{aligned}$$

It is easier to see the properties of the likelihood function by taking the logarithm:

$$\begin{aligned} \log P(Y|\beta X) &= -\frac{\sum_t (y_t - \beta x_t - \mu)^2}{2\sigma^2} \\ &\quad -\frac{1}{2} \log 2\pi\sigma^2 \end{aligned}$$

As a function of β , the log-likelihood attains a maximum at the same point where the ordinary least squares (OLS) method finds its optimum value of $\beta = \beta_{OLS}$. Contrary to the OLS, which boils down all the available data to one number, which is then taken as a real objective quan-

tity, the probabilistic framework retains more information about the relationship between Y and X , thereby preserving it in the distribution $P(\beta|XY)$.

In Figure 1 we show the distribution of β , $P(\beta|XY)$, when the dependent instrument Y is the daily change in the price of Microsoft Corporation stock and the independent instrument X is the daily price change of the exchange-traded fund SPY corresponding to the Standard & Poor's 500 index. Three years of daily data are used in the estimates of $P(\beta|XY)$. In Figure 2 the same amount of data is used to estimate $P(\beta|XY)$ when Y is the daily change in the price of the stock of a natural resource company, the Mosaic Company, and X is, again, the set of contemporaneous daily price changes in SPY.

Having obtained distributions of the model parameters β , μ , and σ from the data, the portfolio manager blends likelihoods with opinions about the distribution of the residual returns. The portfolio manager's alpha model is that the expectation of daily returns of Y is μ_0 with the confidence band $\pm\sigma_0$: $\mu_0 \sim N(\mu_0, \sigma_0)$. Combined with subjective opinions, the

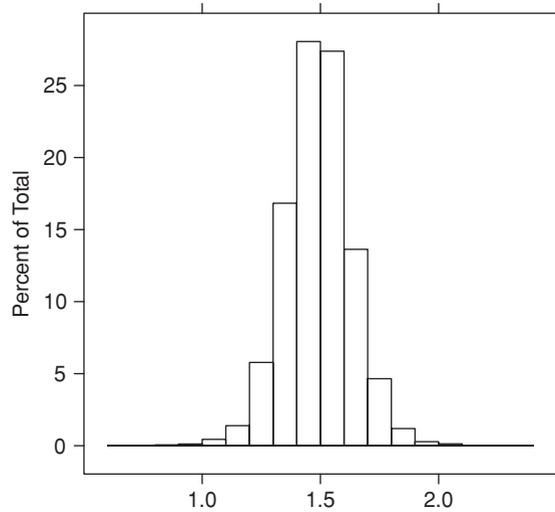


Figure 2 Distribution of β of Daily Price Changes over Three Years for the Mosaic Company (MOS) and the S&P 500 ETF

idiosyncratic distribution is again, normal:

$$\prod P(\epsilon_t) \propto N(\tilde{\mu}, \tilde{\sigma})$$

$$\tilde{\mu} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{\mu N}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}}$$

$$\frac{1}{\tilde{\sigma}^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

In order to overcome the evidence extracted from the data and given by μ , the portfolio manager’s confidence must be greater than the confidence range of the data: The portfolio manager’s confidence is high, that is, when $\sigma_0 \ll \sigma/N$, the posterior expectation of alpha is governed by the portfolio manager’s prognosis. In the opposite case, the data are trusted more than the portfolio manager’s judgment.

The portfolio manager sets risk preferences with the utility function

$$U(C(T), C, \eta) = -\exp\left(-\frac{C(T) - C}{C_\eta}\right)$$

Taking the expectations over one period $T = 1$ we obtain:

$$E(U) = -\int \alpha \beta \exp\left(-\frac{(\tilde{\mu} + \beta \mu_x)w_y + \mu_x w_x + \mu_z w_z}{\eta}\right) + 1/2\eta^2(w_y^2 \tilde{\sigma}^2 + w_x^2 \sigma_x^2 + \sigma_x^z(w_y^2 \beta^2 + 2w_y w_x \beta)) P(\beta)$$

Here

$$w_x = \frac{C_x}{C}, w_z = \frac{C_z}{C}, w_y = 1 - w_x - w_z$$

First, we focus on the problem of optimum allocation when there is no hedging: $w_x = 0, \mu_x = 0$. Define the certainty equivalent (CE) of the investment in Y and Z as such guaranteed change in C that results in the same utility as a risky investment in Y and Z . Mathematically, it is defined as the inverse of the utility function:

$$CE(C(T), C) = U^{-1}(E(U))$$

For the exponential utility function we obtain

$$CE(C(T), C) = -C_\eta \log E(U(C(T), C, \eta))$$

Adopting $P(\beta) = N(\beta_0, \Gamma)$ and integrating expected utility over the model parameter β , we finally arrive at

$$CE(w_y) = \tilde{\mu} w_y + \mu_z(1 - w_y) - \frac{1}{2\eta} \tilde{\sigma}^2 w_y^2 - \frac{1}{2\eta} \frac{\beta_0^2 \sigma_x^2 w_y^2}{1 - \frac{\Gamma^2 \sigma_x^2 w_y^2}{\eta^2}} + \frac{1}{2} \eta \log \left(1 - \frac{\Gamma^2 \sigma_x^2 w_y^2}{\eta^2}\right)$$

The first three terms in this equation represent the certainty equivalent of the investment without the risk model $\beta_0 = 0$ and without the

model risk $\Gamma = 0$. The optimal fraction of the capital invested in Y is a well-known expression (see, for example, Merton, 1969)

$$w_y = \eta \frac{\tilde{\mu} - \mu_z}{\tilde{\sigma}^2}$$

In this case, the fraction invested in the risky instrument is proportional to the portfolio manager's risk tolerance and inversely proportional to the instrument's idiosyncratic risk, which, in the absence of any model, is the total risk of the instrument.

Introduction of the risk model without uncertainty $\beta_0 \neq 0$, $\Gamma = 0$ results in the obvious extension:

$$w_y = \eta \frac{\tilde{\mu} - \mu_z}{\tilde{\sigma}^2 + \beta_0^2 \sigma_x^2}$$

Here $\tilde{\sigma}^2 + \beta_0^2 \sigma_x^2$ is, again, the total risk of Y as given by the model, split into the idiosyncratic part and the part coming from the influence of X .

When $\Gamma \neq 0$, the last two terms in the equation for $CE(w_y)$ represent the model risk. In some situations, the term $\Gamma^2 \sigma_x^2 w_y^2$ can be thought of as the contribution to the expected variance due to the model risk. Indeed, if $\Gamma^2 \sigma_x^2 w_y^2 \ll \eta^2$ (i.e., when the risk tolerance is much greater than possible risk associated with the factor X) in the expression for the certainty equivalent, the model risk is simply added to the total risk:

$$CE(w_y) \approx \tilde{\mu} w_y + \mu_z (1 - w_y) - \frac{1}{2\eta} (\tilde{\sigma}^2 + (\beta_0^2 + \Gamma^2) \sigma_x^2) w_y^2 + O(\sigma_x^1 w_y^1 / \eta^3)$$

In this expression, the last term is proportional to the magnitude of the expression in parentheses and is small in comparison with the preceding terms.

The contribution of the model risk is not so obvious in a general case. Clearly, when $\Gamma^2 \sigma_x^2 \sim \eta^2$, the model risk significantly affects optimal allocations.

Position Hedging

Now the portfolio manager aims to reduce the influence of the factor X on the variability of returns. The portfolio manager adds a position in X to the portfolio. Weight w_x allocated to X is chosen to maximize CE . Positive weight corresponds to a long position in X , whereas a negative weight corresponds to a short position or its equivalent. In the case when X is the daily performance of the Standard & Poor's 500 market index, a short position can be roughly replicated by taking a long position in an exchange-traded fund (ETF) whose daily returns correspond by design to the inverse—up to a constant factor—of the daily performance of the S&P 500 index.

The certainty equivalent of the portfolio is

$$E(w_y) = \tilde{\mu} w_y - \frac{1}{2\eta} \tilde{\sigma}^2 w_y^2 - \frac{1}{2\eta} \frac{\sigma_x^2 (w_x + \beta_0 w_y)^2}{1 - \frac{\Gamma^2 \sigma_x^2 w_y^2}{\eta^2}} + \frac{1}{2} \eta \log \left(1 - \frac{\Gamma^2 \sigma_x^2 w_y^2}{\eta^2} \right)$$

The first two terms in this expression are the idiosyncratic alpha and risk of the instrument Y .

The third term introduces the risk associated with the portfolio returns dependence on X . Let's take a closer look at it. Its structure is similar to the term describing the idiosyncratic risk: variance of the portfolio due to X divided by the portfolio manager's risk tolerance. In the third term, contribution from the risk model comes in two forms. In the numerator $w_x + \beta_0 w_y$ is the total weight of X in the portfolio: the sum of the weight of the position in X , w_x and the estimate of the contribution from exposure to X of the position in Y , $\beta_0 w_y$. The fact that the total contribution of X is the same as in the standard portfolio theory is purely accidental and is due to the choice of the model distribution of β .

In the denominator, the portfolio manager's risk tolerance is augmented by a factor that depends on the uncertainty of β :

$$1 - \frac{\Gamma \sigma_x^2 w_y^2}{\eta^2}$$

This term being less than unity, uncertainty effectively reduces the portfolio manager’s risk tolerance.

The fourth term is the contribution to CE from the model risk. Terms associated with the model risk indicate that when the uncertainty of the model approaches a critical value $\Gamma^2\sigma_x^2w_y^2 \sim \eta^2$ the portfolio becomes unfeasible unless w_y is sufficiently small.

In the absence of a risk model $\beta_0 = 0, \Gamma = 0$ optimal allocations maximizing CE of the portfolio are

$$w_x \sim 0$$

$$w_y \sim \frac{1}{\tilde{\sigma}^2}$$

When the risk model is present, but the uncertainty of the model is much bigger than its prediction $\beta_0 \ll \Gamma$, we obtain another useful result:

$$w_y \sim \frac{1}{\tilde{\sigma}^2 + \Gamma^2\sigma_x^2}$$

In this case the optimal allocation in Y is determined by the total risk of the instrument composed of the idiosyncratic risk and the uncertainty of the model.

When the risk model is present and is absolutely precise $\beta_0 \neq 0, \Gamma = 0$, the usual *hedging ratio* $\frac{w_x}{w_r} = -\beta$ completely eliminates the dependency of portfolio returns and their CE on X —the result conventionally obtained in the traditional formulation of the risk management problem.

From the probabilistic point of view, however, an absolutely precise model is nonsensical. Moreover, situations when both the model’s optimal parameters and the uncertainty of the parameters are of the same order of magnitude are most likely to occur in real applications.

Contribution from the risk model and from the uncertainty of the model become separated and especially simple when the portfolio manager’s risk tolerance is sufficiently large,

$$\Gamma^2\sigma_x^2w_y^2 \ll \eta^2:$$

$$CE(w_y, w_x) \approx \mu w_y - \frac{1}{2\eta}(\sigma^2 - \Gamma^2\sigma_x^2)w_y^2 - \frac{1}{2\eta}\sigma_x^2(w_x - \beta_\mu w_y)^2$$

Note that there is no combination of the instruments Y and X that can eliminate the effect of X . That the effect of the instrument X may never be eliminated completely is a better depiction of the everyday experience of the portfolio manager. Probabilistic decision theory accounting for the model risk, however, gives a reasonable indication of what the portfolio manager can expect from such or another composition of the portfolio when its components are mutually dependent.

In more complicated settings, once the portfolio manager introduces the costs of hedging, the decision whether to hedge or not comes naturally as the consequence of the interplay between the value of hedging and the costs. Let $y|w_xC| = y|\beta_0w_yC|$ be the cost associated with the hedge. Then one should hedge the position if

$$-\frac{1}{2\eta}(\tilde{\sigma}^2 + \Gamma^2\sigma_x^2)w_y^2 - \gamma|\beta_0w_yC| > -\frac{1}{2\eta}\sigma_y^2w_y^2$$

Hedging is justified if the model risk of the hedge plus the cost of implementing the model is smaller than the original risk that the hedge is meant to reduce.

In the equation above all quantities are evaluated from the data and the subjective prior beliefs using the methods of the Bayesian inference. Even when the model and the model parameters are relatively stable, the decision whether to hedge or not to hedge depends on the portfolio manager’s risk tolerance, which in turn can be represented by a combination of external constraints, or be inferred from another model.

A portfolio manager can readily extend the methodology of the preceding sections to more complicated cases of many interrelated instruments and many factors. The probability

distribution of the correlation matrix, however, will not necessarily appear in the calculations in place of the probability distribution of β : Noise models that have no concept of second or higher moments completely rule out correlation matrices in the calculations. Moreover, these distributions naturally lead to decisions being determined by a few extreme outliers. Fortunately, even pathological noise distributions, which seem to be the norm rather than an exception in finance, are treated equally well by the methods of the probabilistic decision theory, which is designed to incorporate all available data plus the portfolio manager's preferences and constraints.

In the next section we will address a problem of the model risk in an investment when the risk profile is different from that of an investment in an equity portfolio.

INVESTMENT IN A RISKY BOND

Let P be the face value of the zero-coupon bond, r the benchmark rate over the period of interest, ρ the multiplicative spread rate for the bond, so that

$$V = \frac{P}{(1+r)(1+\rho)}$$

is the current fair or market price of the bond, possibly unknown. An alternative investment vehicle Z is available as in the previous section, the rate of return for this instrument being r_z .

Let there be two possible states of the world. In the first state the bond is redeemed at the face value at the end of the period. In the second state of the world the bond is redeemed at P_y . The situation when $P_y = 0$ is possible, in which case the investment is a total loss. If the investor purchases N units of the risky bond and the remainder of the capital is preserved in the alternative vehicle, then, at the end of the period, the investor's capital is

$$C_1 = \begin{cases} NP + (1+r_z)(C_0 - NV), & \text{with } (1-p_d) \\ NP_r + (1+r_z)(C_0 - NV), & \text{with } p_d \end{cases}$$

In the traditional formulation the investment is justified if the expected return on capital when $N > 0$ is greater than the expected return when $N = 0$. This translates into the following expression, which links all the input data of the problem and the unknown value of the bond:

$$P(1-p_d) + P_r p_d > (1+r_z)V$$

This traditional approach is a reasonably good approximation under certain conditions. A much richer view along with the set of quantitative tools is required in a general case.

From the probabilistic decision theory viewpoint, the probabilities and other relevant parameters entering the decision-making process must be inferred from the model, from the data, and from the investor's prior beliefs, and are best represented by distributions of possible states of the world. We consider now a simple one-parametric risk model and show how the model risk contributes to the decision process.

Parameter Inference in the Bernoulli Model

In the Bernoulli-like model, the investment vehicle under consideration belongs to a class of essentially similar bonds. They are financial obligations issued by debtors facing essentially the same economic (financial, market, etc.) conditions. Given these conditions, it is customary to assume that the failure of each instrument is a random event. Failures in the class occur with the same probability p_d per unit time, which, for simplicity, will coincide in our analysis with the maturity time of the instrument.

The model of the random process, the empirical data, and the investor's prior beliefs determine all that we know about the model parameter p_d .

Assume that the empirical data are the sample of n observations of the class, and m is the number of cases when a debtor defaults. Adopting a beta-distribution of the model parameter, we obtain the following posterior distribution

given the data and the prior beliefs:

$$\pi(p_d, \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} P_d^{\alpha-1} (1 - p_d)^{\beta-1}$$

where

$$\begin{aligned} \alpha &= \alpha_0 + m \\ \beta &= \beta_0 + n - m \end{aligned}$$

and α_0, β_0 are the parameters representing the investor's prior beliefs. In the prior distribution, α_0 can be interpreted as number of cases of default and β_0 is the number of cases when the bond was repaid in full. The prior distribution's parameters can come from the investor's own experience, or from the consensus of experts, or be inferred from agency ratings. The magnitude of α_0, β_0 versus n, m determines the relative weight the investor assigns to prior beliefs. Prior beliefs dominate the data when $\alpha_0 + \beta_0 \gg n$.

In Figure 3, the investor's prior beliefs follow the prior probability of default 0.1. Parameters of the prior distribution are $\alpha_0 = 2, \beta_0 = 11$. The newly arriving data point to the probability of default 0.2. Observe the change in the shape of the distribution: Its mode moves from ~ 0.1

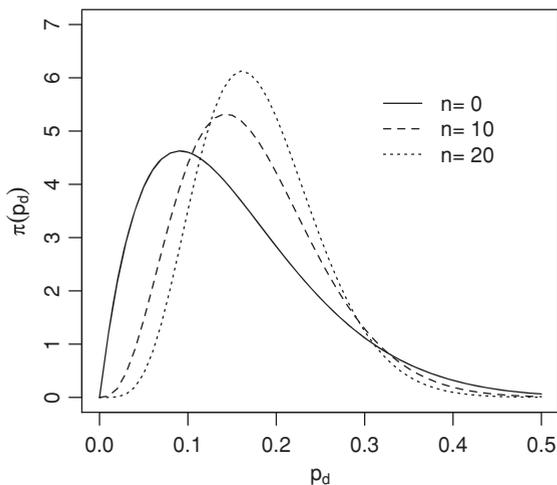


Figure 3 Distribution of the Probability of Default p_d
 Note: Prior distribution is defined by $\alpha_0 = 2, \beta_0 = 11$. Newly arriving data follow the new probability $\frac{n_i}{n} = 0.2$.

to ~ 0.2 as the new data gradually overcome the investor's prior beliefs. Note that the model risk—the width of the distribution—remains relatively high.

In the Bayesian perspective, the distribution of the *probability of default* is a convenient vehicle that carries all that the investor knows from the set of observations and the investor's prior beliefs: what is the most probable state of the world and what is the spread of possible states of the world given the investor's choice of the model of the world.

The rich framework offered by the Bayesian inference of the probability of default consequently brings in a rich set of valuation methods that naturally account for the model risk. In the next sections we will study the valuation effects of the risk of models.

Model Risk Contribution to the Fair Price of the Bond

First, we obtain an interesting estimate of the model risk contribution to the fair price of the bond under the assumption of the infinite risk tolerance. This is a degenerate case most closely resembling the traditional formulation. The utility function is linear if the investor's risk tolerance is infinite.

We obtain formally:

$$P(1 - E(p_d)) + P_r E(p_d) > (1 + r_z)V$$

Assume that the sample size is n of which there are m defaults. A flat prior distribution $\pi(p_d, \alpha_0, \beta_0) = \text{const}$ describes an investor who initially is ignorant. Expectation of the probability of default is then governed by the rule of succession (originally developed by Laplace):

$$E(p_d) = \frac{m + 1}{n + 2}$$

The difference between this posterior expectation and the naïve probability of default $p_d = m/n$ is the contribution of the model risk to the

fair price of the bond:

$$\delta V = \frac{P - P_r}{1 + Y_z} \left\{ E(p_\alpha) - \frac{m}{n} \right\}$$

For example, if the naïve default rate estimate is 0.1 and it is based on 100 observations, the contribution of the model risk to the fair price of the bond can be as big as 78 basis points—not an insignificant amount: The model risk can be a substantial contributor to the overall risk of the investment. Thus, the sampling risk and the prior beliefs bias yield a substantial contribution to the overall risk of the investment.

Even in the simplest Bernoulli-like model, the contribution of the model risk to the value of the bond is nonnegligible. This contribution is especially pronounced when the probability of default is small.

Now we will proceed to a case when the investor's risk tolerance is not infinite. We will show that average probabilities are likely insufficient for making an informed investment decision. Relying on just expected probabilities can result in catastrophic consequences for the investor.

Model Risk of Agency Ratings

Currently financial regulators recommend that expected losses be quantified as the expected probability of default times the exposure at default (see Basel, 2008). Consequently, credit scoring and rating agencies aim at developing models that generate expected probabilities of default. These models are calibrated by minimizing the difference between predicted and empirically observed probabilities of default (see, for example, Korablev and Dwyer, 2007). From the preceding section, it follows that the average rates based on thousands of credit events used in the calibration of the agency model alone are insufficient for making investment decisions concerning a portfolio of an arbitrary, possibly small, subset of instruments. Moreover, the naïve probability of default is likely to be useless in the valuation of a

singular derivative instrument, such as a credit default swap (CDS). For a financial practitioner it is important to know, however, that agencies possess and disclose substantially more information than ratings, scoring, or expected probabilities alone. We will now discuss briefly how this information is used in the probabilistic decision framework.

Korablev and Dwyer (2007) report that for a certain group of companies the Moody's KMV EDFTM model was predicting 2.5% as the mean probability of default in 2002. The value of 1.8% was actually observed. The 10 and 90 percentiles of the distribution of predicted rates were 0.5% and 5.4%. This information is sufficient to reconstruct the parameters of the beta-distribution discussed earlier. An approximate match is $\alpha = 1.12$, $\beta = 56.35$. In Figure 4 we show the set of implied distributions for the four years preceding 2002. The inferred distribution $\pi(p_d, \alpha, \beta)$ for the year 2002 is almost identical to that for 2000.

We will now show how the inferred model of the probability of default is used in the decision-making process.

It appears from the following analysis that due to idiosyncrasies of the distribution of the

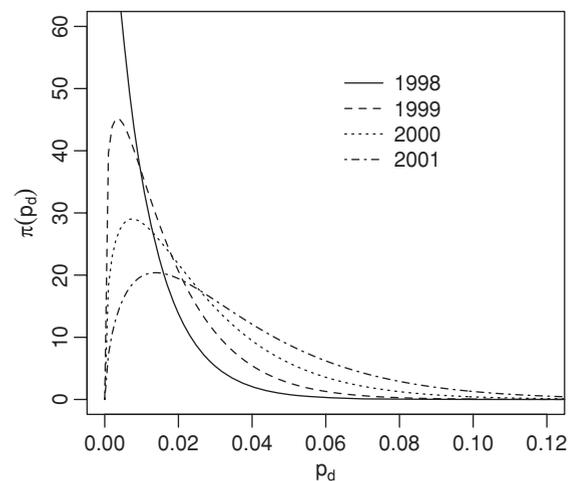


Figure 4 Implied Distribution of the Probability of Default p_d According to the Moody's Data for 1998, 1999, 2000, and 2001

probability of default, the effects of the model risk can be profound, even catastrophic. Describe the investor’s risk preferences with the following disutility function:

$$U(p_d) = -e^{\frac{p_d}{\eta}}$$

This function describes an investor who is progressively reluctant to tolerate deviations from the expected probability of default when these deviations exceed η . Note that disutility of positive deviations from the expected value is growing exponentially, while the beta-distribution of p_d falls off around its mode much slower, approximately as a power function. Using a beta-distributed probability of default $\pi(p_d, \alpha, \beta)$, we find for the certainty equivalent

$$CE(p_d) = U^{-1}(E(U(p_d))) = \eta \log \left(\Gamma(\alpha + \beta) F \left(\alpha, \alpha + \beta, \frac{1}{\eta} \right) \right)$$

where $F(a, b, z)$ is the regularized confluent hypergeometric function $F_1(a, b, z) / \Gamma(b)$ (Weinstein, 2010). The certainty equivalent of $CE(p_d)$ can be interpreted as an equivalent certain probability of default, which supplies the same value for the investor as the uncertain probability of default—given the investor’s risk preferences.

In the limit $\eta \rightarrow \infty$

$$CE(p_d) \rightarrow \frac{\alpha}{\alpha + \beta} \left(1 + \frac{1 + \alpha}{\alpha + \beta + 1} \frac{1}{\eta} \right) + 0(\eta^{-2})$$

At high tolerances $CE(p_d)$ coincides with the mean naïve probability of default. As the investor’s risk tolerance decreases, however, the certainty equivalent grows more and more rapidly. A plot of the exact certainty equivalent probability of default as the function of the model risk tolerance is shown in Figure 5. The parameters of the distribution are $\alpha = 1.45$ and $\beta = 15$, and the dashed line is the asymptote $\alpha / (\alpha + \beta)$.

The catastrophic divergence of the certainty equivalent probability occurs at the values of the tolerance that are close to the width of the distribution $\pi(p_d, \alpha, \beta)$: At the tolerance level

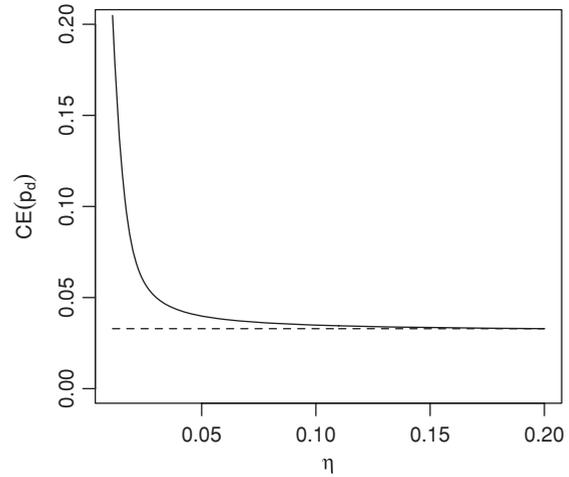


Figure 5 $CE(p_d)$ versus Risk Tolerance η
 Note: Dashed line is the asymptote value $\alpha / (\alpha + \beta)$, $\alpha = 1.45$, $\beta = 45$

$\eta = 0.01$ the $CE(p_d)$ is as big as 0.23, more than seven times the naïve value of the probability. From the practical decision-making standpoint it means that if the investor accepts the price of the bond or associated instruments defined by the naïve probability 0.031, it is likely that the investor is grossly mistaken about the value of the bond given the investor’s risk tolerance and the model risk.

KEY POINTS

- Probabilistic decision theory is a blend of the probabilistic, also called Bayesian, inference and the concept of utility.
- In the probabilistic decision theory optimal decisions maximize the expected value of the user’s utility over all possible states of the world.
- Probabilities of the states of the world are inferred from the empirical data, the model, and the user’s beliefs.
- Uncertainty in the model parameters results in the model risk; a financial model that is free of the model risk is an exception.
- Practical consequences of the model risk are evaluated using the utility function.

- Model risk significantly augments optimal allocations in equity portfolios and can result in a prospective portfolio being ruled out.
- Valuation of a risky bond is significantly affected by the model risk; ratings and expected probabilities of default alone are likely insufficient for the decision-making process.
- Failure to account for the model risk can lead to catastrophic consequences for the investor.
- Unhandled or unknown model risk produces risk exposure that remains indeterminate.

REFERENCES

- Basel Committee on Banking Supervision. (2008). *Guidelines for Computing Capital for Incremental Risk in the Trading Book*. Bank for International Settlements.
- Fama, E. F., and French, K. (1992). The cross-section of expected stock returns. *Journal of Finance* 47, 2: 427–466.
- Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press.
- Korablev, I., and Dwyer, D. (2007). *Power and Level Validation of Moody's KMV EDFTM Credit Measures in North America, Europe and Asia*. Moody's KMV, September 10.
- McKay, D.J.C. (2008). *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press.
- Merton, R. C. (1969). Lifetime portfolio selection under uncertainty. *Review of Economics and Statistics* 51: 247–257.
- Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19, 3: 425–442.
- Weisstein, E. W. (2010). Regularized hypergeometric function. *MathWorld: A Wolfram Web Resource*. <http://mathworld.wolfram.com/RegularizedHypergeometricFunction.html>

Fat-Tailed Models for Risk Estimation

BILIANA GÜNER, PhD

Assistant Professor of Statistics and Econometrics,
Özyeğin University, Turkey

IVAN MITOV, PhD

Head of Quantitative Research, FinAnalytica

BORYANA RACHEVA-YOTOVA, PhD

President, FinAnalytica

Abstract: Accounting for the likelihood of observing extreme returns and for return asymmetry is paramount in financial modeling. In addition to recognizing essential features of the returns' temporal dynamics, such as autocorrelations, volatility clustering, and long memory, a successful univariate model employs a distributional assumption flexible enough to accommodate various degrees of skewness and heavy-tailedness. At the same time, a model's usefulness depends on its scalability and practicality—the extent to which the univariate model can be extended to a multivariate one covering a large number of assets.

Risk models are employed in financial modeling to provide a measure of risk that could be employed in portfolio construction, risk management, and derivatives pricing. A *risk model* is typically a combination of a probability distribution model and a risk measure. In this entry, we discuss alternatives for building the probability distribution model, as well as the pros and cons of various *heavy-tailed* distributional choices. Our focus is univariate models; their multivariate extensions are only briefly mentioned. We start with the fundamentals—the Gaussian distribution. Then, we introduce fat-tailed alternatives, such as the *Student's t distribution* and its asymmetric version and the *Pareto*

stable class of distributions and their tempered extensions. Next, we discuss extreme value theory's risk modeling approach. We conclude with a comparative empirical example to contrast the models' performance over a 10-year period.

THE FUNDAMENTALS: NORMAL DISTRIBUTION

The use of normal (Gaussian) distribution in financial modeling has a long and distinguished tradition. The main reasons for its traditional popularity are several. First, its analytical

tractability means that deriving theoretical results and employing it in applications is (relatively) straightforward; numerical methods are widely available and implementable.¹ Second, certain central results in statistics underlie the importance of the Gaussian distribution.² Third, it has an intuitive appeal—random variables distributed with the Gaussian distribution tend to assume values around the average, with the odds of deviation from the average decreasing exponentially as one moves away from it.

Some of the most prominent financial frameworks built around the normal distribution are Markowitz's modern portfolio theory, the capital asset pricing model, and the Black-Scholes option pricing model. All of them assume (or imply) that asset returns follow a normal distribution and reflect a long-standing paradigm that rational investors' preferences can be described exclusively in terms of expected returns and risk as measured by the variance of the return distribution. However, they are inherently static frameworks. The underlying dynamic is either given exogenously or is based on the assumption that returns have independent and identical distributions. Such characteristics do not fit adequately with the empirically observed features of financial returns and investor choice.

In this section, we describe the fundamentals of a risk modeling approach based on the Gaussian distribution. We start with a review of some of its basic properties and facts.

Basics and Properties of the Gaussian Distribution

The normal distribution is characterized by two parameters—a location (mean) parameter and a scale (volatility, standard deviation) parameter.³ The location parameter serves to displace (shift) the whole distribution, while the scale parameter changes the shape of the distribution. For small values of the scale, the

distribution is narrow and peaked, while for (relatively) larger values, it widens and flattens. Since the normal distribution is symmetric around its mean, the location (mean) coincides with the center of the distribution. Commonly, the mean is denoted by μ and the standard deviation by σ .

Two important properties of the normal distribution are location-scale invariance and summation stability. They are directly related to the central role of the normal distribution in traditional financial modeling.

Location-Scale Invariance Property

Let us suppose a random variable X is normally distributed with parameters μ and σ . Now consider another random variable, Y , obtained as a linear function of X , that is, $Y = aX + b$. The variable Y is also normally distributed with parameters $\mu_Y = a\mu + b$ and $\sigma_Y = a\sigma$. That is, if a normal random variable is multiplied by a constant and/or is shifted, it remains distributed with the normal distribution.

Summation Stability Property

Let us take n independent random variables distributed with the Gaussian distribution with parameters μ_i and σ_i . The sum of the variables is normal as well. The resulting normal distribution has a mean and standard deviation obtained, respectively, as

$$\mu = \mu_1 + \mu_2 + \cdots + \mu_n$$

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2}$$

Location-scale invariance and summation stability are not universal properties of statistical distributions. In financial applications, however, they are clearly desirable properties.

The property of summation stability is often used to justify the predominant use of normal distributions in financial modeling. A statistical result, called the central limit theorem, states that, under certain technical conditions, the sum of a large number of random variables behaves

like a normally distributed random variable. More generally, we say that the normal distribution possesses a domain of attraction. In fact, the normal distribution is not the only distribution with this feature. As we will see later in the entry, it is the class of stable distributions (to which the Gaussian distribution belongs) that is unique with that property. A large sum of random variables can only converge to a stable distribution.

Density Function and Fitting of the Normal Distribution

The density function of a random variable X distributed with the normal distribution with mean μ and standard deviation σ is given by the following expression

$$f(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \quad (1)$$

We denote this distribution as $N(\mu, \sigma)$. The variable X and the parameter μ can take any real value, while σ can only take positive values. A normal random variable with zero mean and standard deviation of one is said to be distributed with the standard normal distribution ($N(0, 1)$). The presence of the exponential function in the normal density implies that the probability of events away from the mean decays at an exponential rate. In contrast, heavy-tailed distributions are characterized by power-law behaviors for large (small) values of the random variables, leading to increased chance for extreme events relative to the Gaussian setting.

Fitting of the Gaussian distribution is usually performed by maximizing the logarithm of the likelihood function given by

$$\ell(\mu, \sigma | x_1, x_2, \dots, x_n) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^2} \quad (2)$$

where x_1, x_2, \dots, x_n is the sample of observed data used for fitting. The resulting estimators of

the mean and the standard deviation are (using standard notation):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3)$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (4)$$

Unconditional models imply that returns are independent and identically distributed (IID), so that (among other implications) the returns' means and variances remain unchanged through time. However, empirical evidence abounds that financial returns exhibit time-series properties such as *autocorrelation* and *volatility clustering*, which make unconditional return modeling inadequate. The time-series properties of returns need to be modeled in a conditional framework with appropriate time series models. We consider conditional normal models next.

Conditional Normal Models and Time-Series Properties of Returns

Properly computing the risk of a portfolio depends on recognizing a number of essential features of the evolution of returns through time. We begin with the two most commonly accounted for by academics and practitioners—autocorrelation and volatility clustering.

Sometimes, a portfolio's past performance influences future performance. Current returns of a financial asset may depend on its past returns. This property of autocorrelation is modeled by including lagged (past) values of the return and/or lagged innovations (informational surprises). The resulting conditional model of the expected return is called an autoregressive moving average (ARMA) model.

Time-varying volatility concerns the empirically observed fact that large returns (of either sign) tend to be followed by large ones and small returns by small ones. To describe this volatility clustering effect, the class

of autoregressive conditional heteroskedasticity (ARCH) models, as well as their generalized (GARCH) extensions are widely used. GARCH models assume that volatility on a given day depends on the volatilities and also squared innovations of the one or more previous days.

The typical approach to building a risk model includes at least the elements of autoregressive component and volatility clustering component by means of a GARCH or alternative ARCH-type processes. A conditional normal ARMA(1,1)-GARCH(1,1) model combines the returns' conditional mean and volatility models with the assumption that returns are distributed with the Gaussian distribution. Analytically, the model is represented as

$$r_t = \mu_t + \epsilon_t \quad (5)$$

$$\mu_t = \phi_0 + \phi_1 r_{t-1} + \theta_1 \epsilon_{t-1} \quad (6)$$

$$\sigma_t^2 = \omega + \alpha \sigma_{t-1}^2 + \beta \epsilon_{t-1}^2 \quad (7)$$

where r_t , μ_t , and σ_t^2 are the return, expected return, and return variance at time t , respectively, and ϵ_t is the innovation at time t . The innovation is normally distributed with mean 0 and variance σ_t^2 .⁴

The standardized fitted residuals, $\hat{\epsilon}_t/\hat{\sigma}_t$, are the original returns with the effects of autocorrelation and volatility clustering removed (filtered out). Since the model innovations are assumed to be Gaussian, if the model is correctly specified, these filtered returns must exhibit the dynamics of a Gaussian white noise with variance one. Therefore, one easy way to determine whether the distributional assumptions are valid is to examine the properties of these residuals. Indeed, numerous studies have confirmed that in the case of financial returns the standardized fitted residuals are not Gaussian. That is, even after removing the autocorrelation and volatility clustering, fat tails, though smaller in magnitude, continue to be present in returns. Time-varying volatility then is not sufficient to explain the extreme events observ-

able in returns.⁵ Therefore, a more realistic risk model should allow for fat-tailed innovations. In the next section, we discuss parametric fat-tailed models, specifically, models based on the classical Student's t distribution and its asymmetric version, as well as on the stable and tempered stable distributions.

INCORPORATING HEAVY TAILS AND SKEWNESS: PARAMETRIC FAT-TAILED MODELS

The Student's t distribution has become the "go to" mainstream alternative of the normal distribution, when attempting to address asset returns' heavy-tailedness. Further below, we introduce an extension, called the skewed Student's t distribution, designed to deal with data asymmetries. First, we turn to discussing the "classical" Student's t distribution.

The "Classical" Student's t Distribution

The Student's t distribution (or simply the t -distribution) is symmetric and mound-shaped, like the normal distribution. However, it is more peaked around the center and has fatter tails. This makes it better suited for return modeling than the Gaussian distribution. Additionally, numerical methods for the t -distribution are widely available and easy to implement.

The t -distribution has a single parameter, called degrees of freedom (DOF), that controls the heaviness of the tails and, therefore, the likelihood for extreme returns. The DOF takes only positive values, with lower values signifying heavier tails. Values less than 2 imply infinite variance, while values less than 1 imply infinite mean. The t -distribution becomes arbitrarily close to the normal distribution as DOF increases above 30.

Density Function of the Student’s *t* Distribution

A random variable *X* (taking any real value) distributed with the Student’s *t* distribution with *v* degrees of freedom has a density function given by

$$f(x | v) = \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})\sqrt{v\pi}} \left(1 + \frac{x^2}{v}\right)^{-(v+1)/2} \tag{8}$$

where “ Γ ” is the Gamma function. We denote this distribution by t_v . The mean of *X* is zero and its variance is given by

$$\text{var}(X) = \frac{v}{v-2} \tag{9}$$

The variance exists for values of *v* greater than two and the mean—for *v* greater than one.

The *t*-distribution above is sometimes referred to as the “standard” Student’s *t* distribution.⁶ In financial applications, it is often necessary to define the Student’s *t* distribution in a more general manner so that we allow for the mean (location) and scale to be different from zero and one, respectively. The density function of such a “scaled” Student’s *t* distribution is described by

$$f(x | v, \mu, \sigma) = \frac{\Gamma(\frac{v+1}{2})}{\sigma\Gamma(\frac{v}{2})\sqrt{v\pi}} \left(1 + \frac{1}{v} \left(\frac{x - \mu}{\sigma}\right)^2\right)^{-(v+1)/2} \tag{10}$$

where the mean μ can take any real value and σ is positive. The variance of *X* is then equal to $\sigma^2 v / (v - 2)$. We denote the distribution above by $t_v(\mu, \sigma)$.⁷

Finally, we make a note of an equivalent representation of the Student’s *t* distribution which is useful for obtaining simulations from it. The $t_v(\mu, \sigma)$ distribution is equivalently expressed as a scale mixture of the normal distribution where the mixing variable distributed with the inverse-gamma distribution,

$$X \sim N(\mu, \sqrt{W}\sigma)$$

$$W \sim \text{Inv-Gamma}\left(\frac{v}{2}, \frac{v}{2}\right)$$

Later in this entry we will again come across mixture representations in the context of our discussion of the skewed Student’s *t*, the stable Paretian, and the classical tempered stable distributions.

Degrees of Freedom Across Assets and Time

The typical approach to risk modeling based on the Student’s *t* distribution includes building an autoregressive and volatility clustering components, as well as assuming that DOF is the same for all assets’ returns. This assumption is essential if we want to extend the classical Student’s *t* model to a multivariate one. It is, however, an empirical fact that assets are not homogeneous with respect to the degree of non-normality of their returns. Moreover, tail thickness and shape are not constant through time either.

Consider the result of an empirical study of constituent stocks of the S&P 500 stock index during the period from January 2, 1991 to June 30, 2011. We calibrate the Student’s *t* distribution after filtering the equity returns for GARCH effects. The estimated DOF is shown in Figure 1. It is evident that tail behavior diverges dramatically across stocks. Around 44% of equity returns are very heavy tailed, with DOF estimate below five. Around 54% of equities have an estimated DOF parameter between five and 10. Only three stocks exhibit characteristics closer to the Gaussian, with a DOF exceeding 15. Obtaining accurate DOF estimates across assets is important in risk management, since these estimates can form the basis of an analysis of portfolio risk contributors and diversifiers.

Not only does tail behavior vary across assets, it also changes through time. In relatively calm periods, asset returns are almost Gaussian, while in turbulent periods, the tails become fatter. Figure 2 illustrates the time-varying behavior of DOF, suggesting temporal tail dynamics for the Dow Jones Industrial

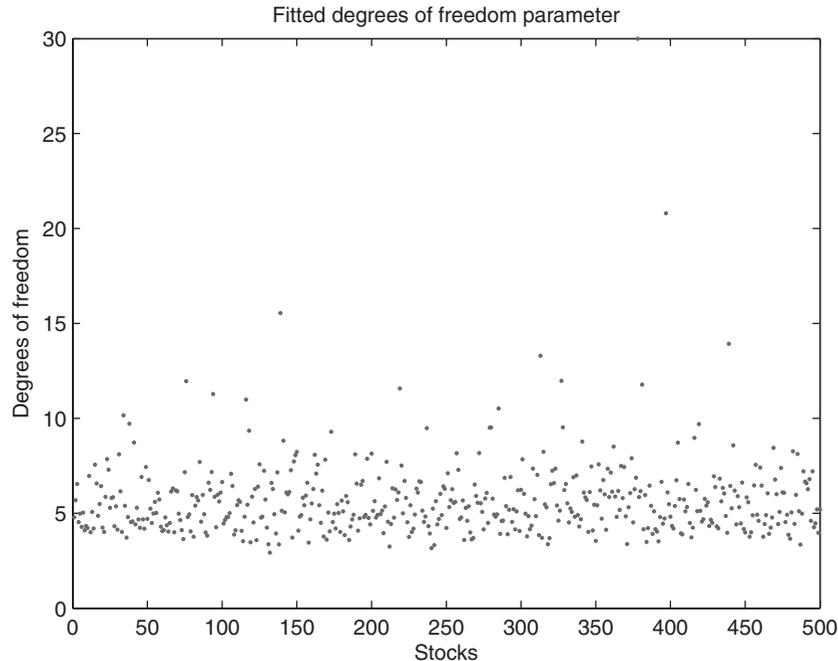


Figure 1 Fitted Degrees-of-Freedom Parameter for S&P 500 Index Stock Returns

Note: The Student's t distribution is calibrated on the residuals from a GARCH model fitted to the returns of the stocks in the S&P 500 index.

Average (DJIA) returns for the period from January 1, 1997 to June 30, 2011. The top and middle plots show the value and return of the DJIA, respectively. The bottom plot shows the DOF parameter estimates.⁸ In periods of “normal” market volatility, returns are almost normally distributed, with a fitted DOF over 30. However, when markets are unsettled, return tails grow heavier. Accounting for that time dynamics is important in risk budgeting and management to serve as an indicator for the transition between different market regimes—from calm to turbulent market or vice versa.

As pointed out earlier, a major limitation of employing the classical Student's t distribution for risk modeling is its symmetry. If there is significant asymmetry in the data, it will not be reflected in the risk estimate. There are at least several versions of the skewed Student's t distribution, depending on the analytical way in which asymmetry is introduced into the

distribution.⁹ Below, we consider the skewed Student's t distribution obtained as a mixture of Gaussian and inverse-gamma distributions.¹⁰

The Skewed Student's t Distribution

Suppose that a random variable X is distributed with the skewed Student's t distribution, obtained as a mixture of a Gaussian distribution and an inverse-gamma distribution,

$$X = \mu + \gamma W + Z\sqrt{W} \quad (11)$$

where

- W is an inverse-gamma random variable with parameters $\nu/2$, $W \sim \text{Inv-Gamma}(\nu/2, \nu/2)$.
- Z is a Gaussian random variable, $Z \sim N(0, \sigma)$, independent of W .

The parameters μ and γ are real-valued. The sign and magnitude of γ control the asymmetry in X . We say that X 's distribution is a mean-variance mixture of the normal

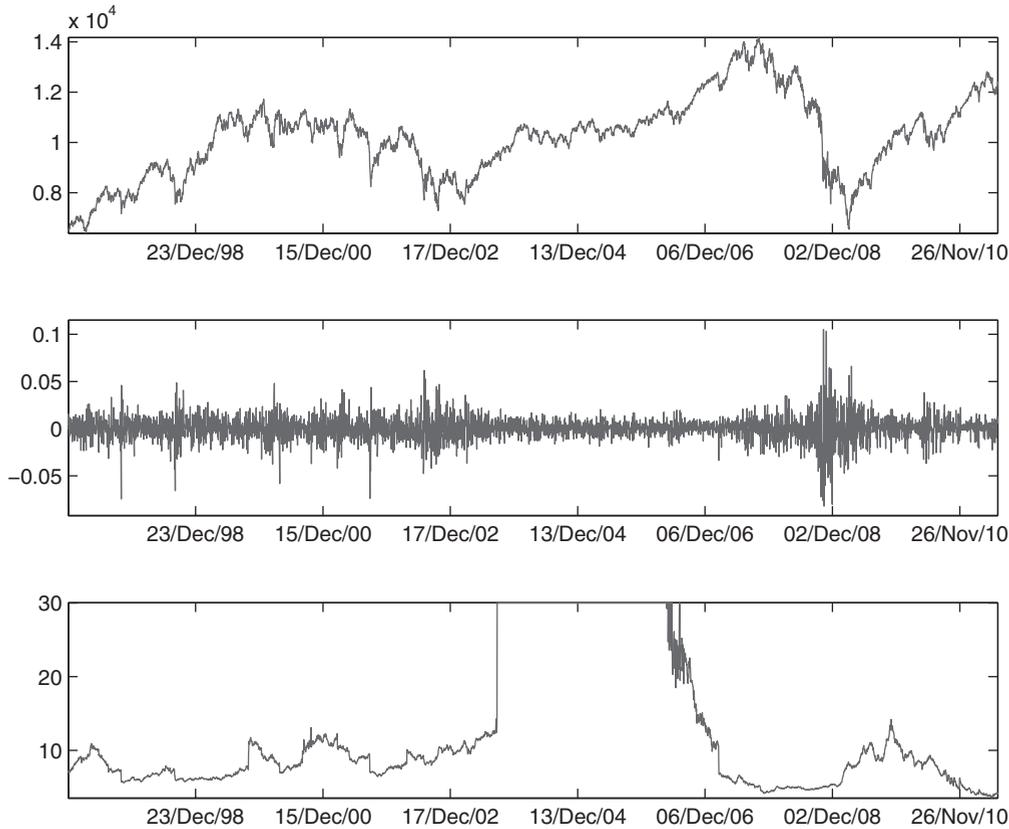


Figure 2 Fitted Degrees-of-Freedom Parameter for DJIA Returns
 Note: The Student’s *t* distribution is calibrated on the residuals from a GARCH model fitted to the return on the DJIA using a 500-day rolling window

distribution, since the mixing variable *W* modifies both the mean and the variance of the Gaussian *Z*. Notice that conditional on the value of *W*, the distribution of *X* is normal:

$$X | W = w \sim N(\mu + \gamma w, \sigma \sqrt{w}) \tag{12}$$

X’s unconditional distribution is what is defined as the skewed Student’s *t* distribution and its density is given by the expression

$$f(x | \mu, \sigma, \gamma, \nu) = A \times \frac{\exp\left(\frac{(x-\mu)\gamma}{\sigma^2}\right)}{\left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{(\nu+1)/2}} \times \frac{K_{(\nu+1)/2}(B)}{B^{-(\nu+1)/2}}$$

where

$$A = \frac{2^{1-(\nu+1)/2}}{\Gamma(\frac{\nu}{2})(\pi\nu)^{1/2}\sigma}$$

$$B = \sqrt{\left(\nu + \frac{(x-\mu)^2}{\sigma^2}\right) \frac{\gamma^2}{\sigma^2}}$$

and $K_\lambda(\cdot)$ is the so-called modified Bessel function with index λ .

Fitting and Simulation of the Classical and Skewed Student’s *t* Distributions

Estimation of the classical and skewed Student’s *t* distributions is carried out using the method of maximum likelihood. Simulations from the two distributions make use of their normal mixture representations. For given

parameters μ, σ , and γ (e.g., the maximum likelihood estimates), generation of t and skewed t observations consists of the steps below:

- Generate an observation w from the inverse-gamma distribution with parameters $\nu/2$.
- Generate an observation z from the normal distribution with mean 0 and variance σ^2 .
- Compute the corresponding observation of the t or skewed t -distribution, respectively, as

$$x = \mu + \sqrt{w}z \quad \text{and} \quad y = \mu + w\gamma + \sqrt{w}z \quad (13)$$

Stable Paretian and Classical Tempered Stable Distributions

Research on stable distributions in the field of finance has a long history.¹¹ In 1963, the mathematician Benoit Mandelbrot first used the stable distribution to model empirical distributions that have *skewness* and fat tails. The practical implementation of stable distributions to risk modeling, however, has only recently been developed. Reasons for the late penetration are the complexity of the associated algorithms for fitting and simulating stable models, as well as the multivariate extensions. To distinguish between Gaussian and non-Gaussian stable distributions, the latter are commonly referred to as stable Paretian, Lévy stable, or α -stable distributions.

Stable Paretian tails decay more slowly than the tails of the normal distribution and therefore better describe the extreme events present in the data. Like the Student's t distribution, stable Paretian distributions have a parameter responsible for the tail behavior, called tail index or index of stability.

Definition of Stable Paretian Distributions

We offer two definitions of the stable Paretian distribution. The first one establishes the stable distribution as having a domain of attraction. That is, (properly normalized) sums of IID random variables are distributed with the α -stable distribution as the number of summands n goes to infinity. Formally, let

Y_1, Y_2, \dots, Y_n be IID random variables and $\{a_n\}$ and $\{b_n\}$ be sequences of real and positive numbers, respectively. A variable X is said to have the stable Paretian distribution if

$$\frac{\sum_{i=1}^n Y_i - a_n}{b_n} \xrightarrow{d} X \quad (14)$$

where the symbol \xrightarrow{d} denotes convergence in distribution.

The density function of the stable Paretian distribution is not available in a closed-form expression in the general case. Therefore, the distribution of a stable random variable X is alternatively defined through its characteristic function. The density function can be obtained through a numerical method, as we explain further below. The characteristic function of the α -stable distribution is given by

$$\varphi_X(t) = \begin{cases} \exp\{i\mu t - \sigma^\alpha |t|^\alpha (1 - i\beta \text{sign}(t) \tan \frac{\pi\alpha}{2})\}, & \alpha \neq 1 \\ \exp\{i\mu t - \sigma |t| (1 - i\beta 2/\pi \text{sign}(t) \log(t))\}, & \alpha = 1 \end{cases} \quad (15)$$

where $\text{sign}(t)$ is 1 if $t > 0$, 0 if $t = 0$, and -1 if $t < 0$. The four parameters uniquely determining the α -stable distribution are:

- α : index of stability or tail index, $0 < \alpha \leq 2$.
- β : skewness parameter, $-1 \leq \beta \leq 1$.
- σ : scale parameter, $\sigma > 0$.
- μ : location parameter, $\mu \in \mathbb{R}$.

We denote the distribution by $S_\alpha(\sigma, \beta, \mu)$. The roles of α and β are illustrated in Figure 3. The sign of β reflects the asymmetry of the distribution. Positive (negative) β implies skewness to the right (left). As noted earlier, the index of stability controls the degree of heavy-tailedness of the distribution. Smaller values imply a fatter tail. The closer the tail index is to two, the more Gaussian-like the distribution is. Indeed, for $\alpha = 2$, we arrive at the Gaussian distribution—if X is distributed with $S_2(\sigma, \beta, \mu)$, then it has the normal distribution with mean equal to μ and variance equal to $2\sigma^2$.

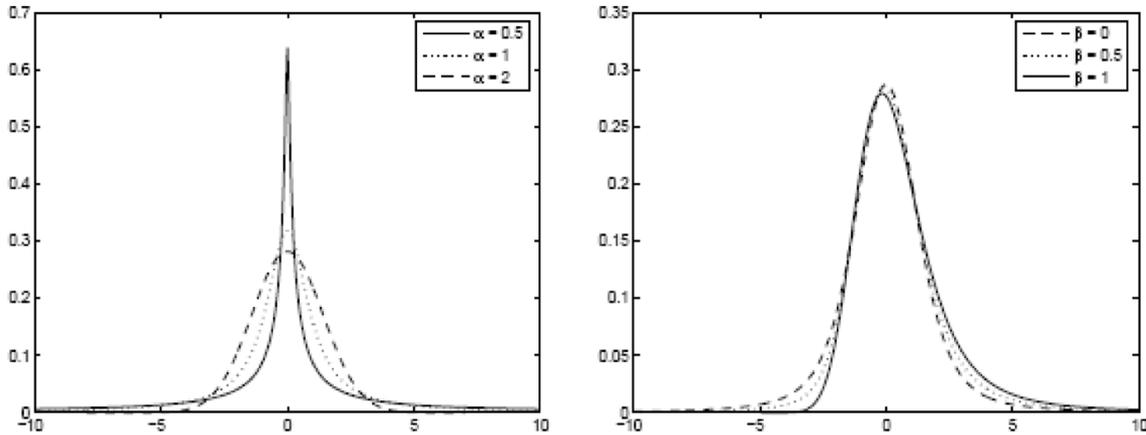


Figure 3 Stable Density: $\mu = 0, \sigma = 1, \beta = 0$, and varying α (left); $\alpha = 1.5, \mu = 0, \sigma = 1$, and varying β (right)

In this case, the parameter β loses its meaning as a skewness parameter and becomes irrelevant. Nevertheless, the normal distribution is usually associated with $\beta = 0$. Apart from the Gaussian distribution, there are two more special cases for which the density function of the stable distribution is available in a closed form: the Cauchy distribution ($\alpha = 1, \beta = 0$) and the completely skewed Lévy distribution ($\alpha = 1/2, \beta = \pm 1$).

Basic Properties of the Stable Distribution

We outline three basic properties of the α -stable distribution:

- *Power-tail decay.* The tail of the stable distribution’s density decays like a power function (slower than the exponential decay). It is this property that allows the stable distribution to capture the occurrence of extreme events. For a constant C , the property can be expressed as

$$P(|X| > x) \propto Cx^{-\alpha}, \quad \text{as } x \rightarrow \infty \quad (16)$$

- *Existence of raw moments.* The magnitude of the tail index determines the order up to which raw moments exist:

$$E|X|^p < \infty, \quad \text{for any } p: 0 < p < \alpha \quad (17)$$

$$E|X|^p = \infty, \quad \text{for any } p: p \geq \alpha$$

This property implies that, for non-Gaussian α -stable distributions ($\alpha < 2$), the variance (as well as higher moments such as skewness and kurtosis) does not exist. When the index of stability has a value less than one, the mean is infinite as well. Since the variance does not exist, one cannot express risk in terms of the variance. However, the scale parameter can play the role of a risk measure, in the same way that the standard deviation does in the normal distribution case.

- *Stability.* The property of stability characterizes the preservation of the distributional form under linear transformations. It is governed by the index of stability α and expressed as follows. Suppose that X_1, X_2, \dots, X_n are IID random variables, independent copies of a random variable X . Then, for a positive constant C_n and a real number D_n , X follows the stable distribution:

$$X_1 + X_2 + \dots + X_n \stackrel{d}{=} C_n X + D_n \quad (18)$$

The notation $\stackrel{d}{=}$ denotes equality in distribution. The constant $C_n = n^{1/\alpha}$ determines the stability property. The stability property means that the “classical” central limit theorem does not apply in the non-Gaussian case. A large sum of appropriately standardized IID random variables is distributed with the

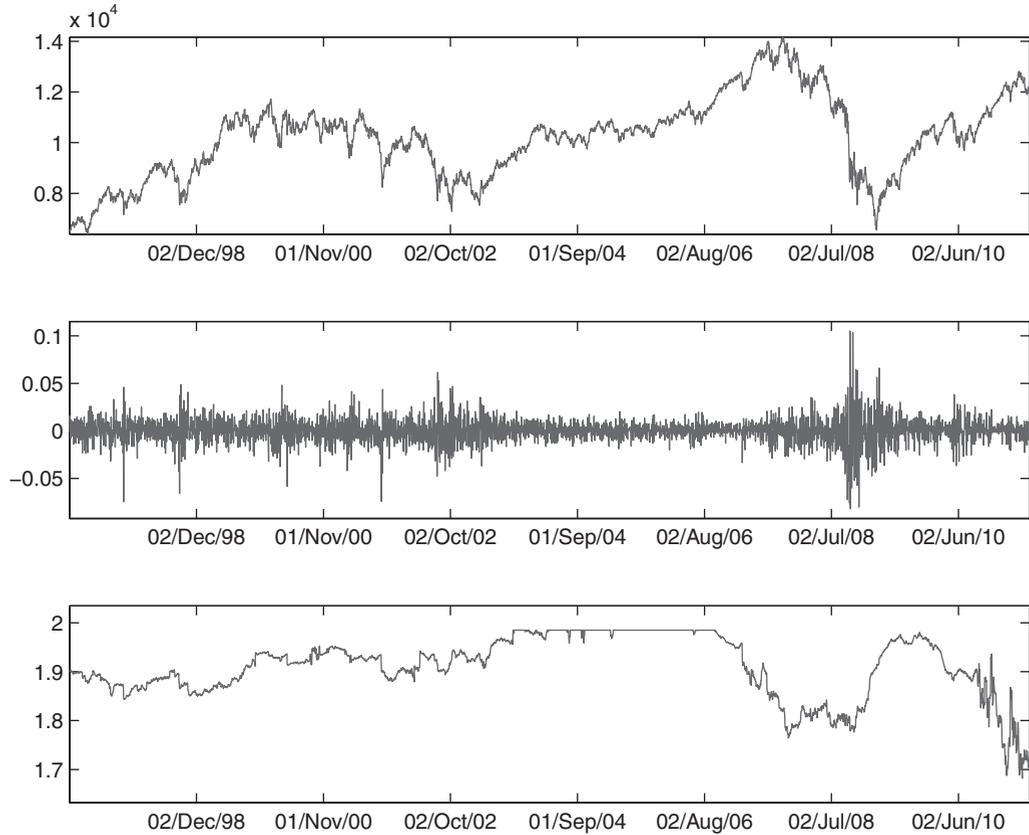


Figure 4 Fitted Stable Tail Index for DJIA Returns
 Note: The Stable Paretian distribution is calibrated on the residuals from a GARCH model fitted to the return on the DJIA using a 500-day rolling window

stable Paretian distribution as the number of terms increases indefinitely, not with the normal distribution.

When the variables $X_i, i = 1, \dots, n$, are themselves distributed with the α -stable distribution, $X_i \sim S_\alpha(\sigma_i, \beta_i, \mu_i)$, the stability property can be extended further:

1. The distribution of $Y = \sum_{i=1}^n X_i$ is α -stable with index of stability α and parameters:

$$\beta = \frac{\sum_{i=1}^n \beta_i \sigma_i^\alpha}{\sum_{i=1}^n \sigma_i^\alpha},$$

$$\sigma = \left(\sum_{i=1}^n \sigma_i^\alpha \right)^{1/\alpha}, \quad \mu = \sum_{i=1}^n \mu_i \quad (19)$$

2. The distribution of $Y = X_1 + a$ for some real constant a is α -stable with index of

stability α and parameters:

$$\beta = \beta_1, \quad \sigma = \sigma_1, \quad \mu = \mu + a \quad (20)$$

3. The distribution of $Y = a X_1$ for some real constant $a (a \neq 0)$ is α -stable with index of stability α and parameters:

$$\beta = \text{sign}(a)\beta_1, \quad \sigma = |a|\sigma_1,$$

$$\mu = \begin{cases} a\mu_1, & \text{for } a \neq 1 \\ a\mu_1 - \frac{2}{\pi}a \ln(a)\sigma_1\beta_1, & \text{for } a = 1 \end{cases}$$

In empirical analysis, the time-varying tail behavior of assets is reflected in the nonconstancy of the tail index of the α -stable distribution, as demonstrated in Figure 4. As in the earlier illustration, the tail index is estimated by fitting a stable distribution to the filtered returns

(after removing the volatility clustering with a GARCH model.) The tail index of the DJIA returns is very close to two in the upward market environment from 2003 to 2005 but starts decreasing right before the 2008 market crash and is smallest at the time of the crash itself. This implies that tail thickness is smallest in the bullish market from 2003 to 2005 and is largest during the crisis period.

As noted above, the variance of non-Gaussian stable distributions does not exist. To address this potentially undesirable feature, smoothly truncated stable distributions and various types of tempered stable distributions have been proposed. They are all obtained with a procedure known as “tempering” applied to the tails of the distribution to ensure that the variance is finite. This procedure replaces the power decay very far out in the tails of the distribution with an exponential (or faster) decay. We discuss the classical tempered stable distributions next.

Definition of Classical Tempered Stable Distributions

The characteristic function of the classical tempered stable (CTS) distribution is given by the following expression:

$$\varphi_X(t) = \exp\{imt - itC\Gamma(1 - \alpha)(\lambda_+^{\alpha-1} - \lambda_-^{\alpha-1}) + C\Gamma(-\alpha)((\lambda_+ - it)^\alpha - \lambda_+^\alpha + (\lambda_- + it)^\alpha - \lambda_-^\alpha)\} \tag{21}$$

We denote the distribution by CTS($\alpha, C, \lambda_+, \lambda_-, m$). The distribution parameters are characterized as follows:

- α : tail index, $\alpha \in (0, 1) \cup (1, 2)$.
- m : location parameter, $m \in \mathbb{R}$.
- C : scale parameter, $C > 0$.
- λ_+ and λ_- : parameters controlling the decay in the right and left tail, respectively; $\lambda_+, \lambda_- > 0$.

The relative magnitudes of λ_+ and λ_- determine the degree of skewness of the CTS distribution. When $\lambda_+ > \lambda_-$ ($\lambda_+ < \lambda_-$), the distribution is skewed to the left (right). Symmetry is obtained for $\lambda_+ = \lambda_-$. Tail heaviness is

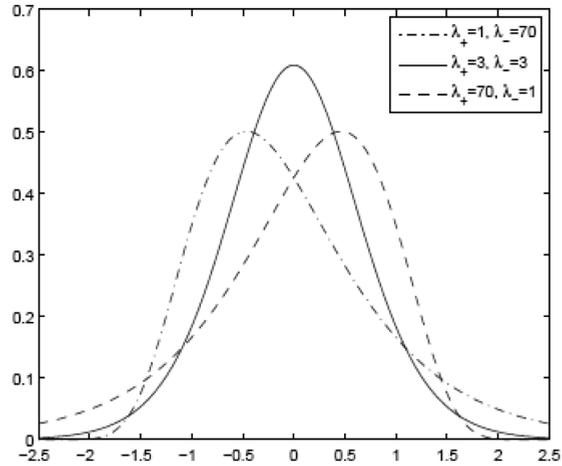


Figure 5 Probability Density of the CTS Distribution: Dependence on λ_+ and λ_-
 Note: CTS Parameter Values: $\alpha = 0.8, C = 1, m = 0$, and varying λ_+ and λ_-

determined in a more flexible manner in the CTS distribution than in the stable Paretian distribution. Three parameters play a role in that: λ_+, λ_- , and α . The former two have the effect of scaling the tails (smaller values correspond to heavier tails), while the latter one, of shaping the tails (as before, small values imply fatter tails). The effect of different values of these three parameters on the CTS distribution can be seen in Figures 5, 6, and 7.

Linear combinations of CTS-distributed random variables are also distributed with the CTS distribution. First, we define the standard CTS distribution. A random variable X has the standard CTS distribution if

$$C = (\Gamma(2 - \alpha) (\lambda_+^{\alpha-2} + \lambda_-^{\alpha-2}))^{-1} \tag{22}$$

The distribution is denoted by $X \sim \text{stdCTS}(\alpha, \lambda_-, \lambda_+)$. Its mean and variance are zero and one, respectively.

For a positive number σ and a real number m , the linear combination $Y = \sigma X + m$ has the CTS distribution:

$$Y \sim \text{CTS} \left(\alpha, \frac{\sigma^\alpha}{\Gamma(2 - \alpha) (\lambda_+^{\alpha-2} + \lambda_-^{\alpha-2})}, \frac{\lambda_+}{\sigma}, \frac{\lambda_-}{\sigma}, m \right) \tag{23}$$

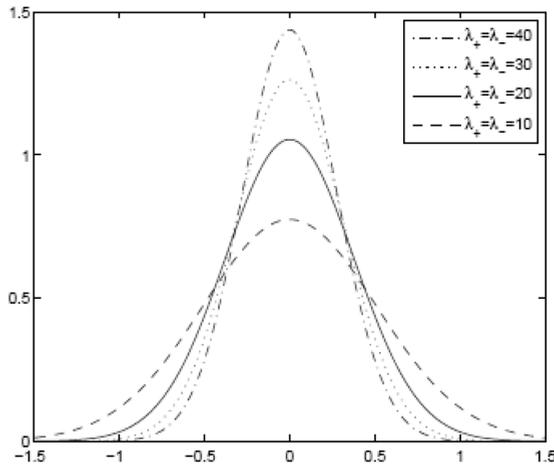


Figure 6 Probability Density of the Symmetric CTS Distribution: Dependence on λ_+ and λ_-
 Note: CTS Parameter Values: $\alpha = 1.1$, $C = 1$, $m = 0$, and varying λ_+ and λ_-

The mean and variance of Y are m and σ^2 , respectively.

Subordinated Representation of the α -Stable and CTS Distributions

Similar to the Student's t distribution, stable distributions can be represented as mixtures of other distributions. More generally,

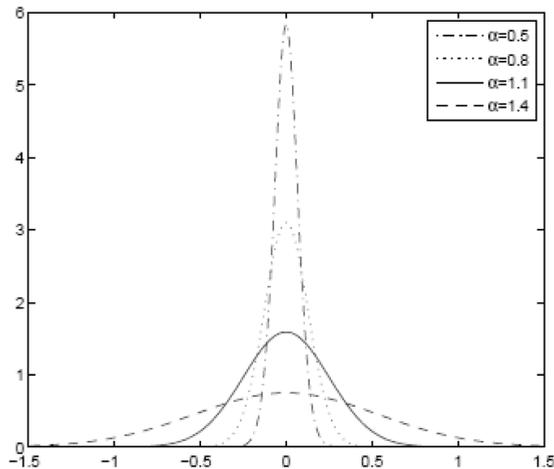


Figure 7 Probability Density of the CTS Distribution: Dependence on α
 Note: CTS Parameter Values: $C = 1$, $\lambda_+ = 50$, $\lambda_- = 50$, $m = 0$, and varying α

(continuous) mixture representations are analyzed within the framework of intrinsic time change and subordination. The price and return dynamics can be considered under two different time scales—the physical (calendar) time and an intrinsic (also called operational, trading or market) time. The intrinsic time is best thought of as the cumulative trading volume process which measures the cumulative trading volume of the transactions up to a point on the calendar-time scale. It is a measure of market activity and a reflection of the empirical observation that price changes are larger when market activity is more intense. Let us denote the intrinsic time process by $T(t)$ and the time-evolving random variable such as price or return by $X(t)$. $X(t)$ is assumed independent of $T(t)$. The compound process $X(T(t))$ is said to be subordinated to X by the intrinsic time $T(t)$ and $T(t)$ is referred to as a subordinator.¹² Since the increments of the intrinsic time $\Delta T(t) = T(t) - T(t - \Delta t)$ are non-decreasing and positive, distributions such as gamma, Poisson, and inverse-Gaussian can be used to describe them in probabilistic terms.¹³ Another distributional alternative is the completely skewed to the right α -stable distribution, $S_\alpha(\sigma, 1, 0)$, for $0 < \alpha < 1$, whose support is the positive real line. Therefore, when $0 < \alpha < 2$, the subordinator is a stable distribution given by $S_{\frac{\alpha}{2}}(\sigma, 1, 0)$.

Subordinated models with random intrinsic time, such as $X(T(t))$, are leptokurtic. They have heavier tails and higher peaks around the mode of zero than the normal distribution. As such, they provide a natural way to model the tail effects observed in prices and returns.

Subordinated representations' usefulness is in allowing for practical ways of simulating random numbers from the corresponding models. Subordinated processes are especially important in multiasset settings, where each marginal distribution has a different tail heaviness. This across-asset heterogeneity can be modeled by having subordinators with different parameters for each asset. As noted earlier, this characteristic of multivariate asset returns is crucially

important for a realistic risk model able to identify tail risk contributors and tail risk diversifiers.

The subordinated representation of the α -stable distribution can be expressed in the following way. Let Z be a standard normal random variable, $Z \sim N(0, 1)$, and Y be a positive $\alpha/2$ -stable random variable independent of Z , $Y \sim S_{\alpha/2}(s, 1, 0)$, where

$$s = \frac{\sigma^2}{2} \cos\left(\frac{\pi\alpha}{4}\right)^{2/\alpha} \tag{24}$$

Then, the variable

$$X = Y^{1/2}Z$$

is symmetric α -stable: $X \sim S_\alpha(\sigma, 0, 0)$. This implies that every symmetric stable variable is conditionally Gaussian (conditional on the value of the stable subordinator). Unconditionally, the symmetric α -stable distribution is expressed as a scale mixture of normal distributions.¹⁴

The CTS distribution has a subordinated representation as well and can be expressed as a mean-scale mixture of Gaussian distributions. For details, see Madan and Yor (2005).

Fitting and Simulations of α -Stable and CTS Distributions

Fitting techniques for the α -stable distributions can be divided into three categories: quantile methods, characteristic function-based methods, and maximum likelihood methods. The quantile method is similar to the method of moments estimation method in that it uses predetermined empirical quantiles to estimate the sample parameters.¹⁵ The characteristic function-based methods also resemble the method of moments and rely on transformations of the sample characteristic function.¹⁶ Finally, the latter method involves maximization of the likelihood function, which is computed numerically. Comparative studies of the three fitting approaches show that the method of maximum likelihood is superior in terms of

estimation accuracy. The quantile method requires the least amount of computational time but is the least accurate one. The second category of methods also have the benefit of computational simplicity but may have a difficulty in estimating the skewness parameter.¹⁷ For these reasons, here we focus on the method of maximum likelihood in some more detail.

In statistical theory, the relationship between the probability density function (pdf) and the characteristic function is expressed as follows:

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itx) \varphi_X(t) dt \tag{25}$$

where $h > 0$ and $f(\cdot)$ and $\varphi(\cdot)$ are the density and characteristic functions, respectively. The pdf of the α -stable and CTS distributions can be computed by numerical evaluation of the integral above. A fast and computationally efficient method of numerical integration is the fast Fourier transform (FFT) algorithm.¹⁸ Consider the pdf computation in (25). The main idea of FFT is to evaluate the integral for a grid of equally-spaced values of the random variable X :

$$x_k = \left(k - 1 - \frac{N}{2}\right)h, \quad k = 1, \dots, N \tag{26}$$

That is, equation (25) can be expressed as

$$f_X(x_k) = \int_{-\infty}^{\infty} \exp\left(-i2\pi\omega\left(k - 1 - \frac{N}{2}\right)h\right) \varphi_X(2\pi\omega)d\omega$$

This integral can be approximated by the so-called Riemann sum, after choosing small enough lower and large enough upper bounds:

$$f_X(x_k) \approx s \sum_{n=1}^N \varphi\left(2\pi s\left(n - 1 - \frac{N}{2}\right)\right) \times \exp\left\{-i2\pi\left(k - 1 - \frac{N}{2}\right)\left(n - 1 - \frac{N}{2}\right)sh\right\} \tag{27}$$

for $k = 1, \dots, N$. Here, the lower and upper bounds equal $-\frac{sN}{2}$ and $\frac{sN}{2}$, respectively. The distance between the grid points $n - 1 - \frac{N}{2}$,

$n = 1, \dots, N$ is s . If $s = \frac{1}{hN}$, we arrive at the following expression for the density, after some algebraic rearrangement:

$$\begin{aligned}
 f_X(x_k) &\approx \frac{(-1)^{k-1+\frac{N}{2}}}{hN} \sum_{n=1}^N (-1)^{n-1} \varphi \\
 &\times \left(\frac{2\pi}{hN} \left(n-1 - \frac{N}{2} \right) \right) \\
 &\times \exp \left(-\frac{i2\pi(n-1)(k-1)}{N} \right), \\
 k &= 1, \dots, N \tag{28}
 \end{aligned}$$

To compute the sum above, one can use the FFT implemented by many numerical analysis software packages. The parameters of the FFT method are N , the number of summands in the Riemann sum, and h , the grid spacing. Their values can be chosen appropriately, so as to achieve a balance between approximation accuracy and computational burden.¹⁹ Finally, the maximum-likelihood estimates of the parameters of the α -stable and CTS distributions are obtained by numerical maximization of the log-likelihood function.

Simulations of α -stable distribution can be accomplished using an algorithm called the Chambers-Mallows-Stuck generator. To generate a random number from $S_\alpha(\sigma, \beta, \mu)$, the steps are as follows:

- Generate two independent random numbers from an exponential distribution with mean 1, $E \sim \exp(1)$, and from a uniform distribution, $U \sim U(-\frac{\pi}{2}, \frac{\pi}{2})$.
- If $\alpha \neq 1$, compute

$$\begin{aligned}
 Z &= s_{\alpha,\beta} \frac{\sin(\alpha(U + b_{\alpha,\beta}))}{(\cos U)^{1/\alpha}} \\
 &\times \left(\frac{\cos(U - \alpha(U + b_{\alpha,\beta}))}{E} \right)^{(1-\alpha)/\alpha} \tag{29}
 \end{aligned}$$

where

$$\begin{aligned}
 s_{\alpha,\beta} &= \left[1 + \beta^2 \tan^2 \frac{\pi\alpha}{2} \right]^{\frac{1}{2\alpha}} \\
 b_{\alpha,\beta} &= \frac{\arctan(\beta \tan \frac{\pi\alpha}{2})}{\alpha} \tag{30}
 \end{aligned}$$

- If $\alpha = 1$, compute

$$\begin{aligned}
 Z &= \frac{2}{\pi} \left[\left(\frac{\pi}{2} + \beta U \right) \tan U \right. \\
 &\quad \left. - \beta \log \left(\frac{E \cos U}{\frac{\pi}{2} + \beta U} \right) \right] \tag{31}
 \end{aligned}$$

- The random variable Z has a standardized stable distribution with location parameter equal to zero and scale parameter equal to one, $Z \sim S_\alpha(1, \beta, 0)$. To obtain an observation from $S_\alpha(\sigma, \beta, \mu)$ with arbitrary values of σ and μ , transform Z according to²⁰

$$S = \sigma Z + \mu \tag{32}$$

Conditional Parametric Fat-Tailed Models

A fat-tailed parametric model includes the following main components:

- An autoregressive component to capture autoregressive behavior.
- A volatility clustering component, usually a GARCH-type model.
- A fat-tailed distribution (stable Paretian or skewed Student's t) to explain the heavy tails and the skewness of the residuals from the ARMA-GARCH model.
- Tail thickness changing with time and across assets addressed.

INCORPORATING HEAVY TAILS AND SKEWNESS: SEMI-PARAMETRIC FAT-TAILED MODELS

In this section, we review semi-parametric models, which combine an empirical distribution for the body of the data distribution where plenty of observations are available and extend the tail by a parametric model based on extreme value theory (EVT). EVT has a long history of applications in modeling the occurrence of severe weather, earthquakes, and other extreme natural phenomena. In general terms, extreme value distributions are the asymptotic

distributions for the normalized largest observations of IID random variables. There are two main categories of models for extreme values: block maxima models and threshold exceedances models.

In the financial applications context, block maxima could refer to the maximal observations within certain predefined periods of time. For example, daily return data could be divided into quarterly (or semiannual or yearly) blocks and the largest daily observations within these blocks collected and analyzed. The distribution of the maximal values is generally not known. However, when the block size is large, so that block maxima are independent (irrespective of whether the underlying data are dependent), the limit distribution is given by EVT.²¹ The number of blocks determines the size of the data sample available for analysis and fitting. In contrast, in threshold exceedances models, the sample size is not predetermined but, naturally, depends on the a priori selected threshold level.

The first model category is represented by the so-called generalized extreme value (GEV) distribution. Its distribution function has the form

$$F_X(x | \xi, \mu, \sigma) = \begin{cases} \exp\left(-\left(1 + \xi \frac{x-\mu}{\sigma}\right)^{-1/\xi}\right), & \xi \neq 0 \\ \exp(-e^{-x}), & \xi = 0 \end{cases} \quad (33)$$

where $1 + \xi(x - \mu)/\sigma > 0$. The parameters $\xi \in \mathbb{R}$, $\mu \in \mathbb{R}$, and $\sigma > 0$ are the shape, location, and scale parameters, respectively. The value of ξ determines the three distributions encompassed by the parametric form above: the Weibull distribution ($\xi < 0$), the Gumbel distribution ($\xi = 0$), and the Fréchet distribution ($\xi > 0$). Of the three, the latter one has the heaviest tails,²² while the first one is short-tailed, with a finite right endpoint and, thus, not favored in modeling financial losses.²³

The block maxima method’s major drawback is its “wastefulness” of data: all but the largest observation within each block are discarded. For this reason, a more common approach

to EVT modeling is the threshold exceedance method. In it, the extreme events exceeding a predetermined high level (that is, events in the tail) are modeled with the generalized Pareto distribution (GPD). Its distribution function is given by

$$F_X(x | \xi, \sigma) = \begin{cases} 1 - \left(1 + \xi \frac{x}{\sigma}\right)^{-1/\xi}, & \xi \neq 0 \\ 1 - \exp\left(-\frac{x}{\sigma}\right), & \xi = 0 \end{cases} \quad (34)$$

where $\sigma > 0$ and $x \geq 0$ when $\xi \geq 0$ and $0 < x < -\sigma/\xi$ when $\xi < 0$. The parameters ξ and σ are the shape and scale parameters, respectively. Like the GEV, the GPD contains several special cases defined by the value of ξ . When $\xi > 0$, we get the Pareto distribution with parameters $\alpha = 1/\xi$ and $k = \sigma/\xi$, whose tails exhibit slow, power-law decay. The exponential distribution is obtained for $\xi = 0$; its tails decay at an exponential rate. A short (finite)-tailed distribution, called Pareto type II distribution, arises when $\xi < 0$.²⁴

Fitting and Simulations of the GPD

In empirical modeling, there is generally a perceived trade-off between fitting the bulk and the tails of the data. Data around the mode are numerous and relatively easy to fit, while data in the tails are sparse and present an estimation challenge. Most commonly, the choice of model is based on how well it fits the bulk of the data, with the tails relegated to a somewhat secondary role. The semiparametric approach we consider in this section is to describe the majority of the data in a nonparametric fashion and use the GPD to fit the tails. Since the GPD describes the excess distribution over a threshold, we now define formally this concept.

For a random variable with cumulative distribution function G , the excess distribution over the threshold u is denoted by G_u and is given by

$$G_u(x) = P(X - u \leq x | X > u) = \frac{G(x + u) - G(u)}{1 - G(u)}$$

for $0 \leq x \leq x_F - u$, where x_F is the right endpoint (a finite number or infinity) of X 's

distribution function G .²⁵ A statistical result known as the Pickand-Balkema-de Haan theorem implies that the excess distributions of a large class of underlying distributions converge to a GPD as the threshold level increases. That is, GPD is the limiting distribution as u increases to infinity.

Denote the available data sample by X_1, \dots, X_N and define an upper and a lower threshold level u_U and u_L , respectively. The data points beyond the threshold levels constitute the tails of the data distribution that are to be modeled with EVT. Naturally, separate modeling of the two tails has the purpose of accounting for the potential skewness in the data distribution. Let us define the exceedances of u_U by $Y_{k,U} = X_k - u_U$, where $X_k > u_U$ and the exceedances of u_L by $Y_{k,L} = u_L - X_k$, where $X_k < u_L$, $k = 1, \dots, K$.²⁶ The estimates of the scale and shape parameters are most conveniently obtained by maximizing the GPD log-likelihood function for each of the sets of data $Y_{k,U}$ and $Y_{k,L}$.²⁷ It is written as

$$\begin{aligned} \ln L(\xi, \sigma | Y_1, \dots, Y_K) &= \sum_{k=1}^K \ln f_Y(\xi, \sigma) \\ &= -K \ln \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{k=1}^K \ln \left(1 + \xi \frac{Y_k}{\sigma}\right) \end{aligned} \quad (35)$$

where $f(\xi, \sigma)$ denotes the GPD density function.

The empirical distribution is usually estimated using kernel density estimation approach. The kernel density estimate can be roughly thought of as a smoothed-out histogram. A parameter, called bandwidth or window width, controls the degree of smoothness of the resulting density estimate. More formally, the kernel density estimate is defined as

$$f(x, x_i, h) = \frac{1}{hn} \sum_{i=1}^n K_h \left(\frac{x - x_i}{h} \right) \quad (36)$$

where $x_i = (x_1, x_2, \dots, x_n)$ is data sample coming from some unspecified distribution and assumed to be IID. The bandwidth, h , takes pos-

itive values and K is called the kernel, a symmetric function that integrates to one. The normal density is often chosen as the kernel in (36). The bandwidth's value can be selected in an optimal way.²⁸

The approach to scenario generation from a model based on GPD is also semiparametric—the body of the distribution is simulated from the empirical density and GPD tails are attached to it. Generating observations from a GPD with a given shape parameter ξ , a scale parameter σ , and a threshold level u can be accomplished in the following three steps:

- Generate an observation U from a uniform distribution on the interval $(0, 1)$.
- Compute the quantity

$$Z = \frac{U^{-\xi} - 1}{\xi} \quad (37)$$

- Compute the GPD realization as

$$Y = \xi + \sigma \times Z \quad (38)$$

Scenarios from the body of the distribution are generated nonparametrically, via historical simulation known as bootstrapping (or resampling, more generally). The procedure involves drawing randomly, with replacement, from the set of historically observed data points. The simulated tails of the distribution are then “attached” to the scenarios from the body to obtain semiparametric scenarios from the whole data distribution.

Threshold Selection

We consider two of the most popular tools for selection of the threshold level—the mean excess function plot and the Hill plot. Both of them rely on visual inspection to determine the threshold.

The mean excess function is closely related to the concept of excess distribution. It describes the average exceedance above a threshold u , as a function of u .²⁹ Formally, it is defined as

$$m(u) = E(X - u | X > u) \quad (39)$$

In the case of the GPD, $m(u)$ can be shown to equal

$$m(u) = \frac{\sigma}{1 - \xi} + \frac{\xi}{1 - \xi} u$$

where $0 \leq u < \infty$ if $0 \leq \xi < 1$ and $0 \leq u \leq -\beta/\xi$ if $\xi < 0$. The excess mean function does not exist for $\xi \geq 1$. The mean excess function is linear in the threshold level. This linearity is used to motivate a graphical check that the data conform to a GPD model: If the plot is approximately linear for high threshold values, the GPD may be employed to describe the distribution of the exceedances. The level above which linearity is evident may be taken as the threshold level.

Plots of the Hill estimator are another EVT model selection method. The Hill approach offers a way to estimate the tail index $\alpha = 1/\xi$. Denote the i th order statistics of the data sample by $X_{(i)}$.³⁰ The Hill estimator of α is defined as

$$H_{m,n} = \left(\frac{1}{m} \sum_{i=1}^m \ln X_{(i)} - \ln X_{(m)} \right)^{-1} \quad (40)$$

where $2 \leq m \leq n$ and m is a sufficiently high number. For $\xi > 0$, the Hill estimator is equal to α asymptotically, as the sample size n and the number of extremes m increase without bound. In practical applications, the Hill estimator is computed for different values of m and plotted against these values. The plot is expected to stabilize above a certain value of m , so that the Hill estimates constructed from a different number of order statistics remain approximately the same. The threshold level u is then estimated by $X_{(m)}$.

The semiparametric approach described in this section is a source of two major challenges. First, in order to obtain a sufficiently large number of observations in the tail, a large sample of historical data is needed. Second, even though the plots of the Hill estimator and the mean excess function provide a method for threshold identification, such identification is intrinsically subjective, as it is based on visual inspection.

Moreover, it is difficult to automate it for large-scale applications.³¹

Conditional GPD Approach

The semiparametric approach described above is unconditional, since it implicitly assumes that the observed data is IID. A typical conditional GPD approach involves the components:

- Autoregressive model to capture linear dependencies in the data.
- GARCH-type model to capture the volatility clustering in the data.
- Semi-parametric model applied to the standardized residuals (which are approximately IID) to explain the data's heavy-tailedness and asymmetry.

COMPARISON AMONG RISK MODELS

Using the DJIA daily returns from February 7, 1992 to June 30, 2011, we conduct a back-testing analysis to compare the three fat-tailed distribution models—stable Paretian, Student's t , and EVT—alongside the normal distribution model. The data used in all models are first filtered for autoregression and volatility clustering using ARMA-GARCH.

The particular models we use in this section are the univariate analogs of the typical approaches to modeling in the multivariate case. A short discussion will help clarify what this means. Earlier we explained that, in a multi-asset setting, taking into account the varying tail behavior of the returns of different assets is of principal importance for risk analysis and management. However, employing the classical Student's t distribution in the multivariate case necessarily implies the same value of the DOF parameter for all assets. That value would "average out" the tail-fatness of assets, so that the risk of some risk drivers will be underestimated, while the risk of others, overestimated. To reflect this typical multivariate application, in our backtesting analysis we

choose to fix the DOF of the Student's t distribution to four.

In the case of the stable Paretian model, similar considerations about the heterogeneity of tail behavior across risk drivers lead us to use the subordinated representation of the α -stable distribution. As mentioned in an earlier section, that representation allows for modeling the individual tail behavior of assets.

The backtesting analysis in this section, therefore, can be understood as a comparison among models with increasing degree of sophistication. We start with the classical parametric approach (the "Gaussian model"). Then, a "non-sophisticated" fat-tailed model, represented by the fixed-DOF Student's t model (the "T-model") is tested. Finally, a state-of-the-art fat-tailed model—the stable subordinated model (the "stable model")—is considered. For each of the four models, exceedances of *value-at-risk* (VaR)—the number of times the realized loss is larger than the predicted VaR level—are tracked.³² We run the backtest with the following settings:

- Backtest period: January 2, 2004, to June 30, 2011.
- VaR confidence level: 99%.
- Time window: 500 rolling days for normal, classical Student's t , and stable Paretian distributions and 3,000 rolling days for EVT.³³
- EVT threshold: 1.02% (as suggested by Goldberg, Miller, and Weinstein (2008)).

The number of exceedances of the daily 99% VaR in the backtesting analysis for the four models is summarized below:

Model	Number of Exceedances
Stable	21
Student's t	26
Gaussian (normal)	42
EVT	1

The number of exceedances is compared using a 95% confidence interval estimated to be [10, 27]. The results show that the Gaussian model is too optimistic—with 42 exceedances,

its VaR forecasts are too low. In contrast, the EVT approach is overly pessimistic: Its predicted VaR is only exceeded once in the backtesting period. The Student's t model and the stable model both produce exceedances within the confidence interval, with the latter model being very close to the upper bound.

The DJIA performance during the backtest period is presented in Figure 8. Figure 9 plots the daily 99% VaR forecast produced by the Gaussian model, the Student's t model, and the stable model against the daily DJIA returns for the full backtest period. Since the EVT model's VaR predictions are too conservative, we have excluded it from the exhibit for the sake of presentation clarity. It can be seen from the figure that in times of low market volatility, the VaR forecasts of the three models are almost indistinguishable. However, during periods of greater market turmoil, differences in predicted risk levels are substantial across models. This point is further elaborated in Figure 10, which shows the spreads between the 99% VaR forecasts for the Student's t Gaussian and the stable-Gaussian model pairs, along with the values and returns of DJIA. We observe that the stable-Gaussian VaR spread stays at zero for the period from 2004 to late-2006, suggesting "normal" market conditions. (The estimated tail index parameter of the α -stable distribution is close or equal to two during that period.) This is an essential feature of the stable model: Despite being a fat-tailed approach, it does not overpenalize the portfolio by assessing unnecessarily high risk estimates during calm market periods. On the other hand, even in times of severe market circumstances the number of exceedances of the stable VaR is within an acceptable range. For the period from June 1, 2008 to June 1, 2009, the stable VaR has one exceedance, which is within the 95% confidence interval for the number of exceedances ([0, 4]). By comparison, the Student's t model's VaR is exceeded four times, while the Gaussian model has seven exceedances.

It is interesting to analyze whether the VaR forecasts can anticipate the transition from a



Figure 8 Dow Jones Industrial Average Performance: January 2, 2004–June 30, 2011

calm market regime to a turbulent one. To investigate, we “zoom in” on the VaR spread dynamics for the two-year period leading up to the September 2008 crash. Figure 11 shows

the VaR spreads for the period September 1, 2006–September 1, 2008, relative to the Gaussian VaR forecast. We notice that the stable-Gaussian relative spread starts increasing in

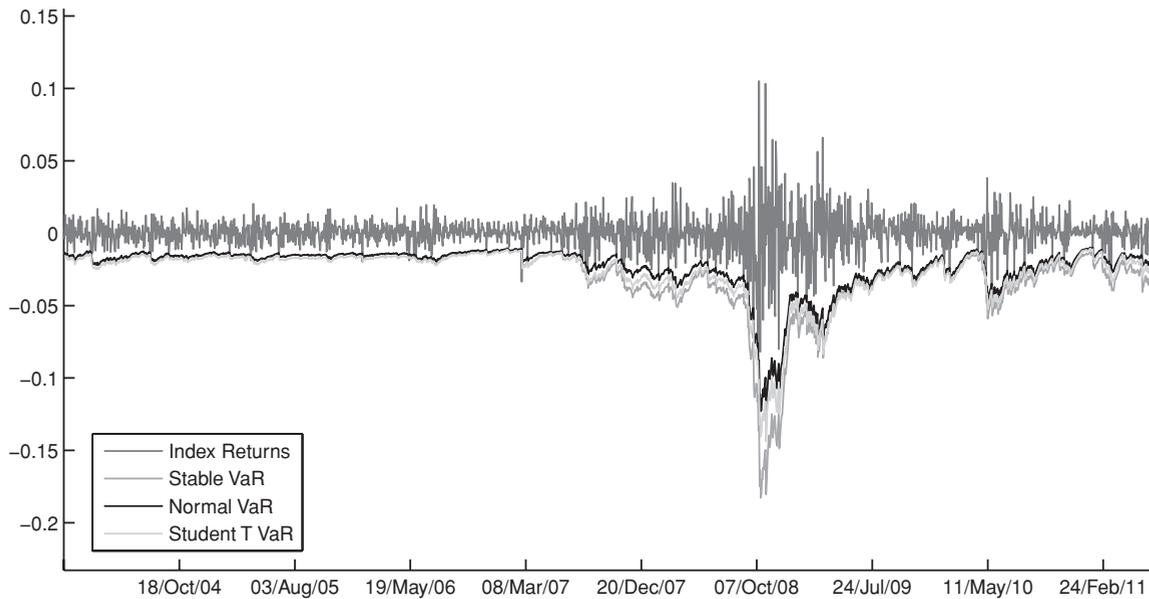


Figure 9 Backtest of the 99% Daily VaR Predicted by Different Distributional Methodologies

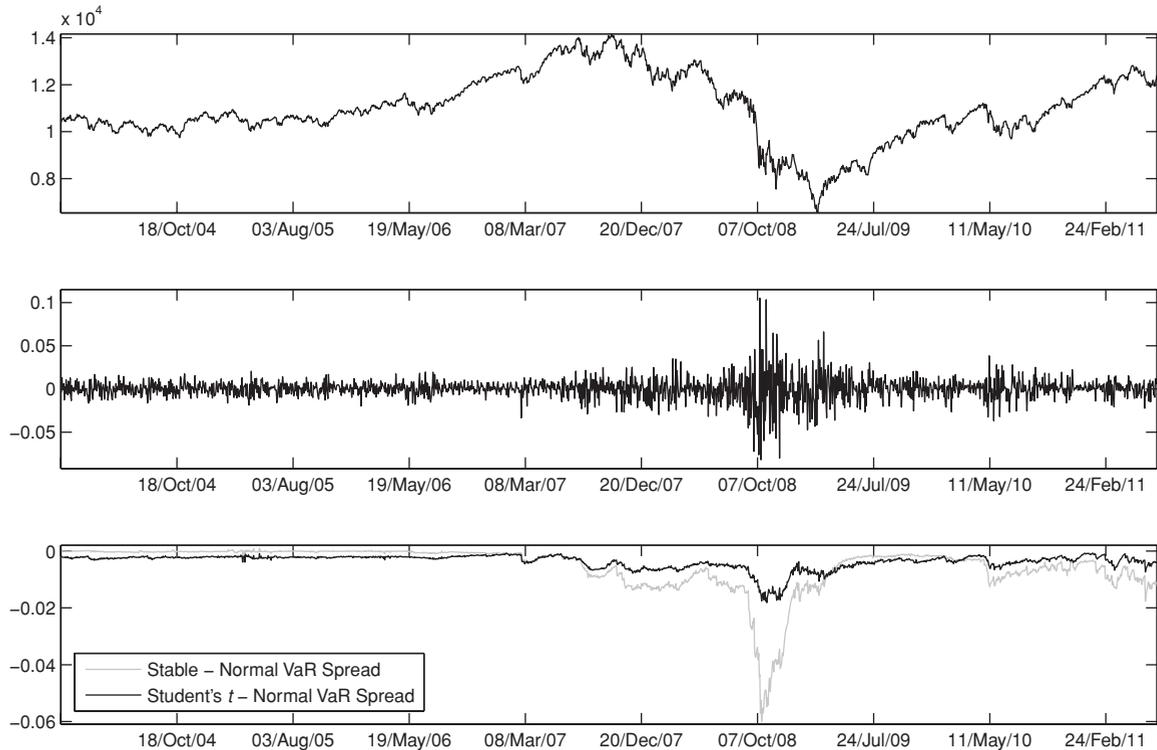


Figure 10 Spreads between 99% VaR Predictions for Student's t Gaussian and Stable-Gaussian Model Pairs: Full Period

late 2006. This is the result of the increased tail-fatness estimated by the stable model (α decreases). At the same time, the Student's t -Gaussian relative spread is fairly constant due to the fact that the DOF (and, therefore, the tail-fatness) is fixed.³⁴ Over the two-year period, we can see a pronounced increase in the stable-Gaussian VaR relative spread. There are two time segments (in spring 2007 and spring 2008) in which the spread actually decreases. Both are associated with periods following major negative news and market tremors.³⁵ In these periods, the Gaussian model's VaR "catches up" post factum due to the increase in the estimates of the conditional GARCH volatility.

In general, one can interpret the upward trend of the stable-Gaussian VaR relative spread as an indicator of markets accumulating higher probability of extreme events before the actual market volatility goes up. This predictive behavior is only possible due to the time-varying

estimates of the tail-fatness (the α parameter in the stable model). Thus, in the fixed-DOF Student's t model, such a predictive trend cannot be observed. During the two-year period, the number of exceedances is eight for the stable model, ten for the Student's t model, and 16 for the Gaussian model, while the 95% confidence interval is $[0, 9]$.

Finally, to test the significance of the stable-Gaussian VaR relative spread, we build a confidence interval for it. We do that by altering the tail index α at each point in time during the backtesting period with plus and minus one standard deviation of α and then re-computing the stable VaR and the associated stable-Gaussian relative spread. The standard deviation of α is estimated using parametric bootstrap, based on 200 bootstrap samples of 500 random draws each generated from an α -stable model with the corresponding α .³⁶ Figure 12 shows the confidence bounds of the

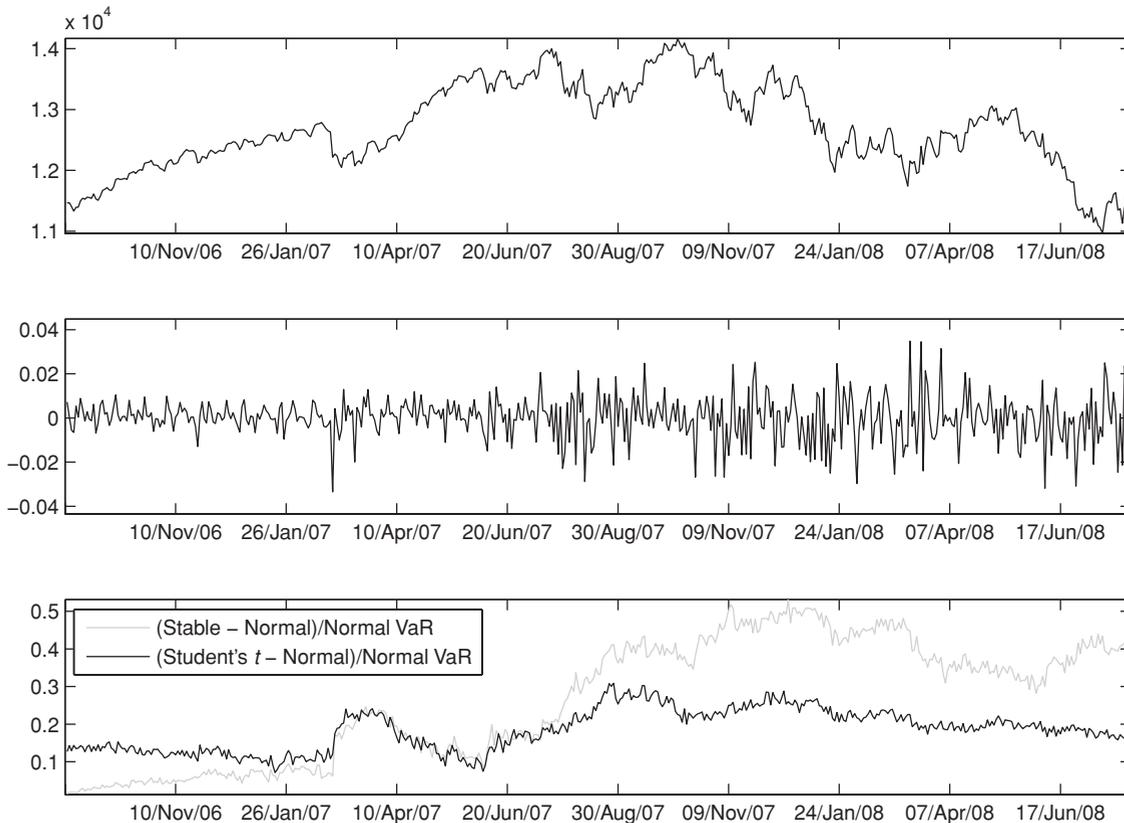


Figure 11 Relative Spreads between 99% VaR Predictions for Student's t Gaussian and Stable-Gaussian Model Pairs: September 1, 2006–September 1, 2008

stable-Gaussian VaR relative spread for the two-year period running up to September 2008, together with the Student's t -Gaussian VaR relative spread. Even the lower bound of the confidence interval of the stable-Gaussian VaR relative spread is more indicative than the Student's t -Gaussian relative spread over this time period. Although the upward trend of the lower confidence bound is not as strong as that of the upper confidence bound, the results support the conclusion that the stable model's VaR forecasts have the ability to anticipate a switch from a calm to a volatile market regime.

KEY POINTS

- The Gaussian distribution is not adequate to describe the empirical features of asset returns. The standardized residuals from a conditional Gaussian model exhibit heavy-tailedness and asymmetry.
- The Student's t distribution has fatter tails than the normal distribution. To account for skewness, however, the "classical" Student's t distribution needs to be modified.
- The skewed Student's t distribution can be represented as a mean-scale mixture of normal distributions; that is, normal distribution with random mean and variance.
- The tails of the stable Paretian distributions decay more slowly than the tails of the normal distribution and therefore better describe the extreme events present in the data.
- In the non-Gaussian case, a large sum of appropriately standardized IID random variables is distributed with the stable Paretian distribution in the limit.

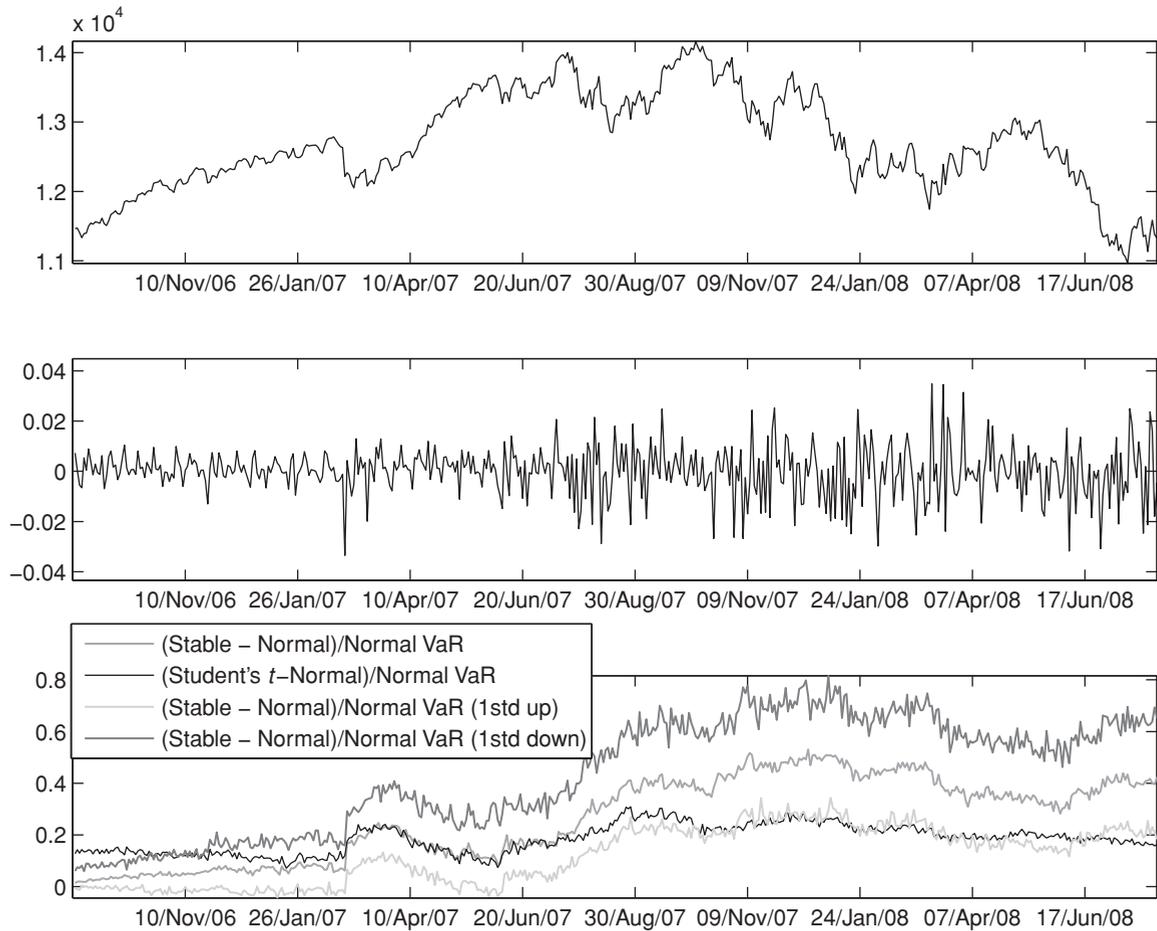


Figure 12 Relative Spreads between 99% VaR Predictions for Student's t Gaussian and Stable-Gaussian Model Pairs with Confidence Bounds: September 1, 2006–September 1, 2008

- To address the issue of infinite variance, the stable Paretian distribution may be modified by tempering of the distribution's tail. This gives rise to the tempered stable distributions.
- There are two main categories of distributions for extreme values—block maxima models and threshold exceedances models. The latter category is more often employed in risk modeling, since it is less “wasteful” of historical data than the former category.
- Selection of the threshold from where the tail of the data distribution starts is based on a subjective judgement and, together with data scarcity, is the main bottleneck in EVT applications.
- In all cases, before applying a fat-tailed model, an ARMA-GARCH filter should be

used to remove the temporal dependence in asset returns.

- A realistic distributional assumption for a model should allow for tail-fatness that changes over time and from asset to asset. Such models can serve as early warning indicators when moving to a new market regime (from calm to turbulent and vice versa) and can identify tail-risk contributors and tail-risk diversifiers.

NOTES

1. The Gaussian distribution's analytical tractability in the multivariate setting is an additional important factor behind its widespread use. See, for example, Kotz,

- Johnson, and Balakrishnan (2000) for details on the multivariate normal distribution.
2. However, later in this entry we provide an important clarification regarding a statistical result, called the central limit theorem.
 3. In general, the scale parameter does not always coincide with the standard deviation (volatility), as we will see, for instance, in our discussion of the scaled Student's t distribution later in the entry.
 4. See, for example, Rachev, Stoyanov, Biglova, and Fabozzi (2005).
 5. The conditional distribution of returns according to the model above is Gaussian. The unconditional distribution, however, is not normal but a mixture of normal distributions (due to the time-varying mean and variance). Its tails are fatter than those of the normal distribution but not fat enough to account for the empirically observed heavy tails.
 6. Note that this is not the same as "standardized," since the standard deviation of X is not one.
 7. Notice that the Student's t distribution defined in equation (8) has a location of zero and a scale of one.
 8. More precisely, we estimate a GARCH model on a 500-day rolling window of returns and then fit a t -distribution to the (standardized) GARCH residuals.
 9. Skewed Student's t models have been proposed by Fernandez and Steel (1998), Azzalini and Capitanio (2003), and Rachev and Rüschenendorf (1994), among others.
 10. The skewed Student's t distribution belongs to a more general class of distributions called generalized hyperbolic distributions and introduced by Barndorff-Nielsen (1978). It contains the Student's t and normal distributions as limiting cases.
 11. See Rachev and Mittnik (2000) and Samorodnitsky and Taqqu (1994). A detailed description of the stable methodology is available in Rachev, Martin, Racheva-Yotova, and Stoyanov (2009).
 12. For details on the statistical properties of subordinated processes, see Feller (1966) and Clark (1973). Rachev and Mittnik (2000) provide discussions of subordinated processes in financial applications.
 13. More generally, the family of infinitely divisible distributions to which the gamma, Poisson, inverse-Gaussian, and all stable Paretian distributions belong is a natural choice of distributions for the increments of the intrinsic time process $T(t)$.
 14. In a multivariate setting, Z would be distributed with a multivariate normal distribution and Y can be a vector whose components are stable subordinators with different tail-fatness. The resulting distribution is a generalization of the multivariate sub-Gaussian stable distribution.
 15. McCulloch (1986)'s estimation procedure generalized the quantile method for symmetric α -stable distributions of Fama and Roll (1971).
 16. See Press (1972). Kogon and Williams (1998) and Koutrouvelis (1980) suggested regression-type estimator algorithms, also based on the characteristic function.
 17. Comparison among the three types of estimation categories is provided in Stoyanov and Racheva-Yotova (2004).
 18. A detailed description of the stable fitting methodology is available in Rachev and Mittnik (2000).
 19. Rachev and Mittnik (2000) show that selecting $h = 0.01$ and $N = 2^{13}$ reduces the approximation error in computing the α -stable pdf to the satisfactory level of 10^{-6} .
 20. The algorithm for simulations from the CTS distribution is rather involved and described in detail in Rachev, Kim, Bianchi, and Fabozzi (2011).
 21. The role of EVT in modeling maxima of random variables is similar to the one the central limit theorem plays in modeling the sums of random variables. Both characterize the limiting distributions.

22. The tails of the Fréchet distribution decay like a power function at a rate $\alpha = 1/\xi$, the tail index parameter characterizing the α -stable distribution.
23. Comprehensive discussion of EVT is available in McNeil, Frey, and Embrechts (2005).
24. For more details on GPD, see Embrechts, Klüppelberg, and Mikosch (1997).
25. Notice that the threshold level u is in fact the location parameter in the GPD distribution.
26. The thresholds u_L and u_U are usually defined in terms of symmetric empirical quantiles, for example, the 5% and 95% quantiles. In that case, the number of observed data points in each tail is equal to K .
27. To be precise, when fitting the left tail, (35) is maximized over the absolute values of $Y_{k,L}$, $k = 1, \dots, K$.
28. See, for example, Silverman (1986).
29. $m(u)$ is also known as mean residual life function in survival analysis and characterizes the expected residual lifetime of a component that has function for u units of time already.
30. The i th order statistic is the i th largest observation in a data sample. The first order statistic is the maximum of the sample, while the n th order statistic is the minimum of a sample of size n .
31. See Rachev, Racheva-Yotova, and Stoyanov (2010) for a detailed discussion of these challenges.
32. Value-at-risk is defined as the minimum loss at a given confidence level for a pre-defined time horizon.
33. Goldberg, Miller, and Weinstein (2008) use time windows ranging from approximately 1,500 to 7,600 days.
34. The small variations are due to the variability in the estimates of the Student's t GARCH model.
35. In February 2007, Freddie Mac announced that it would no longer buy the most risky subprime mortgages and mortgage-related securities and in April 2007, New Century Financial Corporation, a leading subprime mortgage lender, filed for Chapter 11 bankruptcy protection. Then, in the spring of 2008, we observed the collapse of Bear Stearns and the associated events.
36. The bootstrap sample size is 500, since this is the length of the time window used to calibrate the model.

REFERENCES

- Azzalini, A., and Capitanio, A. (2003). Distributions generated by a perturbation of symmetry with emphasis on multivariate skew t -distribution. *Journal of the Royal Statistical Society, Series B* 65, 2: 367–389.
- Barndorff-Nielsen, O. (1978). Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of Statistics* 5: 151–157.
- Clark, P. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica* 41, 1: 135–155.
- Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modeling Extremal Events for Insurance and Finance*. Springer.
- Fama, E., and Roll, R. (1971). Parameter estimates for symmetric stable distributions. *Journal of the American Statistical Association* 66: 331–338.
- Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*. Hoboken, NJ: Wiley.
- Fernandez, C., and Steel, M. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association* 93, 441: 359–371.
- Goldberg, L., Miller, G., and Weinstein, J. (2008). Beyond value-at-risk: Forecasting portfolio loss at multiple horizons. *Journal of Investment Management* 6, 2: 73–98.
- Hull, P., and Welsh, A. (1985). Adaptive estimates of parameters of regular variation. *Annals of Statistics* 12, 1: 331–341.
- Kogon, S., and Williams, D. (1998). Characteristic function based estimation of stable parameters. In R. Adler, R. Feldman, and M. Taqqu (eds.), *A Practical Guide to Heavy Tails*, Birkhauser.
- Kotz, S., Johnson, N. L., and Balakrishnan, N. (2000). *Continuous Multivariate Distributions: Models and Applications*. Hoboken, NJ: Wiley & Sons.
- Koutrouvelis, I. (1980). Regression-type estimation of the parameters of stable laws. *Journal of the American Statistical Association* 7: 918–928.
- Madan, D., and Yor, M. (2006). CGMY and Meixner Subordinators are Absolutely

- Continuous with Respect to One Sided Stable Subordinators. Cornell University Library, arXiv:math/0601173v2 [math.PR].
- Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business* 26: 394–419.
- McCulloch, J. (1986). Simple consistent estimators of stable distribution parameters. *Communications in Statistics. Simulation and Computation* 15: 1109–1136.
- McNeil, A., Frey, R., and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools*. Princeton: Princeton University Press.
- Press, S. (1972). Estimation in univariate and multivariate stable distribution. *Journal of the American Statistical Association* 67: 842–846.
- Rachev, S., Kim, Y., Bianchi, M., and Fabozzi, F. (2011). *Financial Models with Levy Processes and Volatility Clustering*. Hoboken, NJ: Wiley.
- Rachev, S., Martin, D., Racheva-Yotova, B., and Stoyanov, S. (2009). Stable ETL, optimal portfolios and extreme risk management, in G. Bol (ed.), *Risk Assessment: Decisions in Banking and Finance*, Physica, Springer.
- Rachev, S., and Mittnik, S. (2000). *Stable Paretian Models in Finance*. Hoboken, NJ: Wiley & Sons.
- Rachev, S., Racheva-Yotova, B., and Stoyanov, S. (2010). Capturing fat tails. *Risk Magazine*, May: 72–76.
- Rachev, S., and Rüschenendorf, L. (1994). On the Cox, Ross, and Rubinstein model for option pricing. *Theory of Probability and its Applications* 39: 150–190.
- Rachev, S., Stoyanov, S., Biglova, A., and Fabozzi, F. (2005). An empirical examination of daily stock return distributions for U.S. stocks. In D. Baier, R. Decker, L. Schmidt-Thieme (eds.), *Data Analysis and Decision Support*, Springer.
- Samorodnitsky, G., and Taqqu, M. (1994). *Stable Non-Gaussian Random Processes, Stochastic Models with Infinite Variance*. London: Chapman & Hall.
- Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.
- Stoyanov, S., and Racheva-Yotova, B. (2004). Univariate stable laws in the field of finance—Parameter estimation. *Journal of Concrete and Applicable Mathematics* 2, 4: 24–49.

Index

- Absence of arbitrage principle, *I:99, I:127*. *See also* arbitrage, absence of
- ABS/MBS (asset-backed securities/mortgage-backed securities), *I:258–259, I:267*
- cash flow of, *III:4*
- comparisons to Treasury securities, *III:5*
- modeling for, *III:536*
- Accounting, *II:532, II:542–543*
- Accounting firms, watchdog function of, *II:542*
- Accounts receivable turnover ratio, *II:557–558*
- Active-passive decomposition model, *III:17, III:19–22, III:26*
- Activity ratios, *II:557–558, II:563*
- Adapted mesh, one year to maturity, *II:680f*
- Adjustable rate mortgages (ARMs). *See* ARMs (adjustable rate mortgages)
- Adjustments for changes in net working capital (ANWC), *II:25*
- Adverse selection, *III:76*
- Affine models, *III:554–557*
- Affine process, basic, *I:318–319, I:334n*
- Agency ratings, and model risk, *II:728–729*
- Airline stocks, *II:249–250, II:250f, II:250t, II:252t*
- Akaike Information Criterion (AIC), *II:703, II:717*
- Algorithmic trading, *II:117*
- Algorithms, *II:676–677, II:701–702, III:124*
- Allied Products Corp., cash flow of, *II:576*
- α -stable densities, *III:243f, III:244f*
- α -stable distributions
- defined, *II:738*
- discussion of, *III:233–238*
- fitting techniques for, *II:743–744*
- properties of, *II:739*
- simulations for, *II:750*
- subordinated representation of, *II:742–743*
- usefulness of, *III:242*
- and VaR, *II:748*
- variables with, *II:740*
- α -stable process, *III:499*
- Alternative risk measures proposed, *III:356–357*
- Amazon.com
- cash flows of, *II:568, II:568t*
- American International Group (AIG), stock prices of, *III:238*
- Amortization, *II:611, III:72–73*
- Analysis
- and Barra model, *II:244–248*
- bias in, *II:109*
- common-size, *II:561–563*
- crisis-scenario, *III:379–380*
- to determine integration, *II:514*
- formulas for quality, *II:239*
- fundamental, *II:243, II:248, II:253–254*
- interpretation of results, *III:42–44*
- mathematical, *I:18*
- model-generated, *III:41–42*
- multivariate, *II:48*
- statistical, *I:140, II:353–354*
- sum-of-the-parts, *II:43–44*
- vertical *vs.* horizontal common-size, *II:562*
- Analytics, aggregate, *II:269t*
- Anderson, Philip W., *III:275*
- Annual percentage rate (APR), *II:598, II:615–616*
- Annual standard deviation, *vs.* volatility, *III:534*
- Annuities
- balances in deferred, *II:610f*
- from bonds, *I:211–212*
- cash flows in, *II:604–607*
- future value factor, *II:605–606*
- ordinary, *II:605*
- present value factor, *II:605, II:606–607*
- valuation of due, *II:608–609*
- valuing deferred, *II:609–611*
- Anticipation, in stochastic integrals, *III:475*
- Approximation, quality of, *II:330–331*
- APT (arbitrage pricing theory), *I:116*
- Arbitrage
- absence of, *I:56, I:135, II:473*
- in continuous time, *I:121–123*
- convertible bond, *I:230*
- costless profits, *I:442*
- costless trades, *I:428t*
- defined, *I:99, I:119, I:123*
- in discrete-time, continuous state, *I:116–119*
- and equivalent martingale measures, *I:111–112*
- in multiperiod finite-state setting, *I:104–114*
- in one-period setting, *I:100–104*
- pricing of, *I:124, I:134–135, II:476*
- profit from, *I:221–222*
- and relative valuation models, *I:260*
- and state pricing, *I:55–56, I:102, I:130*
- test for costless profit, *I:441*
- trading strategy with, *I:105*
- types of, *I:55–56*
- using, *I:70–71*
- Arbitrage-free, *III:577, III:593–594*
- Arbitrage opportunities, *I:55, I:56, I:100, I:117, I:260–261, I:437*
- Arbitrage pricing theory (APT), *I:116*
- application of, *I:60–61*
- development of, *II:468, II:475–476*
- factors in, *II:138*
- key points on, *II:149–150*
- and portfolio optimization, *I:40*

- ARCH (autoregressive conditional heteroskedastic) models and behavior of errors, *II:362*
 defined, *I:176*
 in forecasting, *II:363*
 reasons for, *III:351*
 type of, *II:131*
 use of, *II:733–734*
- ARCH/GARCH models
 application to VaR, *II:365–366*
 behavior of, *II:361–362*
 discussion of, *II:362–366*
 generalizations of, *II:367–373*
 usefulness of, *II:366–367*
- ARCH/GARCH processes, *III:277*
- Area, approximation of, *II:589–590*, *II:589f*
- ARIMA (autoregressive integrated moving average) process, *II:509–510*
- ARMA (autoregressive moving average) models
 defined, *II:519*
 and Hankel matrices, *II:512*
 linearity of, *II:402*
 and Markov coefficients, *II:512*
 multivariate, *II:510–511*, *II:513–514*
 nonuniqueness of, *II:511*
 representations of, *II:508–512*
 and time properties, *II:733*
 univariate, *II:508–510*
- ARMA (autoregressive moving average) processes, *III:276–277*
- ARMs (adjustable rate mortgages), *III:25*, *III:71–72*, *III:72f*, *III:74*
- Arrays, in MATLAB and VBA, *III:420–421*, *III:457–458*, *III:466*
- Arrow, Kenneth, *II:467*, *II:699*
- Arrow-Debreu price, *I:53–55*. *See also* state price
- Arrow-Debreu securities, *I:458*, *I:463*
- Arthur, Bryan, *II:699*
- Artificial intelligence, *II:715*
- Asian fixed calls, with finite difference methods, *II:670t*
- Asian options, pricing, *III:642–643*
- Asset allocation
 advanced, *I:36*
 building blocks for, *I:38*
 modeling of, *I:42*
 standard approach to, *I:37–38*
- Asset-backed securities (ABS), *I:258*
- Asset-liability management (ALM), *II:303–304*, *III:125–126*
- Asset management, focus of, *I:35*
- Asset prices
 codependence of, *I:92*
 multiplicative model for, *I:86–87*, *I:88*
 negative, *I:84*, *I:88*
 statistical inference of models, *I:560*
- Asset pricing, *I:3*, *I:56–59*, *I:59–60*, *I:65–66*, *II:197*
- Asset return distributions, skewness of, *III:242*
- Asset returns
 characteristics of, *III:392*
 errors in estimation of, *III:140–141*
 generation of correlated, *I:380–381*
 log-normal distribution applied to, *III:223–225*
 models of, *III:381*
 normal distribution of, *I:40*
 real-world, *III:257*
 simulated vector, *I:380–381*
- Assets
 allocation of, *I:10*
 on the balance sheet, *II:533–534*
 carry costs, *I:424–425*
 correlation of company, *I:411*
 current *vs.* noncurrent, *II:533*
 deliverable, *I:483*
 discrete flows of, *I:425–426*
 expressing volatilities of, *III:396–397*
 financing of, *II:548*
 funding cost of, *I:531*
 future value of, *I:426t*, *I:427t*
 highly correlated, *I:192*
 intangible, *II:534*
 liquid, *II:551*
 management of, *II:558*
 market prices of, *I:486*
 new fixed, *II:25*
 prices of, *I:60*
 redundant, *I:51*
 representation of, *II:515*
 risk-free, *I:112–113*
 risky *vs.* risk-free, *I:5–6*
 shipping, *I:555*
 storage of physical, *I:439*, *I:442–443*, *I:560–561*
 values of after default events, *I:350*
- Asset swaps, *I:227–230*
- Assumptions
 about noise, *II:126*
 under CAPM, *I:68–69*
 errors in, *III:399*
 evaluation of, *II:696*
 homoskedasticity *vs.* heteroskedasticity, *II:360*
 importance of, *III:62*
 for linear models, *II:310–311*
 for linear regression models, *II:313*
 in scenario analysis, *II:289*
 simplification of, *III:397*
 using inefficient portfolio analysis, *I:288t*
 violations of, *I:475*
 zero mean return, *III:397*
- Attribution analysis, *II:188–189*
- AT&T stock, binomial experiment, *I:146–148*
- Audits, of financial statements, *II:532*
- Augmented Dickey-Fuller test (ADF), *II:387*, *II:389*, *II:390t*, *II:514*
- Autocorrelation, *II:328–329*, *II:503*, *II:733*
- Autoregressive conditional duration (ACD) model, *II:370*
- Autoregressive conditional heteroskedastic (ARCH) models. *See* ARCH (autoregressive conditional heteroskedastic) models
- Autoregressive integrated moving average (ARIMA) process, *II:509–510*
- Autoregressive models, *II:360–362*
- Autoregressive moving average (ARMA) models. *See* ARMA (autoregressive moving average) models
- AVaR. *See* average value at risk (AVaR)
- Average credit sales per day, calculation of, *II:553*
- Average daily volume (ADV), *II:63*
- Averages, equally weighted, *III:397–409*
- Average value at risk (AVaR) measure
 advantages of, *III:347*
 back-testing of, *III:338–340*
 boxplot of fluctuation of, *III:338f*
 and coherent risk measures, *III:333–334*
 computation of in practice, *III:336–338*
 computing for return distributions, *III:334–335*
 defined, *III:331–335*
 estimation from sample, *III:335–336*
 and ETL, *III:345–347*
 geometrically, *III:333f*
 graph of, *III:347f*
 higher-order, *III:342–343*
 historical method for, *III:336–337*
 hybrid method for, *III:337*
 minimization formula for, *III:343–344*
 Monte Carlo method for, *III:337–338*
 with the multivariate normal assumption, *III:336*
 of order one, *III:342–343*
 for stable distributions, *III:344–345*
 tail probability of, *III:332–333*
- Axiomatic systems, *III:152–153*

- Bachelier, Louis, *II:121–122, II:467, II:469–470, III:241–242, III:495*
- Back propagation (BP), *II:420*
- Back-testing
- binomial (Kupiec) approach, *III:363*
 - conditional testing (Christoffersen), *III:364–365*
 - diagnostic, *III:367–368*
 - example of, *II:748–751*
 - exceedance-based statistical approaches, *III:362–365*
 - in-sample *vs.* out-sample, *II:235–236*
 - need for, *III:361–362*
 - statistical, *III:362*
 - strengths/weaknesses of
 - exceedance-based, *III:365*
 - tests of independence, *III:363–364*
 - trading strategies, *II:236–237*
 - use of, *III:370*
 - using normal approximations, *III:363*
 - of VaRs, *III:365–367*
- Backward induction pricing technique, *III:26*
- Bailouts, *I:417*
- Balance sheets
- common-size, *II:562, II:562f*
 - information in, *II:533–536*
 - sample, *II:534t, II:546t*
 - structure of, *II:536*
 - XYZ, Inc. (example), *II:29t*
- Balls, drawing from urn, *III:174–177, III:175f, III:179–180*
- Bandwidth, *II:413–414, II:746*
- Bank accounts, and volatility, *III:472*
- Bank for International Settlements (BIS), definition of operational risk, *III:82*
- Bankruptcy, *I:350, I:366–369, II:577*
- Banks, use of VaR measures, *III:295*
- Barclays Global Risk Model, *II:173, II:193n, II:268*
- Barra models
- E3, *II:256, II:257t, II:261*
 - equity, *II:245–246*
 - fundamental data in, *II:246t*
 - fundamental factor, *II:244–248, II:248–250*
 - risk, *II:256*
 - use of, *II:254n*
- Barrier options, *II:683*
- Basel II Capital Accord, on operational risk, *III:86–87*
- Basic earning power ratio, *II:547, II:549*
- Bayes, Thomas, *I:140, I:196*
- Bayesian analysis
- empirical, *I:154–155*
 - estimation, *I:189*
 - hypothesis comparison, *I:156–157*
 - in parameter estimation, *II:78*
 - and probability, *I:140, I:148*
 - steps of decision making in, *I:141*
 - testing, *I:156–157*
 - use of, *I:18*
- Bayesian inference, *I:151, I:157–158, II:719*
- Bayesian Information Criterion (BIC), *II:703, II:717*
- Bayesian intervals, *I:156, I:170*
- Bayesian methods, and economic theory, *III:142*
- Bayes' theorem, *I:143–148, I:152*
- Behaviors, patterns of, *II:707–710, III:34–35*
- BEKK(1,1,K) model, *II:372*
- Beliefs
- about long-term volatility, *III:408–409*
 - posterior, *I:151–152*
 - prior, *I:152, I:159*
- Bellman's principle, *II:664–665*
- Benchmarks
- choice of, *II:114–115*
 - effect of taxes on, *II:74*
 - fair market, *III:626*
 - modeling of, *II:696*
 - portfolio, *II:272t*
 - for risk, *II:265, III:350, III:354–355*
 - risk in, *II:259*
 - tracking of, *II:67*
 - for trades, *II:117, III:624*
 - use of, *I:41–42, II:66–69*
- Benchmark spot rate curves, *I:222–223*
- Berkowitz transformation, application of, *III:366–367, III:368*
- Bernoulli model, parameter inference in, *II:726–727*
- Bernoulli trials, *I:81, III:170, III:174*
- Bessel function of the third kind, *III:232*
- Best bids/best asks, *II:449–450*
- Best practices, *I:416*
- Beta function, *III:222*
- Betas
- β_{1963} , *I:74–75*
 - β_{1964} , *I:75*
 - β_{1963} *vs.* β_{1964} , *I:76–77*
 - distribution of, *III:222*
 - meanings of, *I:74*
 - in portfolios, *II:273*
 - pricing model, *I:60–61, I:71–72*
 - propositions about, *I:75–77*
 - robust estimates of, *II:442–443*
 - in SL-CAPM models, *I:66–67*
 - two beta trap, *I:74–77*
- Bets, unintended, *II:261, II:263–264, II:264, II:265*
- Better building blocks, *I:36*
- Bias
- from data, *II:204*
 - discretization error, *III:641*
 - estimator, *III:641*
 - survivorship (look-ahead), *II:202, II:204, II:712–713, II:718*
- Bid-ask bounce, *II:455–457*
- Bid-ask spread
- aspects of, *III:597*
 - average hourly, *II:454f*
 - defined, *II:454*
 - under market conditions, *II:455f*
 - risk in, *III:372*
- Binomial experiment, *I:146–148*
- Black, Fischer, *II:468, II:476*
- Black and Scholes
- assumptions of, *I:510*
- Black-Derman-Toy (BDT) model
- defined, *I:492*
 - discussion of, *III:608–609*
 - features of, *III:549*
 - interest rate model, *III:616f*
 - as no arbitrage model, *III:604*
 - use of, *III:300*
- Black-Karasinski (BK) model, *III:548, III:607–608*
- binomial lattice, *III:611*
 - defined, *I:493*
 - features of, *III:604*
 - forms of, *III:600t*
 - interest rate trinomial lattice, *III:615f*
 - trinomial lattice, *III:616f*
- Black-Litterman model
- assumptions with, *I:196–197*
 - derivation of, *I:196–197*
 - discussion of, *I:195–201*
 - with investor's views and market equilibrium, *I:198–199*
 - mixed estimation procedure, *I:200*
 - use of for forecasting returns, *I:193–194, II:112*
 - use of in parameter estimation, *II:78*
 - variance of, *I:200*
- Black-Scholes formula
- for American options, *II:674*
 - with change of time, *III:522, III:524–525*
 - and diffusion equations, *II:654*
 - and Gaussian distribution, *II:732*
 - and Girsanov's theorem, *I:132–133*
 - statistical concepts for, *III:225*
 - use of, *I:126–127, I:136*
 - use of in MATLAB, *III:423–427, III:447*
 - use of with VBA, *III:462–463*
 - and valuation models, *I:271*
- Black-Scholes-Merton stock option pricing formula, *I:557*

- Black-Scholes model
 assumptions of, *I:512, III:655*
 and calibration, *II:681–682*
 for European options, *II:660–662, III:639–640*
 and hedging, *I:410*
 and Merton's model, *I:343*
 for pricing options, *I:487, I:509–510, I:522*
 usefulness of, *I:475*
 use of, *I:272*
 volatility in, *III:653*
- Black volatility, *III:548, III:550*
- Bohr, Niels, *I:123*
- Bond-price valuation model, *III:581–583*
- Bonds
 analytical models for, *I:271–273*
 annuities from, *I:211–212*
 calculating yields on, *II:618*
 callable, *I:24f, I:244–245, III:302–303, III:302f*
 capped floating rate, valuation of, *I:249f*
 changes in prices, *I:373–374*
 computing accrued interest and clean price of, *I:214–215*
 convertible, *I:230, I:271*
 corporate, *I:279, III:598–599*
 coupon-paying, *III:584–586*
 default-free, *I:223*
 determination of value of, *I:211–213*
 discount, *I:212*
 effective duration/convexity of, *I:255, I:256f*
 European convertible, *I:272*
 in European-style calls, *I:440*
 floating-coupon, *I:246–248, I:247f*
 floating-rate callable capped, *I:248*
 floating valuation, *I:253f*
 full (dirty) price, *I:214, I:370*
 futures contracts on, *I:498*
 general principles of valuation, *I:209–216*
 inflation-indexed, *I:278, I:279, I:283–290, I:290–294*
 input information for example, *III:613t*
 interest rate tree for, *I:244f*
 loading of specific, *II:279*
 modeling prices of, *I:490–494*
 and modified or effective duration, *III:299*
 nonpar, *I:232n*
 option-free, *I:241f, I:243*
 options on, *I:252–253, I:498–501, I:501–502*
 planned amortization class (PAC), *III:6*
 plot of convertible functions, *I:273f*
 prediction of yield spreads, *II:336–344*
 price/discount rate relationship, *I:215–216, I:215f*
 prices of, *I:213–214, I:278, I:382, II:727–728*
 prices with effective duration/convexity, *III:300t, III:301t*
 pricing for, *I:498–503, III:588*
 putable, effective duration of, *III:303–304, III:304f*
 regression data for spread application, *II:338–343t*
 relation to CDSs, *I:525–526*
 risk-free, *I:316*
 risk-neutral, *III:586*
 risk-neutral/equilibrium models for, *III:597–598*
 security levels of, *I:375t*
 spreads over time, *I:402f*
 straight, duration of, *III:301–302, III:301f*
 time path of, *I:216*
 valuation of, *I:213–215, I:216–223, I:223, II:730, III:576*
 valuing of, *I:213–214, I:244–246, I:246f*
 volatility of, *I:279*
- Book value, of companies, *II:535*
- Bootstrapping
 parametric, *II:428*
 of spot rate curve, *I:217–220*
 technique for, *II:711–712, II:746*
 usefulness of, *III:325*
 use of, *I:223, III:408*
- Borel functions, *III:508–509*
- Borel measures, *III:199, III:498*
- Borrowers, *III:5, III:70–71, III:74–75, III:598–600*
- Borrowing, *I:72–73, I:479–480*
- Boundary conditions, *II:660*
 need for, *II:661*
- Box-and-whiskers diagrams, use of, *III:329–330n*
- Boxplots, use of, *III:329–330n*
- Brennen-Schwartz model, *III:549*
- Brown, Robert, *III:476*
- Brownian motion, geometric (GBM), *I:95, III:656*
- Brownian motion (BM)
 arithmetic, *I:125, III:492, III:503*
 in binomial models, *I:114*
 bounds of, *III:473–474*
 canonical, *III:478*
 conditions defining, *III:483n*
 defined, *I:95, III:476–479*
 with drift, *III:491*
 early work on, *II:470*
 excursion of, *III:480*
 fractal properties of, *III:479–480, III:480f*
 generated by random walk, *III:479f*
 generating paths for in VBA, *III:463–465*
 geometric, *III:492–493, III:503, III:524*
 and Girsanov's theorem, *I:131, I:263*
 in Ito processes, *III:487–488*
 in Ito's formula, *III:488–489*
 and the Merton model, *I:306*
 one-dimensional standard, *III:477–478*
 paths of, *III:486, III:501–502, III:502f*
 path with deviation zones, *III:537f*
 process of, *I:269–270n*
 properties of, *III:479–481, III:501, III:536*
 in randomness calculations, *III:534–535*
 and stochastic integrals, *III:473*
 time-changed, *III:503–505*
 usefulness of, *III:495–496*
 use of, *I:262*
 variants of, *III:506*
- Bubbles, discovering, *II:396–399*
- Burmeister-Ibbotson-Roll-Ross (BIRR) model, *II:140*
- Burnout effect, *III:17–18, III:24, III:74*
- Burnout factor
 initializing of, *III:22*
- Business cycles, *I:351–352, I:408, II:430–431, II:432–433*
- Businesses, correlation within sectors, *I:411*
- Butterfly's wings, effect of, *II:645*
- Calculus, stochastic, *I:94–97*
- Calendarization, *II:43, II:487–488*
- Calibration
 of derivatives, *I:494*
 effect of, *III:619*
 under GIG model, *II:524*
 of local volatility, *II:681–685*
 need for, *III:604*
 to short forward curve, *III:545–546*
- Callable bonds, *I:462*
- Call options
 defined, *I:439*
 discrepancy measures across maturities of, *II:525t*
 early exercise of American-style, *I:442–443, I:449–450*
 European, *I:125, I:127–129, I:501, I:511, II:522–525, II:679*
 1998 prices of, *II:524t*
 value of, *I:259*

- Calls
 American-style, *I:441–442, I:449*
 error on value of, *II:668f*
 European-style, *I:440–441, I:448–449, I:448t*
- Canonical correlation analysis, *I:556*
- Capital asset pricing model (CAPM).
See CAPM (capital asset pricing model)
- Capital expenditures coverage ratio,
II:575–576
- Capital gains, taxes on, *II:73*
- Caplets, *I:249, III:589–590*
- CAPM
 multifactor, *II:475*
- CAPM (capital asset pricing model).
See also Roy CAPM; SL-CAPM
 application of, *I:60–61*
 areas of confusion, *I:67–68*
 for assessing operational risk,
III:92–93
 in asset pricing, *II:474*
 defined, *I:394*
 and discount factor model, *I:65–66*
 and investor risk, *I:73–74*
 using assumptions under, *I:68–69*
- Caps
 defined, *I:248–251*
 value of, *I:248, III:552–553*
 valuing of, with floors, *I:249–250, I:256*
- Carry, *I:423–426, I:481*
- Carry costs, *I:424–425, I:426, I:435, I:437–438, I:455n, I:481. See also* net cost of carry
- CART (classification and regression trees)
 defined, *II:375*
 example, input variables for, *II:379t*
 example, out-of-sample performance, *II:381t*
 fundamentals of, *II:376–377*
 in stock selection, *II:378–381*
 strengths and weaknesses of,
II:377–378
 uses of, *II:381*
- Cash-and-carry trade, *I:480, I:481, I:487*
- Cash concept, *II:567*
- Cash flows
 accounting for, *III:306*
 analysis of, *II:574–577, III:4–5*
 for bond class, *III:9t*
 of bonds, *I:211*
 cash flow at risk (CFaR), *III:376–378*
 classification of, *II:567*
 defined, *I:209–210, II:539, III:4*
 direct *vs.* indirect reporting method,
II:567
 discounted, *I:225*
 discrete, *I:429*
 distribution analysis *vs.* benchmark,
III:310
 estimation of, *I:209–210, II:21–23*
 expected, *I:211*
 factors in, *III:31–32, III:377*
 form residential mortgage loans,
III:62
 futures *vs.* forwards, *I:431t*
 future value of, *II:603f*
 influences on, *III:44*
 interest coverage ratio of, *II:561, II:575*
 interim, *I:482*
 for loan pool, *III:9t*
 measurement of, *II:565–566, III:14*
 monthly, *III:52–54, III:53t*
 net free (NFCF), *II:572–574, II:578*
 in OAS analysis, *I:259*
 perpetual stream of, *II:607–608*
 sources of, *II:540–541, II:569t*
 in state dependent models,
I:351–352
 statement of, *II:539–541, II:566–567*
 time patterns of, *II:607–611*
 and time value of money, *II:595–596*
 time value of series of, *II:602–607*
 for total return receivers, *I:542*
 for Treasuries, *I:219, III:564–565*
 types of in assessing liquidity risk,
III:378
 use of information on, *II:576–577*
 valuation of, *II:618–619*
vs. free cash flow, *II:22–23*
- Cash flow statements
 example of, *II:541*
 form of, *II:26t*
 information from, *II:577–578*
 reformatting of, *II:569t*
 restructuring of, *II:568*
 sample, *II:547t*
 use of, *II:24–26*
- Cash flow-to-debt ratio, *II:576*
- Cash-out refinancing, *III:66, III:69*
- Cash payments, *I:486–487, III:377*
- Categorizations, determining usefulness of, *II:335*
- Cauchy, Augustin, *II:655*
- Cauchy initial value problem, *II:655, II:656, II:656f, II:657*
- CAViAR (conditional autoregressive value at risk), *II:366*
- CDOs (collateralized debt obligations), *I:299, I:525, III:553, III:645*
- CDRs (conditional default rates)
 in cash flow calculators, *III:34*
 defaults measured by, *III:58–59*
 defined, *III:30–31*
 monthly, *III:62t*
 projections for, *III:35f*
 in transition matrices, *III:35f*
- CDSs (credit default swaps)
 basis, *I:232*
 bids on, *I:527*
 cash basis, *I:402*
 discussion of, *I:230–232*
 fixed premiums of, *I:530–531*
 hedging with, *I:418*
 illustration of, *I:527*
 initial value of, *I:538*
 maturity dates, *I:526*
 payoff and payment structure of,
I:534f
 premium payments, *I:231f, I:533–535*
 pricing models for, *I:538–539*
 pricing of by static replication,
I:530–532
 pricing of single-name, *I:532–538*
 quotations for, *I:413*
 risk and sensitivities of, *I:536–537*
 spread of, *I:526*
 unwinding of, *I:538*
 use of, *I:403, I:413, II:284*
 valuation of, *I:535–536*
 volume of market, *I:414*
- Central limit theorem
 defined, *I:149n, III:209–210, III:640*
 and the law of large numbers,
III:263–264
 and random number generation,
III:646
 and random variables, *II:732–733*
- Central tendencies, *II:353, II:354, II:355*
- Certainty equivalents, *II:723–724, II:724–725*
- CEV (constant elasticity of variance),
III:550, III:551f, III:654–655
- Chambers-Mallows-Stuck generator,
II:743–744
- Change of measures, *III:509–517, III:516t*
- Change of time methods (CTM)
 applications of, *III:522–527*
 discussion of, *III:519–522*
 general theory of, *III:520–521*
 main idea of, *III:519–520, III:527*
 in martingale settings, *III:522–523*
 in stochastic differential equation setting, *III:523*
- Chaos, defined, *II:653*
- Chaos: *Making a New Science* (Gleick),
II:714
- Characteristic function
vs. probability density function,
II:743

- Characteristic lines, *II:316, II:318t, II:344–348, II:345–347t*
- Chebyshev inequalities, *III:210, III:225*
- Chen model, *I:493*
- Chi-square distributions, *I:388–389, III:212–213*
- Cholesky factor, *I:380*
- Chow test, *II:336, II:343, II:344, II:350*
- CID (conditionally independent defaults) models, *I:320, I:321–322, I:333*
- CIR model, *I:498, I:500–501, I:502*
- Citigroup, *I:302, I:408f, I:409f*
- CLA (critical line algorithm), *I:73*
- Classes
criteria for, *II:494*
- Classical tempered stable (CTS) distribution, *II:741–742, II:741f, II:742f, II:743–744, III:512*
- Classification, and Bayes' Theorem, *I:145*
- Classification and regression trees (CART). *See* CART (classification and regression trees)
- Classing, procedure for, *II:494–498*
- Clearinghouses, *I:478*
- CME Group, *I:489–490*
- CMOs (collateralized mortgage obligations), *III:598, III:645*
- Coconut markets, *I:70*
- Coefficients
binomial, *III:171, III:187–191*
of determination, *II:315*
estimated, *II:336–337*
- Coherent risk measures, *III:327–329*
and VaR, *III:329*
- Coins, fair/unfair, *III:169, III:326–327*
- Cointegrated models, *II:503*
- Cointegration
analysis of, *II:381t*
defined, *II:383*
empirical illustration of, *II:388–393*
technique of, *II:384–385*
testing for, *II:386–387*
test of, *II:394t, II:396t*
use of, *II:397*
- Collateralized debt obligations (CDOs), *I:299, I:525, III:553, III:645*
- Collateralized mortgage obligations (CMOs), *III:598, III:645*
- Collinearity, *II:329–330*
- Commodities, *I:279, I:556, I:566*
- Companies. *See* firms
- Comparison principals, *II:676*
- Comparisons *vs.* testing, *I:156*
- Complete markets, *I:103–104, I:119, I:133, I:461*
- Complexity, profiting from, *II:57–58*
- Complexity (Waldrop), *II:699*
- Complex numbers, *II:591–592, II:592f*
- Compounding. *See also* interest and annual percentage rates, *II:616*
continuous, *II:599, II:617*
determining number of periods, *II:602*
discrete *vs.* continuous, *III:570–571*
formula for growth rate, *II:8*
more than once per year, *II:598–599*
and present value, *II:618*
- Comprehensive Capital Analysis and Review, *I:300*
- Comprehensive Capital Assessment Review, *I:412*
- Computational burden, *III:643–644*
- Computers. *See also* various software applications
increased use of, *III:137–138*
introduction of into finance, *II:480*
modeling with, *I:511, II:695*
random walk generation of, *II:708*
in stochastic programming, *III:124, III:125–126*
- Concordance, defined, *I:327*
- Conditional autoregressive value at risk (CAViaR), *II:366*
- Conditional default rate (CDR). *See* CDRs (conditional default rates)
- Conditionally independent defaults (CID) models, *I:320, I:321–322, I:323*
- Conditioning/conditions, *I:24, II:307–308, II:361, II:645*
- Confidence, *I:200, I:201, II:723, III:319*
- Confidence intervals, *II:440, III:338t, III:399–400, III:400f*
- Conglomerate discounts, *II:43*
- Conseco, debt restructure of, *I:529*
- Consistency, notion of, *II:666–667*
- Constant elasticity of variance (CEV), *III:550, III:551f, III:654–655*
- Constant growth dividend discount model, *II:7–9*
- Constraints, portfolio
cardinality, *II:64–65*
common, *III:146*
commonly used, *II:62–66, II:84*
holding, *II:62–63*
minimum holding/transaction size, *II:65*
nonnegativity, *I:73*
real world, *II:224–225*
round lot, *II:65–66*
setting, *I:192*
turnover, *II:63*
on weights of, *I:191–192*
- Constraint sets, *I:21, I:28, I:29*
- Consumer Price Index (CPI), *I:277–278, I:291f, I:292, I:292f*
- Consumption, *I:59–60, II:360, III:570*
- Contagion, *I:320, I:324, I:333*
- Contingent claims
financial instruments as, *I:462*
incomplete markets for, *I:461–462*
unit, *I:458*
use of analysis, *I:463*
utility maximization in markets, *I:459–461*
value of, *I:458–459*
- Continuity, formal treatment of, *II:583–584*
- Continuous distribution function (c.d.f.), *III:167, III:196, III:205, III:345–346, III:345f*
- Continuous distribution function $F(a)$, *III:196*
- Continuous time/continuous state, *III:578*
- Continuous-time processes, change of measure for, *III:511–512*
- Control flow statements in VBA, *III:458–460*
- Control methods, stochastic, *I:560*
- Convenience yields, *I:424, I:439*
- Convergence analysis, *II:667–668*
- Conversion, *I:274, I:445*
- Convexity
in callable bonds, *III:302–303*
defined, *I:258–259, III:309*
effective, *III:13, III:300–304, III:617t*
measurement of, *III:13–14, III:304–305*
negative, *III:14, III:49, III:303*
positive, *III:13*
use of, *III:299–300*
- Convex programming, *I:29, I:31–32*
- Cootner, Paul, *III:242*
- Copulas
advantages of, *III:284*
defined, *III:283*
mathematics of, *III:284–286*
usefulness of, *III:287*
visualization of bivariate independence, *III:285f*
visualization of Gaussian, *III:287f*
- Corner solutions, *I:200*
- Correlation coefficients
relation to R^2 , *II:316*
and Theil-Sen regression, *II:444*
use of, *III:286–287*
- Correlation matrices, *II:160t, II:163t, III:396–397*
- Correlations
in binomial distribution, *I:118*
computation of, *I:92–93*

- concept of, *III:283*
- drawbacks of, *III:283–284*
- between periodic increments, *III:540t*
- and portfolio risk, *I:11*
- robust estimates of, *II:443–446*
- serial, *II:220*
- undesirable, *I:293*
- use of, *II:271*
- Costs, net financing, *I:481*
- Cotton prices, model of, *III:383*
- Countable additivity, *III:158*
- Counterparts, robust, *II:81*
- Countries, low- vs. high inflation, *I:290*
- Coupon payments, *I:212, III:4*
- Coupon rates, computing of, *III:548–549*
- Courant-Friedrichs-Lewy (CFL) conditions, *II:657*
- Covariance
 - calculation of between assets, *I:8–9*
 - estimators for, *I:38–40, I:194–195*
 - matrix, *I:38–39, I:155, I:190*
 - relationship with correlation, *I:9*
 - reliability of sample estimates, *II:77*
 - use of, *II:370–371*
- Covariance matrices
 - decisions for interest rates, *III:406*
 - eigenvectors/eigenvalues, *II:160t*
 - equally weighted moving average, *III:402–403*
 - frequency of observations for, *III:404*
 - graphic of, *II:161t*
 - residuals of return process of, *II:162t*
 - of RiskMetrics™ Group, *III:412–413*
 - statistical methodology for, *III:398–399*
 - of ten stock returns, *II:159t*
 - use of, *II:158–159, II:169*
 - using EWMA in, *III:411*
- Coverage ratios, *II:560–561*
- Cox-Ingersoll-Ross (CIR) model, *I:260, I:491–492, I:547, I:548, III:546–547, III:656*
- Cox processes, *I:315–316, II:470–471*
- Cox-Ross-Rubenstein model, *I:510, I:522, II:678*
- CPI (Consumer Price Index), *I:277–278, I:291f, I:292, I:292f*
- CPRs (conditional prepayment rates). *See* prepayment, conditional
- CPR vector, *III:74. See also* prepayment, conditional
- Cramer, Harald, *II:470–471*
- Crank-Nicolson schemes, *II:666, II:669, II:674, II:680*
- Crank Nicolson-splitting (CN-S) schemes, *II:675*
- Crashmetrics, use of, *III:379, III:380*
- Credible intervals, *I:156*
- Credit-adjusted spread trees, *I:274*
- Credit crises
 - of 2007, *III:74*
 - of 2008, *III:381*
 - data from and DTS model, *I:396*
 - in Japan, *I:417*
- Credit curing, *III:73*
- Credit default swaps (CDSs). *See* CDSs (credit default swaps)
- Credit events
 - and credit loss, *I:379*
 - in default swaps, *I:526, I:528–530*
 - definitions of, *I:528*
 - descriptions of most used, *I:528t*
 - exchanges/payments in, *I:231f*
 - in MBS turnover, *III:66*
 - prepayments from, *III:49–50*
 - protection against, *I:230*
 - and simultaneous defaults, *I:323*
- Credit hedging, *I:405*
- Credit inputs, interaction of, *III:36–38*
- Credit loss
 - computation of, *I:382–383*
 - distribution of, *I:369f*
 - example of distribution of, *I:386f*
 - simulated, *I:389*
 - steps for simulation of, *I:379–380*
- Credit models, *I:300, I:302, I:303*
- Credit performance, evolution of, *III:32–36*
- Credit ratings
 - categories of, *I:362*
 - consumer, *I:302*
 - disadvantages of, *I:300–301*
 - implied, *I:381–382*
 - maturity of, *I:301*
 - reasons for, *I:300*
 - risks for, *II:280–281, II:280t*
 - use of, *I:309*
- Credit risk
 - common, *I:322*
 - counterparty, *I:413*
 - in credit default swaps, *I:535*
 - defined, *I:361*
 - distribution of, *I:377*
 - importance of, *III:81*
 - measures for, *I:386f*
 - modeling, *I:299–300, I:322, III:183*
 - quantification of, *I:369–372*
 - reports on, *II:278–281*
 - shipping, *I:566*
 - and spread duration, *I:391–392*
 - vs. cash flow risk, *III:377–378*
- Credit scores, *I:300–302, I:301–302, I:309, I:310n*
- Credit spreads
 - alternative models of, *I:405–406*
 - analysis with stock prices, *I:305t*
 - applications of, *I:404–405*
 - decomposition, *I:401–402*
 - drivers of, *I:402*
 - interpretation of, *I:403–404*
 - model specification, *I:403*
 - relationship with stock prices, *I:304*
 - risk in, *II:279t*
 - use of, *I:222–223*
- Credit support, evaluation of, *III:39–40*
- Credit value at risk (CVaR). *See* CVaR
- Crisis situations, estimating liquidity in, *III:378–380*
- Critical line algorithm (CLA), *I:73*
- Cross-trading, *II:85n*
- Cross-validation, leave-one-out, *II:413–414*
- Crude oil, *I:561, I:562*
- Cumulation, defined, *III:471*
- Cumulative default rate (CDX), *III:58*
- Cumulative frequency distributions, *II:493f, II:493t, II:498–499*
- formal presentation of, *II:492–493*
- Currency put options, *I:515*
- Current ratio, *II:554*
- Curve imbalances, *II:270–271*
- Curve options, *III:553*
- Curve risk, *II:275–278*
- CUSIPs/ticker symbols, changes in, *II:202–203*
- CVaR (credit value at risk), *I:384–385, I:385–386, II:68, II:85n, III:392t. See also* value at risk (VaR)
- Daily increments of volatility, *III:534*
- Daily log returns, *II:407–408*
- Dark pools, *II:450, II:454*
- Data. *See also* operational loss data
 - absolute, *II:487–488*
 - acquisition and processing of, *II:198*
 - alignment of, *II:202–203*
 - amount of, *I:196*
 - augmentation of, *I:186n*
 - availability of, *II:202, II:486*
 - backfilling of, *II:202*
 - bias of, *II:204, II:713*
 - bid-ask aggregation techniques for, *II:457f*
 - classification of, *II:499–500*
 - collection of, *II:102, II:103f*
 - cross-sectional, *II:201, II:488, II:488f*
 - in forecasting models, *II:230*
 - frequency of, *II:113, II:368, II:462–463, II:500*
 - fundamental, *II:246–247*
 - generation of, *II:295–296*

- Data (*Continued*)
 high-frequency (HFD) (*See*
 high-frequency data (HFD))
 historical, *II:77–78, II:122, II:172*
 housing bubble, *II:397–399*
 importing into MATLAB,
III:433–434
 industry-specific, *II:105*
 integrity of, *II:201–203*
 levels and scale of, *II:486–487*
 long-term, *III:389–390*
 in mean-variance, *I:193–194*
 misuse of, *II:108*
 on operational loss, *III:99*
 from OTC business, *II:486*
 patterns in, *II:707–708*
 pooling of, *III:96*
 of precision, *I:158*
 preliminary analysis of, *III:362*
 problems in for operational risk,
III:97–98
 qualitative *vs.* quantitative,
II:486
 quality of, *II:204, II:211, II:452–453,*
II:486, II:695
 reasons for classification of,
II:493–494
 for relative valuation, *II:34–35*
 restatements of, *II:202*
 sampling of, *II:459f, II:711*
 scarcity of, *II:699–700, II:703–704,*
II:718
 sorting and counting of, *II:488–491*
 standardization of, *II:204, III:228*
 structure/sample size of, *II:703*
 types of, *II:486–488*
 underlying signals, *II:111*
 univariate, defined, *II:485*
 working with, *II:201–206*
- Databases
 Compustat Point-In-Time, *II:238*
 Factiva, *II:482*
 Institutional Brokers Estimate
 System (IBES), *II:238*
 structured, *II:482*
 third-party, *II:198, II:211n*
- Data classes, criteria for, *II:500*
 Data generating processes (DGPs),
II:295–296, II:298f, II:502, II:702,
III:278
- Data periods, length of, *III:404*
 Data series, effect of large number of,
II:708–709
- Data sets, training/test, *II:710–711*
 Data snooping, *II:700, II:710–712,*
II:714, II:717, II:718
- Datini, Francesco, *II:479–480*
 Davis-Lo infectious defaults model,
I:324
- Days payables outstanding (DPO),
 calculation of, *II:553–554*
- Days sales outstanding (DSO),
 calculation of, *II:553*
- DCF (discounted cash flow) models,
II:16, II:44–45
- DDM (dividend discount models). *See*
 dividend discount models
 (DDM)
- Debt
 long-term, in financial statements,
II:542
 models of risky, *I:304–307*
 restructuring of, *I:230*
 risky, *I:307–308*
- Debt-to-assets ratio, *II:559*
 Debt-to-equity ratio, *II:559*
- Decomposition models
 active/passive, *III:19*
- Default correlation, *I:317–318*
 contagion, *I:353–354*
 cyclical, *I:352, I:353*
 linear, *I:320–321*
 measures of, *I:320–321*
 tools for modeling, *I:319–333*
- Default intensity, *III:225*
- Default models, *I:321–322, I:370f*
- Default probabilities
 adjustments in real time, *I:300–301*
 between companies, *I:412–413*
 cyclical rise and fall, *I:408f, I:409f*
 defined, *I:299–300*
 effect of business cycle on, *I:408*
 effect of rating outlooks on,
I:365–366
 empirical approach to, *I:362–363*
 five-year (Bank of America and
 Citigroup), *I:301f, I:302f*
 merits of approaches to, *I:365*
 Merton's approach to, *I:363–365*
 probability of, *II:727, II:727f, II:728f*
 and survival, *I:533–535*
 and survival probability, *I:323–324*
 term structure of, *I:303*
 time span of, *I:302–303*
vs. ratings and credit scores,
I:300–302
 for Washington Mutual, *I:415f,*
I:416f
 of Washington Mutual, *I:415f,*
I:416f
- Defaults
 annual rates of, *I:363*
 and Bernoulli distributions,
III:169–170
 calculation of monthly, *III:61t*
 clustering of, *I:324–325*
 contagion, *I:320*
 copulas for times, *I:329–331*
- correlation of between companies,
I:411
 cost of, *I:401, I:404f*
 dollar amounts of, *III:59f*
 effect of, *I:228, III:645*
 event *vs.* liquidation, *I:349*
 factors influencing, *III:74–75*
 first passage model of, *I:349*
 historical database of, *I:414*
 intensity of, *I:330, I:414*
 looping, *I:324–325*
 measures of, *III:58–59*
 in Merton approach, *I:306*
 Moody's definition of, *I:363*
 predictability of, *I:346–347*
 and prepayments, *III:49–50,*
III:76–77
 process, relationship to recovery
 rate, *I:372*
 pseudo intensities, *I:330*
 rates of cumulative/conditional,
III:63
 recovery after, *I:316–317*
 risk of, *I:210*
 simulation of times, *I:322–324, I:325*
 threshold of, *I:345–346*
 times simulation of, *I:319*
 triggers for, *I:347–348*
 variables in, *I:307–308*
- Default swaps
 assumptions about, *I:531–532*
 and credit events, *I:530*
 digital, *I:537*
 discussion of, *I:526–528*
 market relationship with cash
 market, *I:530*
 and restructuring, *I:528–529*
 value of spread, *I:534*
- Default times, *I:332*
- Definite covariance matrix, *II:445*
- Deflators, *I:129, I:136*
- Degrees, in ordinary differential
 equations, *II:644–645*
- Degrees of freedom (DOF)
 across assets and time, *II:735–736*
 in chi-square distribution, *III:212*
 defined, *II:734*
 for Dow Jones Industrial Average
 (DJIA), *II:735–737, II:737f*
 prior distribution for, *I:177*
 range of, *I:187n*
 for S&P 500 index stock returns,
II:735–736, II:736f
- Delinquency measures, *III:57–58*
- Delivery date, *I:478*
- Delta, *I:509, I:516–518, I:521*
- Delta-gamma approximation, *I:519,*
III:644–645
- Delta hedging, *I:413, I:416, I:418, I:517*

- Delta profile, *I:518f*
- Densities
- beta, *III:108f*
 - Burr, *III:110f*
 - closed-form solutions for, *III:243*
 - exponential, *III:105–106, III:105f*
 - gamma, *III:108f*
 - Pareto, *III:109f*
 - posterior, *I:170f*
 - two-point lognormal, *III:111f*
- Density curves, *I:147f*
- Density functions
- asymmetric, *III:205f*
 - of beta distribution, *III:222f*
 - chi-square distributions, *III:213f*
 - common means, different variances, *III:203f*
 - computing probabilities from, *III:201*
 - discussion of, *III:197–200*
 - of *F*-distribution, *III:217f*
 - histogram of, *III:198f*
 - of log-normal distribution, *III:223f*
 - and normal distribution, *II:733*
 - and probability, *III:206*
 - rectangular distributions, *III:220*
 - requirements of, *III:198–200*
 - symmetric, *III:204f*
 - of *t*-distribution, *III:214f*
- Dependence, *I:326–327, II:305–308*
- Depreciation, *II:22*
- accumulated, *II:533–534*
 - expense *vs.* book value, *II:539f*
 - expense *vs.* carrying value, *II:540f*
 - in financial statements, *II:537–539*
 - on income statements, *II:536*
 - methods of allocation, *II:537–538*
- Derivatives
- construction of, *II:586–587*
 - described, *II:585–586*
 - embedded, *I:462*
 - energy, *I:558*
 - exotic, *I:558, I:559–560*
 - of functions, defined, *II:593*
 - and incomplete markets, *I:462*
 - interest rate, *III:589–590*
 - nonlinearity of, *III:644–645*
 - OTC, *I:538*
 - pricing of, *I:58, III:594–596*
 - pricing of financial, *III:642–643*
 - relationship with integrals, *II:590*
 - for shipping assets, *I:555, I:558, I:565–566*
 - use of instruments, *I:477*
 - valuation and hedging of, *I:558–560*
 - vanilla, *I:559*
- Derman, Emanuel, *II:694*
- Descriptors, *II:140, II:246–247, II:256*
- Determinants, *II:623*
- Deterministic methods
- usefulness of, *II:685*
- Diagonal VEC model (DVEC), *II:372*
- Dice, and probability, *III:152, III:153, III:155–156, III:156t*
- Dickey-Fuller statistic, *II:386–387*
- Dickey-Fuller tests, *II:514*
- Difference, notation of, *I:80*
- Differential equations
- classification of, *II:657–658*
 - defined, *I:95, II:644, II:657*
 - first-order system of, *II:646*
 - general solutions to, *II:645*
 - linear, *II:647–648*
 - linear ordinary, *II:644–645*
 - partial (PDE), *II:643, II:654–657*
 - stochastic, *II:643–644*
 - systems of ordinary, *II:645–646*
 - usefulness of, *II:658*
- Diffusion, *III:539, III:554–555*
- Diffusion invariance principle, *I:132*
- Dimensionality, curse of, *II:673, III:127*
- Dirac measures, *III:271*
- Directional measures, *II:428, II:429*
- Dirichlet boundary conditions, *II:666*
- Dirichlet distribution, *I:181–183, I:186–187n*
- Discounted cash flow (DCF) models, *II:16, II:44–45*
- Discount factors, *I:57–58, I:59–62, I:60, II:600–601*
- Discount function
- calculation of, *III:571*
 - defined, *III:563*
 - discussion of, *III:563–565*
 - forward rates from, *III:566–567*
 - graph of, *III:563f*
 - for on-the-run Treasuries, *III:564–565*
- Discounting, defined, *II:596*
- Discount rates, *I:211, I:212, I:215–216, II:6*
- Discovery heuristics, *II:711*
- Discrepancies, importance of small, *II:696*
- Discrete law, *III:165–169*
- Discrete maximum principle, *II:668*
- Discretization, *I:265, II:669f, II:672*
- Disentangling, *II:51–56*
- complexities of, *II:55–56*
 - predictive power of, *II:54–55*
 - return revelation of, *II:52–54*
 - usefulness of, *II:52, II:58*
- Dispersion measures, *III:352, III:353–354, III:357*
- Dispersion parameters, *III:202–205*
- Distress events, *I:351*
- Distributional measures, *II:428*
- Distribution analysis, cash flow, *III:310*
- Distribution function, *III:218f, III:224f*
- Distributions
- application of hypergeometric, *III:177–178*
 - beliefs about, *I:152–153*
 - Bernoulli, *III:169–170, III:185t*
 - beta, *I:148, III:108*
 - binomial, *I:81f, III:170–174, III:185t, III:363*
 - Burr, *III:109–110*
 - categories for extreme values, *II:752*
 - common loss, *III:112t*
 - commonly used, *III:225*
 - conditional, *III:219*
 - conditional posterior, *I:178–179, I:182–183, I:184–185*
 - conjugate prior, *I:154*
 - continuous probability, *III:195–196*
 - discrete, *III:185t*
 - discrete cumulative, *III:166*
 - discrete uniform, *III:183–184, III:185t, III:638f*
 - empirical, *II:498, III:104–105, III:105f*
 - exponential, *III:105–106*
 - finite-dimensional, *II:502*
 - of Fréchet, Gumbel and Weibull, *III:267f*
 - gamma, *III:107–108, III:221–222*
 - Gaussian, *III:210–212*
 - Gumbel, *III:228, III:230*
 - heavy-tailed, *I:186n, II:733, III:109, III:260*
 - hypergeometric, *III:174–178, III:185t*
 - indicating location of, *III:235*
 - infinitely divisible, *III:253–256, III:253t*
 - informative prior, *I:152–153*
 - inverted Wishart, *I:172*
 - light- *vs.* heavy-tailed, *III:111–112*
 - lognormal, *III:106, III:106f, III:538–539*
 - mixture loss, *III:110–111*
 - for modeling applications, *III:257*
 - multinomial, *III:179–182, III:185t*
 - non-Gaussian, *III:254*
 - noninformative prior, *I:153–154*
 - normal (*See* normal distributions)
 - parametric, *III:201*
 - Poisson, *I:142, III:182–183, III:185t, III:217–218*
 - Poisson probability, *III:187t*
 - posterior, *I:147–148, I:165, I:166–167, I:169–170, I:177, I:183–184*
 - power-law, *III:262–263*
 - predictive, *I:167*
 - prior, *I:177, I:181–182, I:196*
 - proposal, *I:183–184*
 - representation of stable and CTS, *II:742–743*

- Distributions (*Continued*)
 spherical, II:310
 stable, III:238, III:242, III:264–265, III:384 (*See also* α -stable distributions)
 subexponential, III:261–262
 tails of, III:112*f*, III:648
 tempered stable, III:257, III:382
 testing applied to truncated, III:367
- Diversification, II:57–58
 achieving, I:10
 and cap weighting, I:38
 and credit default swaps, I:413–414
 example of, I:15
 international, II:393–396
 Markowitz's work on, II:471
- Diversification effect, III:321
- Diversification indicators, I:192
- Dividend discount models (DDM)
 applied to electric utilities, II:12*t*
 applied to stocks, II:16–17
 basic, II:5
 constant growth, II:7–9, II:17–18
 defined, II:14
 finite life general, II:5–7
 free cash flow model, II:21–23
 intuition behind, II:18–19
 multiphase, II:9–10
 non-constant growth, II:18
 predictive power of, II:54
 in the real world, II:19–20
 stochastic, II:10–12, II:12*t*
- Dividend payout ratio, II:4, II:20
- Dividends
 expected growth in, II:19
 forecasting of, II:6
 measurement of, II:3–4, II:14
 per share, II:3–4
 reasons for not paying, II:27
 required rate of return, II:19
 and stock prices, II:4–5
- Dividend yield, II:4, II:19
- Documentation
 of model risk, II:696, II:697
- Dothan model, I:491, I:493
- Dow Jones Global Titans 500 (DJGTI), II:490*t*, II:491*t*
- Dow Jones Industrial Average (DJIA)
 in comparison of risk models, II:747–751
 components of, II:489*t*
 fitted stable tail index for, II:740*f*
 frequency distribution in, II:489*t*
 performance (January 2004 to June 2011), II:749*f*
 relative frequencies, II:491*t*
 stocks by share price, II:492*t*
- Drawing without replacement, III:174–177
- Drawing with replacement, III:170, III:174, III:179–180
- Drift
 effects of, III:537
 of interest rates, I:263
 in randomness calculations, III:535
 in random walks, I:84, I:86
 time increments of, I:83
 of time series, I:80
 as variable, III:536
- DTS (duration times spread), I:392, I:393–394, I:396–398
- Duffie-Singleton model, I:542–543
- Dupire's formula, II:682–683, II:685
- DuPont system, II:548–551, II:551*f*
- Duration
 calculations of real yield and inflation, I:286
 computing of, I:285
 defined, I:284, III:309
 effective, III:300–304, III:617*t*
 effective/option adjusted, III:13
 empirical, of common stock, II:318–322, II:319–322*t*
 estimation of, II:323*t*
 measurement of, III:12–13, III:304–305
 models of, II:461
 modified *vs.* effective, III:299
- Duration/convexity, effective, I:255, I:256*f*
- Duration times spread (DTS). *See* DTS (duration times spread)
- Durbin-Watson test, III:647
- Dynamical systems
 equilibrium solution of, II:653
 study of, II:651
- Dynamic conditional correlation (DCC) model, II:373
- Dynamic term structures, III:576–577, III:578–579, III:591
- Early exercise, I:447, I:455. *See* calls, American-style; options
- Earnings before interest, taxes, depreciation and amortization (EBITDA), II:566
- Earnings before interest and taxes (EBIT), II:23, II:547, II:556
- Earnings growth factor, II:223
- Earnings per share (EPS), II:20–21, II:38–39, II:537
- Earnings revisions factor, II:207, II:209*f*
- EBITDA/EV factor
 correlations with, II:226
 examples of, II:203, II:203*f*, II:207, II:208*f*
 in models, II:232, II:238–239
 use of, II:222–223
- Econometrics
 financial, II:295, II:298–300, II:301–303
 modeling of, II:373, II:654
- Economic cycles, I:537, II:42–43
- Economic intuition, II:715–716
- Economic laws, changes in, II:700
- Economy
 states of, I:49–50, II:518–519, III:476
 term structures in certain, III:567–568
 time periods of, II:515–516
- Economy as an Evolving Complex System, The* (Anderson, Arrow, & Pines), II:699
- Educated guesses, use of, I:511
- EE (explicit Euler) scheme, II:674, II:677–678
- Effective annual rate (EAR), interest, II:616–617
- Efficiency
 in estimation, III:641–642
- Efficient frontier, I:13–14, I:17*f*, I:289*f*
- Efficient market theory, II:396, III:92
- Eggs, rotten, I:457–458
- Eigenvalues, II:627–628, II:705, II:706–707*f*, II:707*t*
- Einstein, Albert, II:470
- Elements, defined, III:153–154
- Embedding problem, and change of time method, III:520
- Emerging markets, transaction costs in, III:628
- EM (expectation maximization) algorithm, II:146, II:165
- Empirical rule, III:210, III:225
- Endogenous parameterization, III:580–581
- Energy
 cargoes of, I:561–562
 commodity price models, I:556–558
 forward curves of, I:564–565
 power plants and refineries, I:563
 storage of, I:560–561, I:563–564
- Engle-Granger cointegration test, II:386–388, II:391–392, II:395
- Entropy, III:354
- EPS (earnings per share), II:20–21, II:38–39, II:537
- Equally weighted moving average, III:400–402, III:406–407, III:408–409
- Equal to earnings before interest and taxes (EBIT), II:23, II:547, II:556
- Equal-variance assumption, I:164, I:167
- Equations
 difference, homogenous *vs.* nonhomogenous, II:638

- difference *vs.* differential, *II:629*
 diffusion, *II:654–656, II:658n*
 error-correction, *II:391, II:395f*
 homogeneous linear difference,
 II:639–642, II:641f
 homogenous difference, *II:630–634,*
 II:631–632f, II:633–634f, II:642
 linear, *II:623–624*
 linear difference, systems of,
 II:637–639
 matrix characteristics of, *II:628*
 no arbitrage, *III:612, III:617–619*
 nonhomogeneous difference,
 II:634–637, II:635f, II:637–638f
 stochastic, *III:478*
- Equilibrium**
 and absolute valuation models,
 I:260
 defined, *II:385–386*
 dimensions of, *III:601*
 in dynamic term structure models,
 III:576
 expectations for, *II:112*
 expected returns from, *II:112*
 modeling of, *III:577, III:594*
 in supply and demand, *III:568*
- Equilibrium models**
 use of, *III:603–604*
- Equilibrium term structure models,**
 III:601
- Equities, I:279**
 investing in, *II:89–90*
- Equity**
 on the balance sheet, *II:535*
 changes in homeowner, *III:73*
 in homes, *III:69*
 as option on assets, *I:304–305*
 shareholders', *II:535*
- Equity markets, II:48**
- Equity multipliers, II:550**
- Equity risk factor models, II:173–178**
- Equivalent probability measures,**
 I:111, III:510–511
- Ergodicity, defined, II:405**
- Erlang distribution, III:221–222**
- Errors. See also estimation error;**
 standard errors
 absolute percentages of, *II:525f,*
 II:526f
 estimates of, *II:676*
 in financial models, *II:719*
 a posteriori estimates, *II:672–673*
 sources of, *II:720*
 terms for, *II:126*
 in variables problem, *II:220*
- Esscher transform, III:511, III:514**
- Estimates/estimation**
 confidence in, *I:199*
 consensus, *II:34–35*
 equations for, *I:348–349*
 in EVT, *III:272–274*
 factor models in, *II:154*
 with GARCH models, *II:364–365*
 in-house from firms, *II:35*
 maximum likelihood, *II:311–313*
 methodology for, *II:174–176*
 and PCA, *II:167f*
 posterior, *I:176*
 posterior point, *I:155–156*
 processes for, *I:193, II:176*
 properties of for EWMA, *III:410–411*
 robust, *I:189*
 techniques of, *II:330*
 use of, *II:304*
- Estimation errors**
 accumulation of, *II:78*
 in the Black-Litterman model, *I:201*
 covariance matrix of, *III:139–140*
 effect of, *I:18*
 pessimism in, *III:143*
 in portfolio optimization, *II:82,*
 III:138–139
 sensitivity to, *I:191*
 and uncertainty sets, *III:141*
- Estimation risk, I:193**
 minimizing, *III:145*
- Estimators**
 bias in, *III:641*
 efficiency in, *III:641–642*
 equally weighted average,
 III:400–402
 factor-based, *I:39*
 terms used to describe, *II:314*
 unbiased, *III:399*
 variance, *II:313*
- ETL (expected tail loss), III:355–356**
- Euler approximation, II:649–650,**
 II:649f, II:650f
- Euler constant, III:182**
- Euler schemes, explicit/implicit, II:666**
- Europe**
 common currency for, *II:393*
 risk factors of, *II:174*
- European call options**
 Black-Scholes formula for,
 III:639–640
 computed by different methods,
 III:650–651, III:651f
 explicit option pricing formula,
 III:526–527
 pricing by simulation in VBA,
 III:465–466
 pricing in Black-Scholes setting,
 III:649
 simulation of pricing, *III:444–445,*
 III:462–463
 and term structure models,
 III:544–545
- European Central Bank, I:300**
- Events**
 defined, *III:85, III:162, III:508*
 effects of macroeconomic, *II:243–244*
 extreme, *III:245–246, III:260–261,*
 III:407
 identification of, *II:516*
 mutually exclusive, *III:158*
 in probability, *III:156*
 rare, *III:645*
 rare *vs.* normal, *I:262*
 tail, *III:88n, III:111, III:118*
 three- δ , *III:381–382*
- EVT (extreme value theory). See**
 extreme value theory (EVT)
- EWMA (exponentially weighted**
 moving averages), III:409–413
- Exceedance observations, III:362–363**
- Exceedances, of VaR, III:325–326,**
 III:339
- Excel**
 accessing VBA in, *III:477*
 add-ins for, *I:93, III:651*
 data series correlation in, *I:92–93*
 determining corresponding
 probabilities in, *III:646*
 Excel Link, *III:434*
 Excel Solver, *II:70*
 interactions with MATLAB, *III:448*
 macros in, *III:449, III:454–455*
 notations in, *III:477n*
 random number generation in,
 III:645–646
 random walks with, *I:83, I:85, I:87,*
 I:90
 @RISK in, *II:12f*
 syntax for functions in, *III:456*
- Exchange-rate intervention, study on,**
 III:177–178
- Exercise prices, I:452, I:484, I:508**
- Expectation maximization (EM)**
 algorithm, *II:146, II:165*
- Expectations, conditional, I:122,**
 II:517–518, III:508–509
- Expectations hypothesis, III:568–569,**
 III:601n
- Expected shortfall (ES), I:385–386,**
 III:332. See also average value at
 risk (AVaR)
- Expected tail loss (ETL), III:291,**
 III:293f, III:345–347, III:347f,
 III:355–356
- Expected value (EV), I:511**
- Expenses, noncash, II:25**
- Experiments, possibility of, II:307**
- Explicit costs, defined, III:623**
- Explicit Euler (EE) scheme, II:674,**
 II:677–678
- Exponential density function, III:218f**

- Exponential distribution, *III*:217–219
 applications in finance, *III*:219
- Exponentially weighted moving averages (EWMA)
 discussion of, *III*:409–413
 forecasting model of, *III*:411
 properties of the estimates, *III*:410–411
 standard errors for, *III*:411–412
 statistical methodology in, *III*:409
 usefulness of, *III*:413–414
 volatility estimates for, *III*:410*f*
- Exposures
 calculation of, *II*:247*t*
 correlation between, *II*:186
 distribution of, *II*:250*f*, *II*:251*f*, *II*:254
 management of, *II*:182–183
 monitoring of portfolio, *II*:249–250
 name-specific, *II*:188
- Extrema, characterization of local, *I*:23
- Extremal random variables, *III*:267
- Extreme value distributions, generalized, *III*:269
- Extreme value theory (EVT), *II*:744–746, *III*:95, *III*:228
 defined, *III*:238
 for IID processes, *III*:265–274
 in IID sequences, *III*:275
 role of in modeling, *II*:753*n*
- Factor analysis
 application of, *II*:165
 based on information coefficients, *II*:222
 defined, *II*:141, *II*:169
 discussion of, *II*:164–166
 importance of, *II*:238
vs. principal component analysis, *II*:166–168
- Factor-based strategies
vs. risk models, *II*:236
- Factor-based trading, *II*:196–197
 model construction for, *II*:228–235
 performance evaluation of, *II*:225–228
- Factor exposures, *II*:247–248, *II*:275–283
- Factorials, computing of, *III*:456
- Factorization, defined, *II*:307
- Factor mimicking portfolio (FMP), *II*:214
- Factor model estimation, *II*:142–147, *II*:150
 alternative approaches and extensions, *II*:145–147
 applied to bond returns, *II*:144–145
 computational procedure for, *II*:142–144
 fixed N, *II*:143
 large N, *II*:143–144
- Factor models
 in the Black-Litterman framework, *I*:200
 commonly used, *II*:150
 considerations in, *II*:178
 cross-sectional, *II*:220–221
 defined, *II*:153
 fixed income, *II*:271–272
 in forecasting, *II*:230–231
 linear, *II*:154–156, *II*:168
 normal, *II*:156
 predictive, *II*:142
 static/dynamic, *II*:146–147, *II*:155
 in statistical methodology, *II*:141
 strict, *II*:155–156
 types of, *II*:138–142
 usefulness of, *II*:154, *II*:503
 use of, *I*:354, *II*:137, *II*:150, *II*:168, *II*:219–225
- Factor portfolios, *II*:224–225
- Factor premiums, cross-sectional methods for evaluation of, *II*:214–219
- Factor returns, *II*:191*t*, *II*:192*t*
 calculation of, *II*:248
- Factor risk models, *II*:113, *II*:119
- Factors
 adjustment of, *II*:205–206
 analysis of data of, *II*:206–211
 categories of, *II*:197
 choice of, *II*:232–235
 defined, *II*:196, *II*:211
 desirable properties of, *II*:200
 development of, *II*:198
 estimation of types of, *II*:156
 graph of, *II*:166*f*
 known, *II*:138–139
 K systematic, *II*:138–139
 latent, *II*:140–141, *II*:150
 loadings of, *II*:144, *II*:145*t*, *II*:155, *II*:166*t*, *II*:167*f*, *II*:168*t*
 market, *II*:176
 orthogonalization of, *II*:205–206
 relationship to time series, *II*:168*f*
 sorting of, *II*:215
 sources for, *II*:200–201
 statistical, *II*:197
 summary of well-known, *II*:196*t*
 transformations applied to, *II*:206
 use of multiple, *II*:141–142
- Failures, probability of, *II*:726–727
- Fair equilibrium, between multiple accounts, *II*:76
- Fair value
 determination of, *III*:584–585
 Fair value, assessment of, *II*:6–7
- Fama, Eugene, *II*:468, *II*:473–474
- Fama-French three-factor model, *II*:139–140, *II*:177
- Fama-MacBeth regression, *II*:220–221, *II*:224, *II*:227–228, *II*:228*f*, *II*:237, *II*:240*n*
- Fannie Mae/Freddie Mac, writedowns of, *III*:77*n*
- Fast Fourier transform algorithm, *II*:743
- Fat tails
 of asset return distributions, *III*:242
 in chaotic systems, *II*:653
 class \mathcal{L} , *III*:261–263
 comparison between risk models, *II*:749–750
 effects of, *II*:354
 importance of, *II*:524
 properties of, *III*:260–261
 in Student's *t* distribution, *II*:734
- Favorable selection, *III*:76–77
- F*-distribution, *III*:216–217
- Federal Reserve
 effects of on inflation risk premium, *I*:281
 study by Cleveland Bank, *III*:177–178
 timing of interventions of, *III*:178
- Feynman-Kac formulas, *II*:661
- FFAs (freight forward agreements), *I*:566
- Filtered probability spaces, *I*:314–315, *I*:334*n*
- Filtration, *II*:516–517, *III*:476–477, *III*:489–490, *III*:508
- Finance, three major revolutions in, *III*:350
- Finance companies, captive, *I*:366–369
- Finance theory
 development of, *II*:467–468
 effect of computers on, *II*:476
 in the nineteenth century, *II*:468–469, *II*:476
 in the 1960s, *II*:476
 in the 1970s, *II*:476
 stochastic laws in, *III*:472
 in the twentieth century, *II*:476
- Financial assets, price distribution of, *III*:349–350
- Financial crisis (2008), *III*:71
- Financial date, pro forma, *II*:542–543
- Financial distress, defined, *I*:351
- Financial institutions, model risk of, *II*:693
- Financial leverage ratios, *II*:559–561, *II*:563
- Financial modelers, mistakes of, *II*:707–710

- Financial planning, *III*:126–127, *III*:128, *III*:129
- Financial ratios, *II*:546, *II*:563–564
- Financial statements
 assumptions used in creating, *II*:532
 data in, *II*:563
 information in, *II*:533–542, *II*:543
 pro forma, *II*:22–23
 time statements for, *II*:532
 usefulness of, *II*:531
 use of, *II*:204–205, *II*:246
- Financial time series, *I*:79–80, *I*:386–387, *II*:415–416, *II*:503–504
- Financial variables, modeling of, *III*:280
- Find, in MATLAB, *III*:422
- Finite difference methods, *II*:648–652, *II*:656–657, *II*:665–666, *II*:674–675, *II*:676–677, *III*:19
- Finite element methods, *II*:669–670, *II*:672, *II*:679–681
- Finite element space, *II*:670–672
- Finite life general DDM, *II*:5–7
- Finite states, assumption of, *I*:100–101
- Firms
 assessment of, *II*:546–547
 and capital structure, *II*:473
 characteristics of, *II*:94, *II*:176–177, *II*:201
 clientele of, *II*:36
 comparable, *II*:34, *II*:35–36
 geographic location of, *II*:36
 history *vs.* future prospects, *II*:92
 phases of, *II*:9–10
 retained earnings of, *II*:20
 valuation of, *II*:26–27, *II*:473
 value of, *II*:27–31, *II*:39
vs. characteristics of group, *II*:90–91
- First boundary problem, *II*:655–656, *II*:657f
- First Interstate Bancorp, *I*:304
 analysis of credit spreads, *I*:305t
 debt ratings of, *I*:410
- First passage models (FPMs), *I*:342, *I*:344–348
- Fischer-Tippett theorem, *III*:266–267
- Fisher, Ronald, *I*:140
- Fisherian, defined, *I*:140
- Fisher's information matrix, *I*:160n
- Fisher's law, *II*:322–323
- Fixed-asset turnover ratio, *II*:558
- Fixed-charge coverage ratio, *II*:560–561
- Flesaker-Hughston (FH) model, *III*:548–549
- Flows, discrete, *I*:448–453
- FMP (factor mimicking portfolio), *II*:214
- Footnotes, in financial statements, *II*:541–542
- Ford Motor Company, *I*:408f, *I*:409f
- Forecastability, *II*:132
- Forecastability, concept of, *II*:123
- Forecast encompassing
 defined, *II*:230–231
- Forecasts
 of bid-ask spreads, *II*:456–457
 comparisons of, *II*:420–421
 contingency tables, *II*:429t
 development of, *II*:110–114
 directional, *II*:428
 effect on future of, *II*:122–123
 errors in, *II*:422f
 evaluation of, *II*:428–430, *III*:368–370
 machine-learning approach to, *II*:128
 measures of, *II*:429–430, *II*:430
 need for, *II*:110–111
 in neural networks, *II*:419–420
 one-step ahead, *II*:421f
 parametric bootstraps for, *II*:428–430
 response to macroeconomic shocks, *II*:55f
 usefulness of, *II*:131–132
 use of models for, *II*:302
 of volatility, *III*:412
- Foreclosures, *III*:31, *III*:75
- Forward contracts
 advantages of, *I*:430
 buying assets of, *I*:439
 defined, *I*:426, *I*:478
 equivalence to futures prices, *I*:432–433
 hedging with, *I*:429, *I*:429t
 as OTC instruments, *I*:479
 prepaid, *I*:428
 price paths of, *I*:428t
 short *vs.* long, *I*:437–438, *I*:438f
 valuing of, *I*:426–430
vs. futures, *I*:430–431, *I*:433
vs. options, *I*:437–439
- Forward curves
 graph of, *I*:434f
 modeling of, *I*:533, *I*:557–558, *I*:564–565
 normal *vs.* inverted, *I*:434
 of physical commodities, *I*:555
- Forward freight agreements (FFAs), *I*:555, *I*:558, *I*:566
- Forward measure, use of, *I*:543–544
- Forward rates
 calculation of, *I*:491, *III*:572
 defined, *I*:509–510
 from discount function, *III*:566–567
 implied, *III*:565–567
 models of, *III*:543–544
 from spot yields, *III*:566
 of term structure, *III*:586
- Fourier integrals, *II*:656
- Fourier methods, *I*:559–560
- Fourier transform, *III*:265
- FPMs (first passage models), *I*:342, *I*:344–348
- Fractals, *II*:653–654, *III*:278–280, *III*:479–480
- Franklin Tempelton Investment Funds, *II*:496t, *II*:497t, *II*:498t
- Fréchet distribution, *II*:754n, *III*:228, *III*:230, *III*:265, *III*:267, *III*:268
- Fréchet-Hoeffding copulas, *I*:327, *I*:329
- Freddie Mac, *II*:77n, *II*:754n, *III*:49
- Free cash flow (FCF), *II*:21–23
 analysis of, *II*:570–571
 calculation of, *II*:23–24, *II*:571–572
 defined, *II*:569–571, *II*:578
 expected for XYZ, Inc., *II*:30t
 financial adjustments to, *II*:25–26
 statement of, direct method, *II*:24–25, *II*:24t
 statement of, indirect method, *II*:24–25, *II*:24t
vs. cash flow, *II*:22–23
- Freedman-Diaconis rule, *II*:494, *II*:495, *II*:497
- Frequencies
 accumulating, *II*:491–492
 distributions of, *II*:488–491, *II*:499f
 empirical cumulative, *II*:492
 formal presentation of, *II*:491
- Frequentist, *I*:140, *I*:148
- Frictions, costs of, *II*:472–473
- Friedman, Milton, *I*:123
- Frontiers, true, estimated and actual efficient, *I*:190–191
- F_SCORE, use of, *II*:230–231
- F-test, *II*:336, *II*:337, *II*:344, *II*:425, *II*:426
- FTSE 100, volatility in, *III*:412–413
- Fuel costs, *I*:561, *I*:562–563. *See also* energy
- Full disclosure, defined, *II*:532
- Functional, defined, *I*:24
- Functional-coefficient autoregressive (FAR) model, *II*:417
- Functions
 affine, *I*:31
 Archimedean, *I*:329, *I*:330–331, *I*:331
 Bessel, of the third kind, *II*:591
 beta, *II*:591
 characteristic, *II*:591–592, *II*:593
 choosing and calibrating of, *I*:331–333
 Clayton, Frank, Gumbel, and Product, *I*:329

- Functions (*Continued*)
 continuous, *II:581–584, II:582f, II:583, II:592–593*
 continuous / discontinuous, *II:582f*
 convex, *I:24–27, I:25, I:25f, I:26f*
 convex quadratic, *I:26, I:31f*
 copula, *I:320, I:325–333, I:407–408*
 for default times, *I:329–331*
 defined, *I:24, I:333*
 density, *I:141*
 with derivatives, *II:585f*
 elementary, *III:474*
 elliptical, *I:328–329*
 empirical distribution, *III:270*
 factorial, *II:590–591*
 gamma, *II:591, II:591f, III:212*
 gradients of, *I:23*
 Heaviside, *II:418–419*
 hypergeometric, *III:256, III:257*
 indicator, *II:584–585, II:584f, II:593*
 likelihood function, *I:141–143, I:143f, I:144f, I:148, I:176, I:177*
 measurable, *III:159–160, III:160f, III:201*
 minimization and maximization of values, *I:22, I:22f*
 monotonically increasing, *II:587–588, II:588f*
 nonconvex quadratic, *I:26–27*
 nondecreasing, *III:154–155, III:155f*
 normal density, *III:226f*
 optimization of, *I:24*
 parameters of copulas, *I:331–332*
 properties of quasi-convex, *I:28*
 quasi-concave, *I:27–28, I:27f*
 right-continuous, *III:154–155, III:155f*
 surface of linear, *I:33f*
 with two local maxima, *I:23f*
 usefulness of, *I:411–412*
 utility, *I:4–5, I:14–15, I:461*
 Fund management, art of, *I:273*
 Fund separation theorems, *I:36*
 Futures
 Eurodollar, *I:503*
 hedging with, *I:433*
 market for housing, *II:396–397*
 prices of, and interest rates, *I:435n*
 telescoping positions of, *I:431–432*
 theoretical, *I:487*
 valuing of, *I:430–433*
 vs. forward contracts, *I:430–431*
 Futures contracts
 defined, *I:478*
 determining price of, *I:481*
 pricing model for, *I:479–481*
 theoretical price of, *I:481–484*
 vs. forward contracts, *I:433, I:478–479*
 Futures options, defined, *I:453*
 Future value, *II:618*
 determining of money, *II:596–600*
 Galerkin methods, principle of, *II:671*
 Gamma, *I:509, I:518–520*
 Gamma process, *III:498*
 Gamma profile, *I:519f*
 Gapping effect, *I:509*
 GARCH (generalized autoregressive conditional heteroskedastic) models
 asymmetric, *II:367–368*
 exponential (EGARCH), *II:367–368*
 extensions of, *III:657*
 factor models, *II:372*
 GARCH-M (GARCH in mean), *II:368*
 Markov-switching, *I:180–184*
 time aggregation in, *II:369–370*
 type of, *II:131*
 usefulness of, *III:414*
 use of, *I:175–176, I:185–186, II:371, II:733–734, III:388*
 and volatility, *I:179*
 weights in, *II:363–364*
 GARCH (1,1) model
 Bayesian estimation of, *I:176–180*
 defined, *II:364*
 results from, *II:366, II:366f*
 skewness of, *III:390–391*
 strengths of, *III:388–389*
 Student's *t*, *I:182*
 use of, *I:550–551, III:656–657*
 GARCH (1,1) process, *I:551t*
 Garman-Kohlhagen system, *I:510–511, I:522*
 Gaussian density, *III:98f*
 Gaussian model, *III:547–548*
 Gaussian processes, *III:280, III:504*
 Gaussian variables, and Brownian motion, *III:480–481*
 Gauss-Markov theorem, *II:314*
 GBM (geometric Brownian motion), *I:95, I:97*
 GDP (gross domestic product), *I:278, I:282, II:138, II:140*
 General inverse Gaussian (GIG) distribution, *II:523–524*
 Generalized autoregressive conditional heteroskedastic (GARCH) models. *See* GARCH (generalized autoregressive conditional heteroskedastic) models
 Generalized central limit theorem, *III:237, III:239*
 Generalized extreme value (GEV) distribution, *II:745, III:228–230, III:272–273*
 Generalized inverse Gaussian distribution, use of, *II:521–522*
 Generalized least squares (GLS), *I:198–199, II:328*
 Generalized tempered stable (GTS) processes, *III:512*
 Generally accepted accounting principles (GAAP), *II:21–22, II:531–532, II:542–543*
 Geometric mean reversion (GMR) model, *I:91–92*
 computation of, *I:91*
 Gibbs sampler, *I:172n, I:179, I:184–185*
 GIG models, calibration of, *II:526–527*
 Gini index of dissimilarity (Gini measure), *III:353–354*
 Ginnie Mae/Fannie Mae/Freddie Mac, actions of, *III:49*
 Girsanov's theorem
 and Black-Scholes option pricing formula, *I:132–133*
 with Brownian motion, *III:511*
 and equivalent martingale measures, *I:130–133*
 use of, *I:263, III:517*
 Glivenko-Cantelli theorem, *III:270, III:272, III:348n, III:646*
 Global Economy Workshop, Santa Fe Institute, *II:699*
 Global Industry Classification Standard (GICS®), *II:36–37, II:248*
 Global minimum variance (GMV) portfolios, *I:39*
 GMR (geometric mean reversion) model, *I:91–92*
 GMV (global minimum variance) portfolios, *I:15, I:194–195*
 GNP, growth rate of (1947–1991), *II:410–411, II:410f*
 Gradient methods, use of, *II:684*
 Granger causality, *II:395–396*
 Graphs, in MATLAB, *III:428–433*
 Greeks, the, *I:516–522*
 beta and omega, *I:522*
 delta, *I:516–518*
 gamma, *I:518–520*
 rho, *I:521–522*
 theta, *I:509, I:520–521*
 use of, *I:559, II:660, III:643–644*
 vega, *I:521*
 Greenspan, Alan, *I:140–141*
 Growth, *I:283f, II:239, II:597–598, II:601–602*
 Gumbel distribution, *III:265, III:267, III:268–269*

- Hamilton-Jacobi equations, *II:675*
- Hankel matrices, *II:512*
- Hansen-Jagannathan bound, *I:59, I:61–62*
- Harrison, Michael, *II:476*
- Hazard, defined, *III:85*
- Hazard (failure) rate, calculation of, *III:94–95*
- Heat diffusion equation, *II:470*
- Heath-Jarrow-Morton framework, *I:503, I:557*
- Heavy tails, *III:227, III:382*
- Hedge funds, and probit regression model, *II:349–350*
- Hedge ratios, *I:416–417, I:509*
- Hedges
 - importance of, *I:300*
 - improvement using DTS, *I:398*
 - in the Merton context, *I:409*
 - rebalancing of, *I:519*
 - risk-free, *I:532f*
- Hedge test, *I:409, I:411*
- Hedging
 - costs of, *I:514, II:725*
 - and credit default swaps, *I:413–414*
 - determining, *I:303–304*
 - with forward contracts, *I:429, I:429t*
 - of fuel costs, *I:561*
 - with futures, *I:433*
 - gamma, *I:519*
 - portfolio-level, *I:412–413*
 - of positions, *II:724–726*
 - ratio for, *II:725*
 - with swaps, *I:434–435*
 - transaction-level, *I:412*
 - usefulness of, *I:418*
 - use of, *I:125–126*
 - using macroeconomic indices, *I:414–417*
- Hessian matrix, *I:23–24, I:25, I:186n, III:645*
- Heston model, *I:547, I:548, I:552, II:682*
 - with change of time, *III:522*
- Heteroskedasticity, *II:220, II:359, II:360, II:403*
- HFD (high-frequency data). *See* high-frequency data (HFD)
- Higham's projection algorithm, *II:446*
- High-dimensional problems, *II:673*
- High-frequency data (HFD)
 - and bid-ask bounce, *II:454–457*
 - defined, *II:449–450*
 - generalizations to, *II:368–370*
 - Level I, *II:451–452, II:452f, II:453t*
 - Level II, *II:451*
 - properties of, *II:451, II:453t*
 - recording of, *II:450–451*
 - time intervals of, *II:457–462*
 - use of, *II:300, II:481*
 - volume of, *II:451–454*
- Hilbert spaces, *II:683*
- Hill estimator, *II:747, III:273–274*
- Historical method
 - drawbacks of, *III:413*
 - weighting of data in, *III:397–398*
- Hit rate, calculation of, *II:240n*
- HJM framework, *I:498*
- HJM methodology, *I:496–497*
- Holding period return, *I:6*
- Ho-Lee model
 - continuous variant for, *I:497*
 - defined, *I:492*
 - in history, *I:493*
 - interest rate lattice, *III:614f*
 - as short rate model, *III:23*
 - for short rates, *III:605*
 - as single factor model, *III:549*
- Home equity prepayment (HEP) curve, *III:55–56, III:56f*
- Homeowners, refinancing behavior of, *III:25*
- Home prices, *I:412, II:397f, II:399t, III:74–75*
- Homoskedasticity, *II:360, II:373*
- Horizon prices, *III:598*
- Housing, *II:396–399, III:48*
- Howard algorithm (policy iteration algorithm), *II:676–677, II:680*
- Hull-White (HW) models
 - binomial lattice, *III:610–611*
 - for calibration, *II:681*
 - defined, *I:492*
 - interest rate lattice, *III:614f*
 - and short rates, *III:545–546*
 - for short rates, *III:605*
 - trinomial lattice, *III:613, III:616f*
 - usefulness of, *I:503*
 - use of, *III:557, III:604*
 - valuing zero-coupon bond calls with, *I:500*
- Hume, David, *I:140*
- Hurst, Harold, *II:714*
- Hypercubes, use of, *III:648*
- IBM stock, log returns of, *II:407f*
- Ignorance, prior, *I:153–154*
- Implementation risk, *II:694*
- Implementation shortfall approach, *III:627*
- Implicit costs, *III:631*
- Implicit Euler (IE) scheme, *II:674, II:677–678*
- Implied forward rates, *III:565–567*
- Impurity, measures of, *II:377*
- Income, defined for public corporation, *II:21–22*
- Income statements
 - common-size, *II:562–563, II:562t*
 - defined, *II:536*
 - in financial statements, *II:536–537*
 - sample, *II:537t, II:547t*
 - structure of, *II:536*
 - XYZ Inc. (example), *II:28t*
- Income taxes. *See* taxes
- Independence, *I:372–373, II:624–625, III:363–364, III:368*
- Independence function, in VaR models, *III:365–366*
- Independently and identically distributed (IDD) concept, *I:164, I:171, II:127, III:274–280, III:367, III:414*
- Indexes
 - characteristics of efficient, *I:42t*
 - defined, *II:67*
 - of dissimilarity, *III:353–354*
 - equity, *I:15t, II:190t, II:262–263*
 - tail, *II:740–741, II:740f, III:234*
 - tracking of, *II:64, II:180*
 - use of weighted market cap, *I:38*
 - value weighted, *I:76–77*
 - volatility, *III:550–552, III:552f*
- Index returns, scenarios of, *II:190t, II:191t*
- Indifference curves, *I:4–5, I:5f, I:14*
- Industries, characteristics of, *II:36–37, II:39–40*
- Inference, *I:155–158, I:169t*
- Inflation
 - effect on after-tax real returns, *I:286–287*
 - and GDP growth, *I:282*
 - indexing for, *I:278–279*
 - in regression analysis, *II:323*
 - risk of, *II:282*
 - risk premiums for, *I:280–283*
 - seasonal factors in, *I:292*
 - shifts in, *I:285f*
 - volatility of, *I:281*
- Information
 - anticipation of, *III:476*
 - from arrays in MATLAB, *III:421*
 - completeness of, *I:353–354*
 - contained in high volatility stocks, *III:629*
 - and filtration, *III:517*
 - found in data, *II:486*
 - and information propagation, *II:515*
 - insufficient, *III:44*
 - integration of, *II:481–482*
 - overload of, *II:481*
 - prior in Bayesian analysis, *I:151–155, I:152*
 - propagation of, *I:104*

- Information (*Continued*)
 structures of, I:106f, II:515–517
 unstructured *vs.* semistructured,
 II:481–482
- Information coefficients (ICs), II:98–99,
 II:221–223, II:223f, II:227f, II:234
- Information ratios
 defined, II:86n, II:115, II:119, II:237
 determining, II:100f
 for portfolio sorts, II:219
 use of, II:99–100
- Information sets, II:123
- Information structures
 defined, II:518
- Information technology, role of,
 II:480–481
- Ingersoll models, I:271–273, I:275f
- Initial conditions, fixing of, II:502
- Initial margins, I:478
- Initial value problems, II:639
- Inner quartile range (IQR), II:494
- Innovations, II:126
- Insurance, credit, I:413–414
- Integrals, II:588–590, II:593. *See also*
 stochastic integrals
- Integrated series, and trends,
 II:512–514
- Integration, stochastic, III:472, III:473,
 III:483
- Intelligence, general, II:154
- Intensity-based frameworks, and the
 Poisson process, I:315
- Interarrival time, III:219, III:225
- Intercepts, treatment of, II:334–335
- Interest
 accumulated, II:604–605, II:604f
 annual *vs.* quarterly compounding,
 II:599f
 compound, II:597, II:597f
 computing accrued, and clean price,
 I:214–215
 coverage ratio, II:560
 defined, II:596
 determining unknown rates,
 II:601–602
 effective annual rate (EAR),
 II:616–617
 mortgage, II:398
 simple *vs.* compound, II:596
 terms of, II:619
 from TIPS, I:277
- Interest rate models
 binomial, III:173–174, III:174f
 classes of, III:600
 confusions about, III:600
 importance of, III:600
 properties of lattices, III:610
 realistic, arbitrage-free, III:599
 risk-neutral/arbitrage-free, III:597
- Interest rate paths, III:6–9, III:7, III:8t
- Interest rate risk, III:12–14
- Interest rates
 absolute *vs.* relative changes in,
 III:533–534
 approaches in determining future,
 III:591
 binomial model of, III:173–174
 binomial trees, I:236, I:236f, I:237f,
 I:240f, I:244, I:244f, III:174f
 borrowing *vs.* lending, I:482–483
 calculation of, II:613–618
 calibration of, I:495
 caps/caplets of, III:589–590
 caps on, I:248–249
 categories of term structure, III:561
 computing sensitivities, III:22–23
 continuous, I:428, I:439–488
 derivatives of, III:589–590
 determination of appropriate,
 I:210–211
 distribution of, III:538–539
 dynamic of process, I:262
 effect of, I:514–515
 effect of shocks, III:23
 effect on putable bonds, III:303–304
 future course of, III:567, III:573
 and futures prices, I:435n
 importance of models, III:600
 jumps of, III:539–541
 jumpy and continuous, III:539f
 long *vs.* short, III:538
 market spot/forward, I:495t
 mean reversion of, III:7
 modeling of, I:261–265, I:267, I:318,
 I:491, I:503, III:212–213
 multiple, II:599–600
 negative, III:538
 nominal, II:615–616
 and option prices, I:486–487
 and prepayment risk, III:48
 risk-free, I:442
 shocks/shifts to, III:585–596
 short-rate, I:491–494, III:595
 simulation of, III:541
 stochastic, I:344, I:346
 structures of, III:573, III:576
 use of for control, I:489
 volatility of, III:405, III:533
- Intermarket relations, no-arbitrage,
 I:453–455
- Internal consistency rule, in OAS
 analysis, I:265
- Internal rate of return (IRR), II:617–618
 in MBSs, III:36
- International Monetary Fund
 Global Stability Report, I:299
- International Swap and Derivatives
 Association (ISDA). *See* ISDA
- Interpolated spread (I-spread), I:227
- Interrate relationship, arbitrage-free,
 III:544
- Intertemporal dependence, and risk,
 III:351
- Intertrade duration, II:460–461,
 II:462t
- Intertrade intervals, II:460–461
- Intervals, credible, I:170
- Interval scales, data on, II:487
- Intrinsic value, I:441, I:511, I:513,
 II:16–17
- Invariance property, III:328–329
- Inventory, II:542, II:557
- Inverse Gaussian process, III:499
- Investment, goals of, II:114–115
- Investment management, III:146
- Investment processes
 activities of integrated, II:61
 evaluation of results of, II:117–118
 model creation, II:96
 monitoring of performance, II:104
 quantitative, II:95, II:95f
 quantitative equity, II:95f, II:96f,
 II:105
 research, II:95–102
 sell-structured, II:108
 steps for equity investment, II:119
 testing of, II:109
- Investment risk measures, III:350–351
- Investments, I:77–78n, II:50–51,
 II:617–618
- Investment strategies, II:66–67,
 II:198
- Investment styles, quantamental,
 II:93–94, II:93f
- Investors
 behavior of, II:207, II:504
 comfort with risk, I:193
 completeness of information of,
 I:353–354
 focus of, I:299, II:90–91
 fundamental *vs.* quantitative,
 II:90–94, II:91f, II:92f, II:105
 goals/objectives of, II:114–115,
 II:179, III:631
 individual accounts of, II:74
 monotonic preferences of, I:57
 number of stocks considered, II:91
 preferences of, I:5, I:260, II:48, II:56,
 II:92–93
 prior beliefs of, II:727
 real-world, II:132
 risk aversion of, II:82–83, II:729
 SL-CAPM assumptions about, I:66
 sophisticated of, II:108
 in uncertain markets, II:54
 views of, I:197–199
- Invisible hand, notion of, II:468–469

- ISDA (International Swap and Derivatives Association)
 Credit Derivative Definitions (1999), *I:230, I:528*
 Master Agreement, *I:538*
 organized auctions, *I:526–527*
 supplement definition, *I:230*
- I-spread (interpolated spread), *I:227*
- Ito, Kiyosi, *II:470*
- Ito definition, *III:486–487*
- Ito integrals, *I:122, III:475, III:481, III:490–491*
- Ito isometry, *III:475*
- Ito processes
 defined, *I:95*
 generic univariate, *I:125*
 and Girsanov's theorem, *I:131*
 under HJM methodology, *I:497*
 properties of, *III:487–488*
 and smooth maps, *III:493*
- Ito's formula, *I:126, III:488–489*
- Ito's lemma
 defined, *I:98*
 discussion of, *I:95–97*
 in estimation, *I:348*
 and the Heston model, *I:548*
- James-Stein shrinkage estimator, *I:194*
- Japan, credit crisis in, *I:417*
- Jarrow-Turnbull model, *I:307*
- Jarrow-Yu propensity model, *I:324–325*
- Jeffreys' prior, *I:153, I:160n, I:171–172*
- Jensen's inequality, *I:86, III:569*
- Jevons, Stanley, *II:468*
- Johansen-Juselius cointegration tests, *II:391–393, III:395*
- Joint jumps/defaults, *I:322–324*
- Joint survival probability, *I:323–324*
- Jordan diagonal blocks, *II:641–642*
- Jorion shrinkage estimator, *I:194, I:202*
- Jump-diffusion, *III:554–557, III:657*
- Jumps
 default, *I:322–324*
 diffusions, *I:559–560*
 downward, *I:347*
 idiosyncratic, *I:323*
 incorporation of, *I:93–94*
 in interest rates, *III:539–541*
 joint, *I:322–324*
 processes of, *III:496*
 pure processes, *III:497–501, III:506*
 size of, *III:540*
- Kalotay-Williams-Fabozzi (KWF) model, *III:604, III:606–607, III:615f*
- Kamakura Corporation, *I:301, I:307, I:308–309, I:310n*
- Kappa, *I:521*
- Karush-Kuhn-Tucker conditions (KKT conditions), *I:28–29*
- Kendall's tau, *I:327, I:332*
- Kernel regression, *II:403, II:412–413, II:415*
- Kernels, *II:412, II:413f, II:746*
- Kernel smoothers, *II:413*
- Keynes, John Maynard, *II:471*
- Key rate durations (KRD), *II:276, III:311–315, III:317*
- Key rates, *II:276, III:311*
- Kim-Rachev (KR) process, *III:512–513*
- KKT conditions (Karush-Kuhn-Tucker conditions), *I:28–29, I:31, I:32*
- KoBoL distribution, *III:257n*
- Kolmogorov extension theorem, *III:477–478*
- Kolmogorov-Smirnov (KS) test, *II:430, III:366, III:647*
- Kolmogorov equation, use of, *III:581*
- Kreps, David, *II:476*
- Krispy Kreme Doughnuts, *II:574–575, II:574f*
- Kronecker product, *I:172, I:173n*
- Kuiper test, *III:366*
- Kurtosis, *I:41, III:234*
- Lag operator L , *II:504–506, II:507, II:629–630*
- Lagrange multipliers, *I:28, I:29–31, I:30, I:32*
- Lag times, *II:387, III:31*
- Laplace transforms, *II:647–648*
- Last trades, price and size of, *II:450*
- Lattice frameworks
 bushy trees in, *I:265, I:266f*
 calibration of, *I:238–240*
 fair, *I:235*
 interest rate, *I:235–236, I:236–238*
 one-factor model, *I:236f*
 for pricing options, *I:487*
 usefulness of, *I:235*
 use of, *I:240, I:265–266, III:14*
 value at nodes, *I:237–238*
 1-year rates, *I:238f, I:239f*
- Law of iterated expectations, *I:110, I:122, II:308*
- Law of large numbers, *I:267, I:270n, III:263–264, III:275*
- Law of one α , *II:50*
- Law of one price (LOP), *I:52–55, I:99–100, I:102, I:119, I:260*
- LCS (liquidity cost score), *I:402*
 use of, *I:403*
- LDIs (liability-driven investments), *I:36*
- LD (loss on default), *I:370–371*
- Leases, in financial statements, *II:542*
- Least-square methods, *II:683–685*
- Leavens, D. H., *I:10*
- Legal loss data
 Cruz study, *III:113, III:115t*
 Lewis study, *III:117, III:117t*
- Lehman Brothers, bankruptcy of, *I:413*
- Level (parallel) effect, *II:145*
- Lévy-Khinchine formula, *III:253–254, III:257*
- Lévy measures, *III:254, III:254t*
- Lévy processes
 and Brownian motion, *III:504*
 in calibration, *II:682*
 change of measure for, *III:511–512*
 conditions for, *III:505*
 construction of, *III:506*
 from Girsanov's theorem, *III:511*
 and Poisson process, *III:496*
 as stochastic process, *III:505–506*
 as subordinators, *III:521*
 for tempered stable processes, *III:512–514, III:514t*
 and time change, *III:527*
- Lévy stable distribution, *III:242, III:339, III:382–386, III:392*
- LGD (loss given default), *I:366, I:370, I:371*
- Liabilities, *II:533, II:534–535, III:132*
- Liability-driven investments (LDIs), *I:36*
- Liability-hedging portfolios (LHPs), *I:36*
- LIBOR (London Interbank Offered Rate)
 and asset swaps, *I:227*
 changes in, by type, *III:539–540*
 curve of, *I:226*
 interest rate models, *I:494*
 market model of, *III:589*
 spread of, *I:530*
 in total return swaps, *I:541*
 use of in calibration, *III:7*
- Likelihood maximization, *I:176*
- Likelihood ratio statistic, *II:425*
- Limited liability rule, *I:363*
- Limit order books, use of, *III:625, III:632n*
- Lintner, John, *II:474*
- Lipschitz condition, *II:658n, III:489, III:490*
- Liquidation
 effect of, *II:186*
 procedures for, *I:350–351*
 process models for, *I:349–351*
 time of, *I:350*
 vs. default event, *I:349*
- Liquidity
 assumption of, *III:371*
 in backtesting, *II:235*
 changes in, *I:405*

- Liquidity (*Continued*)
 cost of, I:401
 creation of, III:624–625, III:631
 defined, III:372, III:380
 effect of, II:284
 estimating in crises, III:378–380
 in financial analysis, II:551–555
 and LCS, I:404
 and market costs, III:624
 measures of, II:554–555
 premiums on, I:294, I:307
 ratios for, II:555
 in risk modeling, II:693
 shortages in, I:347–348
 and TIPS, I:293, I:294
 and transaction costs, III:624–625
- Liquidity-at-risk (LAR), III:376–378
- Liquidity cost, III:373–374, III:375–376
- Liquidity cost score (LCS), I:402, I:403
- Liquidity preference hypothesis, III:570
- Liquidity ratios, II:563
- Liquidity risk, II:282, III:380
- Ljung-Box statistics, II:407, II:421, II:422, II:427–428
- LnMix models, calibration of, II:526–527
- Loading, standardization of, II:177
- Loan pools, III:8–9
- Loans
 amortization of, II:606–607, II:611–613
 amortization table for, II:612t
 delinquent, III:63
 fixed rate, fully amortized schedule, II:614t
 floating rate, II:613
 fully amortizing, II:611
 modified, III:32
 nonperforming, III:75
 notation for delinquent, III:45n
 recoverability of, III:31–32
 refinancing of, III:68–69
 repayment of, II:612f, II:613f
 term schedule, II:615t
- Loan-to-value ratios (LTVs), III:31–32, III:69, III:73, III:74–75
- Location parameters, I:160n, III:201–202
- Location-scale invariance property (Gaussian distribution), II:732
- Logarithmic Ornstein-Uhlenbeck (log-OU) processes, I:557–558
- Logarithmic returns, III:211–212, III:225
- Logistic distribution, II:350
- Logistic regression, I:307, I:308, I:310
- Logit regression models, II:349–350, II:350
- Log-Laplace transform, III:255–256
- Lognormal distribution, III:222–225, III:392
- Lognormal mixture (LnMix) distribution, II:524–525
- Lognormal variables, I:86
- Log returns, I:85–86, I:88
- London Interbank Offered Rate (LIBOR). *See* LIBOR
- Lookback options, I:114, III:24
- Lookback periods, III:402, III:407
- LOP (law of one price). *See* law of one price (LOP)
- Lorenz, Edward, II:653
- Loss distributions, conditional, III:340–341
- Losses. *See also* operational losses
 allocation of, III:32
 analysis of in backtesting, III:338
 collateral *vs.* tranche, III:36
 computation of, I:383
 defined, III:85
 estimation of cumulative, III:39–40
 expected, I:369–370, I:373–374
 expected *vs.* unexpected, I:369, I:375–376
 internal *vs.* external, III:83–84
 median of conditional, III:348n
 projected, III:37f
 restricting severity of, I:385–386
 severity of, III:44
 unexpected, I:371–372, I:374–375
- Loss functions, I:160n, III:369
- Loss given default (LGD), I:366, I:370, I:371
- Loss matrix analysis, III:40–41
- Loss on default (LD), I:370–371
- Loss severity, III:30–31, III:60–62, III:97–99
- Lottery tickets, I:462
- Lower partial moment risk measure, III:356
- Lundbert, Filip, II:467, II:470–471
- Macroeconomic influences, defined, II:197
- Magnitude measures, II:429–430
- Maintenance margins, I:478
- Major indexes, modeling return distributions for, III:388–392
- Malliavin calculus, III:644
- Management, active, II:115
- Mandelbrot, Benoit, II:653, II:738, III:234, III:241–242
- Manufactured housing prepayment (MHP) curve, III:56
- Marginalization, II:335
- Marginal rate of growth, III:197–198
- Marginal rate of substitution, I:60
- Margin calls, exposure to, III:377
- Market cap *vs.* firm value, II:39
- Market completeness, I:52, I:105
- Market efficiency, I:68–73, II:121, II:473–474
- Market equilibrium
 and investor's views, I:198–199
- Market impact
 costs of, III:623–624, III:627
 defined, II:69
 forecasting/modeling of, III:628–631
 forecasting models for, III:632
 forecasting of, III:628–629, III:629–631
 measurement of, III:626–628
 between multiple accounts, II:75–76
 in portfolio construction, II:116
 and transaction costs, II:70
- Market model regression, II:139
- Market opportunity, two state, I:460f
- Market portfolios, I:66–67, I:72–73
- Market prices, I:57, III:372
- Market risk
 approaches to estimation of, III:380
 in bonds, III:595
 in CAPM, I:68–69, II:474
 importance of, III:81
 models for, III:361–362
 premium for, I:203n, I:404
- Markets
 approach to segmented, II:48–51
 arbitrage-free, I:118
 complete, I:51–52, III:578
 complex, II:49
 effect of uncertainty in on bid-ask spreads, II:455–456
 efficiency of, II:15–16
 frictionless, I:261
 incomplete, I:461–462
 liquidity of, III:372
 models of, III:589
 for options and futures, I:453–454
 perfect, II:472
 properties of modern, III:575–576
 sensitivities to value-related variables, II:54t
 simple, I:70
 systematic fluctuations in, II:172–173
 unified approach to, II:49
 up/down, defined, II:347
- Market sectors, defined, III:560
- Market standards, I:257
- Market structure, and exposure, II:269–270
- Market timing, II:260
- Market transactions, upstairs, III:630–631, III:632n

- Market weights, *II:269t*
- Markov chain approximations, *II:678*
- Markov chain Monte Carlo (MCMC)
methods, *II:410f, II:417–418*
- Markov coefficients, *II:506–507, II:512*
- Markov matrix, *I:368*
- Markov models, *I:114*
- Markov processes
in dynamic term structures, *III:579*
hidden, *I:182*
use of, *III:509, III:517*
- Markov property, *I:82, I:180–181, I:183, II:661, III:193n*
- Markov switching (MS) models
discussion of, *I:180–184*
and fat tails, *III:277–278*
stationarity with, *III:275*
usefulness of, *II:433*
use of, *II:409–411, II:411t*
- Markowitz, Harry M., *I:38, I:140, II:467, II:471–472, III:137, III:351–352*
- Markowitz constraint sets, *I:69, I:72*
- Markowitz diversification, *I:10–11, I:11*
- Markowitz efficient frontiers, *I:191f*
- Markowitz model
in financial planning, *III:126*
- Mark-to-market (MTM)
calculation of value, *I:535–536, I:536t*
defined, *I:535*
and telescoping futures, *I:431–432*
- Marshall and Siegel, *II:694*
- Marshall-Olkin copula, *I:323–324, I:329*
- Martingale measures, equivalent
and arbitrage, *I:111–112, I:124*
and complete markets, *I:133*
defined, *I:110–111*
and Girsanov's theorem, *I:130–133*
and state prices, *I:133–134*
use of, *I:130–131*
working with, *I:135*
- Martingales
with change of time methods
(CTM), *III:522–523*
defined, *II:124, II:126, II:519*
development of concept, *II:469–470*
equivalent, *II:476*
measures of, *I:110–111*
use of conditions, *I:116*
use of in forward rates, *III:586*
- Mathematical theory, importance of
advances in, *III:145*
- Mathworks, website of, *III:418*
- MATLAB
array operations in, *III:420–421*
basic mathematical operations in,
III:419–420
- construction of vectors/matrices,
III:420
- control flow statements in,
III:427–428
- desktop, *III:419f*
- European call option pricing with,
III:444–445
- functions built into, *III:421–422*
- graphs in, *III:428–433, III:429–430f, III:431f*
- interactions with other software,
III:433–434
- M-files in, *III:418–419, III:423, III:447*
- operations in, *III:447*
- optimization in, *III:434–444, III:435t*
- Optimization Tool, *III:435–436, III:436f, III:440f, III:441f*
- overview of desktop and editor,
III:418–419
- quadprog function, *II:70*
- quadratic optimization with,
III:441–444
- random number generation,
III:444
- for simulations, *III:651*
- Sobol sequences in, *III:445–446*
- for stable distributions, *III:344*
- surf function in, *III:432–433*
- syntax of, *III:426–427*
- toolboxes in, *III:417–418*
- user-defined functions in,
III:423–427
- Matrices
augmented, *II:624*
characteristic polynomial of, *II:628*
coefficient, *II:624*
companion, *II:639–640*
defined, *II:622*
diagonal, *II:622–623, II:640*
eigenvalues of random, *II:704–705*
eigenvectors of, *II:640–641*
in MATLAB, *III:422, III:432*
operations on, *II:626–627*
ranks of, *II:623, II:628*
square, *II:622–623, II:626–627*
symmetric, *II:623*
traces of, *II:623*
transition, *III:32–33, III:32t, III:33t, III:35f*
types of, *II:622, II:628*
- Matrix differential equations, *III:492*
- Maturity value (lump sum), from
bonds, *I:211*
- Maxima, *III:265–269, III:266f*
- Maximum Description Length
principle, *II:703*
- Maximum eigenvalue test, *II:392–393*
- Maximum likelihood (ML)
approach, *I:141, I:348*
methods, *II:348–349, II:737–738, III:273*
principal, *II:312*
- Maximum principle, *II:662, II:667*
- Max-stable distributions, *III:269, III:339–340*
- MBA (Mortgage Bankers Association)
refi index, *III:70, III:70f*
- MBS (mortgage-backed securities),
I:258
agency vs. nonagency, *III:48*
cash flow characteristics of, *III:48*
default assumptions about, *III:8*
negative convexity of, *III:49*
performance of, *III:74*
prices of, *III:26*
projected long-term performance of,
III:34f
time-related factors in, *III:73–74*
valuation of, *III:62*
valuing of, *III:645*
- MBS (mortgage-backed securities),
nonagency
analysis of, *III:44–45*
defined, *III:48*
estimation of returns, *III:36–44*
evaluation of, *III:29*
factors impacting returns of,
III:30–32
yield tables for, *III:41t*
- Mean absolute deviation (MAD),
III:353
- Mean absolute moment (MAM(q)),
III:353
- Mean colog (M-colog), *III:354*
- Mean entropy (M-entropy), *III:354*
- Mean excess function, *II:746–747*
- Mean/first moment, *III:201–202*
- Mean residual life function, *II:754n*
- Mean reversion
discussion of, *I:88–92*
geometric, *I:91–92*
in HW models, *III:605*
and market stability, *III:537–538*
models of, *I:97*
parameter estimation, *I:90–91*
risk-neutral asset model, *III:526*
simulation of, *I:90*
in spot rate models, *III:580*
stabilization by, *III:538*
within a trinomial setting, *III:604*
- Mean-reverting asset model (MRAM),
III:525–526
- Means, *I:148, I:155, I:380, III:166–167*
- Mean-variance
efficiency, *I:190–191*
efficient portfolios, *I:13, I:68, I:69–70*

- Mean-variance (*Continued*)
 nonrobust formulation, III:139–140
 optimization, I:192
 constraints on, I:191
 estimation errors and, I:17–18
 practical problems in, I:190–194
 risk aversion formulation, II:70
 Mean variance analysis, I:3, I:15f,
 I:201, II:471–472, III:352
 Measurement levels, in descriptive
 statistics, II:486–487
 Media effects, III:70
 Median, I:155, I:159n, II:40
 Median tail loss (MTL), III:341
 Mencken, H. L., II:57
 Menger, Carl, II:468
 Mercurio-Moraleda model, I:493–494
 Merton, Robert, I:299, I:310, II:468,
 II:475, II:476
 Merton model
 advantages and criticisms of,
 I:344
 applied to probability of default,
 I:363–365
 with Black-Scholes approach,
 I:305–306
 default probabilities with, I:307–308
 discussion of, I:343–344
 drawbacks of, I:410
 with early default, I:306
 evidence on performance, I:308–309
 as first modern structural model,
 I:313, I:341
 in history, I:491
 with jumps in asset values, I:306
 portfolio-level hedging with,
 I:411–413
 with stochastic interest rates, I:306
 and transaction-level hedging,
 I:408–410
 usefulness of, I:410, I:411–412,
 I:417–418
 use of, I:304, I:305, I:510
 variations on, I:306–307
 Methodology, equally weighted,
 III:399
 Methods
 quantile, II:354–356
 Methods pathwise, III:643
 Metropolis-Hastings (M-H) algorithm,
 I:178
 M-H algorithm, I:179
 MIB 30, III:402–403, III:402f, III:403f
 Microsoft, II:722f. *See also* Excel
 Midsquare technique, III:647
 Migration mode
 calculation of expected/unexpected
 losses under, I:376t
 expected loss under, I:373–374
 Miller, Merton, II:467, II:473
 MiniMax (MM) risk measure, III:356
 Minimization problems, solutions to,
 II:683–684
 Minimum-overall-variance portfolio,
 I:69
 Minority interest, on the balance
 sheet, II:536
 Mispricing, risk of, II:691–692
 Model creep, II:694
 Model diagnosis, III:367–368
 Model estimation, in non-IDD
 framework, III:278
 Modeling
 calibration of structure, III:549–550
 changes in mathematical, II:480–481
 discrete *vs.* continuous time, III:562
 dynamic, II:105
 issues in, II:299
 nonlinear time series, II:427–428,
 II:430–433
 quantitative, II:481
 Modeling techniques
 non-parametric/nonlinear, II:375
 Model risk
 of agency ratings, II:728–729
 awareness of, I:145, II:695–696
 with computer models, II:695
 consequences of, II:729–730
 contribution to bond pricing,
 II:727–728
 defined, I:331, II:691, II:697
 discussion of, II:714–715
 diversification of, II:378
 endogenous, II:694–695, II:697
 in financial institutions, II:693
 guidelines for institutions,
 II:696–697
 management of, II:695–697, II:697
 misspecification of, II:199
 and robustness, II:301
 of simple portfolio, II:721–726
 sources of, II:692–695
 Models. *See also* operational risk
 models
 accuracy in, III:321
 adjustment, II:502
 advantages of reduced-form, I:533
 analytical tractability of, III:549–550
 APD, III:18, III:20–22, III:21f, III:26
 application of, II:694
 appropriate use of classes of,
 III:597–598
 arbitrage-free, III:600
 autopredictive, II:502
 averages across, II:715
 bilinear, II:403–404
 binomial, I:114–116, I:119
 binomial stochastic, II:10–11
 block maxima, II:745
 choosing, III:550–552
 comparison of, III:617
 compatibility of, III:373
 complexity of, II:704, II:717
 computer, I:511, II:695
 conditional normal, II:733–734
 conditional parametric fat-tailed,
 II:744
 conditioning, II:105
 construction of, II:232–235
 for continuous processes, I:123
 creation of, II:100–102
 cross-sectional, II:174–175, II:175t
 cumulative return of, II:234
 defined, II:691, II:697
 to describe default processes, I:313
 description and estimation of,
 II:256–257
 designing the next, III:590–591
 determining, II:299–300
 disclosure of, I:410
 documentation of, II:696
 dynamic factor, II:128, II:131,
 III:126–127
 dynamic term structure, III:591
 econometric, II:295, II:304
 equilibrium forms of, III:599–600
 equity risk, II:174, II:178–191, II:192
 error correction, II:381t, II:387–388,
 II:394–395
 evidence of performance, I:308–309,
 II:233
 examples of multifactor, II:139–140
 financial, I:139, II:479–480
 forecasting, II:112, II:303–304
 for forecasting, III:411
 formulation of, III:128–131
 fundamental factor, II:244, II:248
 generally, II:360–362
 Gordon-Shapiro, II:17–18
 Heath-Jarrow-Morton, III:586–587,
 III:589
 hidden-variable, II:128, II:131
 linear, II:264, II:310–311, II:348,
 II:507–508
 linear autoregressive, II:128,
 II:130–131
 linear regression, I:91, I:163–170,
 II:360, II:414–415
 liquidation process, I:342
 martingale, II:127–128, III:520–521
 MGARCH, II:371–372
 model-vetting procedure, II:696–697
 moving average, III:414
 multifactor, II:231–232, III:92
 multivariate extensions of,
 II:370–373
 no arbitrage, III:604

- nonlinear, *II*:402–421, *II*:417–418
 penalty functions in, *II*:703
 performance measurement of, *II*:301
 predictive regressive, *II*:130
 predictive return, *II*:128–131
 for pricing, *II*:127–128
 pricing errors in, *I*:322
 principals for engineering,
II:482–483
 probabilistic, *II*:299
 properties of good, *I*:320
 ranking alternative, *III*:368–370
 recalibration of, *II*:713–714
 reduced form default, *I*:310, *I*:313
 regressive, *II*:128, *II*:129–130
 relative valuation, *I*:260
 return forecasting, *II*:119
 returns of, *II*:233*t*
 robustness of, *II*:301
 selection of, *I*:145, *II*:298, *II*:692–693,
II:699–701
 short-rate, *I*:494
 single-index market, *II*:317–318
 static, *II*:297, *III*:573
 static regressive, *II*:129–130
 static *vs.* dynamic, *II*:295–296, *II*:304
 statistical, *II*:175, *II*:175*t*
 stochastic, *I*:557, *III*:124–125
 structural, *I*:305, *I*:313–314, *I*:341–342
 structural *vs.* reduced, *I*:532–533
 subordinated, *II*:742–743
 temporal aggregation of, *II*:369
 testing of, *II*:126–127, *II*:696–697
 time horizon of, *II*:300–301
 time-series, *II*:175, *II*:175*t*
 tree, *II*:381, *III*:22–23
 tuning of, *III*:580–581
 two-factor, *I*:494
 univariate regression, *I*:165
 usefulness of, *II*:122
 use of in practice, *I*:494–496, *III*:600*t*
- Models, lattice**
 binomial, *III*:610, *III*:610*f*
 Black-Karasinski (BK) lattice, *III*:611
 Hull White binomial, *III*:610–611
 Hull White trinomial, *III*:613
 trinomial, *III*:610, *III*:610*f*,
III:611–612
- Models, selection of**
 components of, *II*:717
 generally, *II*:715–717
 importance of, *II*:700
 machine learning approach to,
II:701–703, *II*:717
 uncertainty/noise in, *II*:716–717
 use of statistical tools in, *II*:230
- Modified Accelerated Cost Recovery
 System (MACRS), *II*:538**
- Modified Restructuring clause, *I*:529**
- Modified tempered stable (MTS)
 processes, *III*:513**
- Modigliani, Franco, *II*:467, *II*:473
- Modigliani-Miller theorem, *I*:343,
I:344, *II*:473, *II*:476
- Moment ratio estimators, *III*:274
- Moments**
 exponential, *III*:255–256
 first, *III*:201–202
 of higher order, *III*:202–205
 integration of, *II*:367–368
 raw, *II*:739
 second, *III*:202
 types of, *II*:125
- Momentum**
 formula for analysis of, *II*:239
 portfolios based on, *II*:181
- Momentum factor, *II*:226–227
- Money, future value of, *II*:596–600
- Money funds, European options on,
I:498–499
- Money markets, *I*:279, *I*:282, *I*:314,
II:244
- Monotonicity property, *III*:327
- Monte Carlo methods**
 advantages of, *II*:672
 approach to estimation, *I*:193
 defined, *I*:273
 examples of, *III*:637–639
 foundations of, *I*:377–378
 for interest rate structure, *I*:494
 main ideas of, *III*:637–642
 for nonlinear state-space modeling,
II:417–418
 stochastic content of, *I*:378
 usefulness of, *I*:389
 use of, *I*:266–268, *III*:651
 of VaR calculation, *III*:324–325
- Monte Carlo simulations**
 for credit loss, *I*:379–380
 effect of sampling process, *I*:384
 in fixed income valuation modeling,
III:6–12
 sequences in, *I*:378–379
 speed of, *III*:644
 use of, *III*:10–11, *III*:642
- Moody's diversity score, use of,
I:332
- Moody's Investors Service, *I*:362
- Moody's KMV, *I*:364–365
- Mortgage-backed securities (MBS). *See*
 MBS (mortgage-backed
 securities)
- Mortgage Bankers Association (MBA)
 method, *III*:57–58
- Mortgagee pools**
 composition of, *III*:52
 defined, *III*:23, *III*:65
 nonperforming loans and, *III*:75
- population of, *III*:19
 seasoning of, *III*:20, *III*:22
- Mortgages, *III*:48–49, *III*:65, *III*:69,
III:71
- Mosaic Company, distribution of price
 changes of, *II*:723*f*
- Mossin, Jan, *II*:468, *II*:474
- Moving averages, infinite, *II*:504–508
- MSCI Barra model, *II*:140
- MSCI EM, historical distributions of,
III:391*f*
- MSCI-Germany Index, *I*:143
- MSCI World Index, *I*:15–17
 analysis of 18 countries, *I*:16*t*
- MS GARCH model, *I*:185–186
 estimation of, *I*:182
 sampling algorithm for, *I*:184
- MSR (maximum Sharpe Ratio), *I*:36–37
- MS-VAR models, *II*:131
- Multiaccount optimization, *II*:75–77
- Multicollinearity, *II*:221
- Multilayer perceptrons, *II*:419
- Multinomial/polynomial coefficients,
III:191–192
- Multivariate normal distribution, in**
 MATLAB, *III*:432–433, *III*:433*f*
- Multivariate random walks, *II*:124
- Multivariate stationary series,
II:506–507
- Multivariate *t* distribution, loss
 simulation, *I*:388–389
- Nadaraya-Watson estimator, *II*:412,
II:415
- Natural conjugate priors, *I*:160*n*
- Navigation, fuel-efficient, *I*:562–563
- Near-misses, management of,
III:84–85
- Net cash flow, defined, *II*:541
- Net cost of carry, *I*:424–425, *I*:428,
I:437, *I*:439–440, *I*:455
- Net free cash flow (NFCF), *II*:572–574,
II:578
- Net profit margin, *II*:556
- Net working capital-to-sales ratio,
II:554–555
- Network investment models,
III:129–130, *III*:129*f*
- Neumann boundary condition, *II*:666,
II:671
- Neural networks, *II*:403, *II*:418–421,
II:418*f*, *II*:701–702
- Newey-West corrections, *II*:220
- NIG distribution, *III*:257*n*
- 9/11 attacks, effects of, *III*:402–403
- No-arbitrage condition, in certain
 economy, *III*:567–568
- No arbitrage models, use of, *III*:604
- No-arbitrage relations, *I*:423

- Noise
 continuous-time, III:486
 in financial models, II:721–722
 in model selection, II:716–717
 models for, II:726
 reduction of, II:51–52
- Noise, white
 defined, I:82, II:297
 qualities of, II:127
 sequences, II:312, II:313
 in stochastic differential equations, III:486
 strict, II:125
vs. colored noise, III:275
- Nonlinear additive AR (NAAR)
 model, II:417
- Nonlinear dynamics and chaos, II:645, II:652–654
- Nonlinearity, II:433
 in econometrics, II:401–403
 tests of, II:421–427
- Non-normal probability distributions, II:480
- Nonparametric methods, II:411–416
- Normal distributions, I:81, I:82*f*, I:177–178, III:638*f*
 and AVaR, III:334
 comparison with α -stable, III:234*f*
 fundamentals of, II:731–734
 inverse Gaussian, III:231–233, III:232*f*, III:233*f* (*See also* Gaussian distribution)
 likelihood function, I:142–143
 for logarithmic returns, III:211–212
 mixtures of for downside risk estimation, III:387–388
 for modeling operational risk, III:98–99
 multivariate, and tail dependence, I:387
 properties of, II:732–733, III:209–210
 relaxing assumption of, I:386–387
 standard, III:208
 standardized residuals from, II:751
 use of, II:752*n*
 using to approximate binomial distribution, III:211
 for various parameter values, III:209*f*
vs. normal inverse Gaussian distribution, III:232–233
- Normal mean, and posterior tradeoff, I:158–159
- Normal tempered stable (NTS)
 processes, III:513
- Normative theory, I:3
- Notes, step-up callable, I:251–252, I:251*f*, I:252*f*
- Novikov condition, I:131–132
- NTS distribution, III:257*n*
- Null hypothesis, I:157, I:170, III:362
- Numeraire, change of, III:588–589
- Numerical approximation, I:265
- Numerical models for bonds, I:273–275
- OAS (option-adjusted spread). *See* option-adjusted spread
- Obligations, deliverable, I:231, I:526
- Observations, frequency of, III:404
- Occam's razor, in model selection, II:696
- Odds ratio, posterior, I:157
- Office of Thrift Supervision (OTS)
 method, III:57–58
- Oil industry, free cash flows of, II:570
- OLS (ordinary least squares). *See* ordinary least squares (OLS)
- Open classes, II:493–494
- Operating cash flow (OCF), II:23
- Operating cycles, II:551–554
- Operating profit margin, II:556
- Operational loss data
 de Fontnouvelle, Rosengren, and Jordan study, III:116–117, III:116*t*
 empirical evidence with, III:112–118
 Moscadelli study, III:113, III:116, III:116*t*
 Müller study, III:113, III:114*f*, III:115*t*
 Reynolds-Syer study, III:117–118
 Rosenberg-Schuermann study, III:118
- Operational losses
 and bank size, III:83
 definitions of types, III:84*t*
 direct *vs.* indirect, III:84–85
 expected *vs.* unexpected, III:85
 histogram of, III:104*f*
 histogram of severity distribution, III:95*f*
 historical data on, III:96
 near-miss, III:84–85
 process of arriving at data, III:96–97
 process of occurrence, III:86*f*
 recording of, III:97
 severity of, III:104*f*
 time lags in, III:96–97
 types of, III:81, III:88
- Operational loss models
 approaches to, III:103–104
 assumptions in, III:104
 nonparametric approach, III:103–104, III:104–105, III:118
 parametric approach, III:104, III:105–110, III:118
 types of, III:118
- Operational risk
 classifications of, III:83–88, III:87–88, III:87*f*, III:88
 defined, III:81–83, III:88
 event types with descriptions, III:86*t*
 indicators of, III:83
 models of, III:91–96
 nature of, III:99
 and reputational risk, III:88
 sources of, III:82
- Operational risk/event/loss types, distinctions between, III:85–87
- Operational risk models
 actuarial (statistical) models, III:95
 bottom-up, III:92*f*, III:94–96, III:99
 causal, III:94
 expense-based, III:93
 income-based, III:93
 multifactor causal models, III:95
 operating leverage, III:93
 process-based, III:94–95
 proprietary, III:96
 reliability, III:94–95
 top down, III:92–94, III:99
 types of, III:91–92
- Operations
 addition, II:625, II:626
 defined, II:628
 inverse and adjoint, II:626–627
 multiplication, II:625–626, II:626
 transpose, II:625, II:626
 vector, II:625–626
- Operators in sets, defined, III:154
- Ophelimity, concept of, II:469
- Opportunity cost, I:435, I:438, I:439, II:596, III:623
- Optimal exercise, I:515–516
- Optimization
 algorithms for, III:124
 complexity of, II:82
 constrained, I:28–34
 defined, III:434–435
 local *vs.* global, II:378
 in MATLAB, III:434–444
 unconstrained, I:22–28
- Optimization theory, I:21
- Optimization Toolbox, in MATLAB, III:435–436, III:436*f*
- Optimizers, using, II:115–116, II:483
- Option-adjusted spread (OAS)
 calculation of, I:253–255
 defined, I:254, III:11
 demonstrated, I:254*f*
 determination of, I:259
 implementation of, I:257
 and market value, I:258
 results from example, III:617*t*
 and risk factors, III:599

- rules-of-thumb for analysis, I:264–265
- usefulness of, III:3
- values of, I:267, I:268
- variance between dealers, I:257–258
- Option premium, I:508–509
 - time/intrinsic values of, I:513
- Option premium profiles, I:512, I:512f
- Option prices
 - components of, I:484–485, I:511–512
 - factors influencing, I:486–487, I:486t, I:487–488, I:522–523
 - models for, I:490
- Options
 - American, II:664–665, II:669–670, II:674–679, II:679–681
 - American-style, I:444, I:454–455, I:490
 - Asian, II:663–664, II:668–669, II:642–643
 - on the average, II:663–664
 - barrier, II:662–663
 - basic properties of, I:507–508
 - basket, II:662, II:672
 - Bermudean, II:663–664, III:597
 - buying assets of, I:439
 - costs of, I:441–442, III:11–12
 - difference from forwards, I:437–439
 - early exercise of, I:442–443, I:447
 - Eurodollar, I:489
 - European, I:125, I:127–129, II:660–664, II:665–674
 - European-style, I:444–445, I:454
 - European-style *vs.* American-style, I:453t, I:455n, I:508, I:515–516
 - and expected volatility, I:486
 - expiration/maturity dates of, I:484
 - factors affecting value of, I:474
 - formulas for pricing, III:522, III:527
 - in/out of/at-the-money, I:485
 - long *vs.* short call, I:437–439, I:438f
 - lookback, II:663, II:672, II:673f
 - on the maximum, II:663
 - models of, I:510–511
 - no-arbitrage futures, I:453
 - price relations for, I:448t
 - pricing of, I:124–129, I:455t, I:484–488, I:507, III:408
 - theoretical valuation of, I:508–509
 - time premiums of, I:485
 - time to expiration of, I:486
 - types of, I:484
 - valuing of, I:252–253, III:639
 - vanilla, II:661, III:655
 - volatility of, I:488
- Orders
 - in differential equations, II:643, II:644–645
 - fleeting limit, III:625
 - limit, III:625, III:631
 - market, III:625, III:631
- Order statistics, III:269–270
 - bivariate, III:293–295
 - joint probability distributions for, III:291–292
 - use of, III:289
 - for VaR and ETL, III:292t
 - in VaR calculations, III:291
- Ordinary differential equations (ODE), II:644–645, II:646–648, II:648–652, II:649f
- Ordinary least squares (OLS)
 - alternate weighting of, II:438–439
 - estimation of factor loadings matrix with, II:165
 - in maximum likelihood estimates, II:313–314
 - pictorial representations of, II:437–438, II:438f
 - squared errors in, II:439–440
 - use of, I:165, I:172n, II:353
 - vs.* Theil-Sen estimates of beta, II:442f
 - vs.* Theil-Sen regression, II:441t
- Ornstein-Uhlenbeck process
 - with change of time, III:523
 - and mean reversion, I:263, I:264f
 - solutions to, III:492
 - use of, I:89, I:95
 - and volatility, III:656
- Outcomes, identification and
 - evaluation of worst-case, III:379–380
- Outliers
 - in data sets, II:200
 - detection and management of, II:206
 - effect of, II:355f, II:442–443
 - and market crashes, II:503
 - in OLS methods, II:354
 - in quantile methods, II:355–356
 - and the Thiel-Sen regression algorithm, II:440
- Out-of-sample methodology, II:238
- Pair trading, II:710
- P-almost surely (P-a.s.) occurring events, III:158
- Parallel yield curve shift assumption, III:12–13
- Parameters
 - calibration of, II:693
 - density functions for values, III:229f, III:230f, III:231f
 - distributions of, II:721
 - estimation of for random walk, I:83
 - robust estimation of, II:77–78
 - stable, III:246f
- Parametric methods, use of, II:522
- Parametric models, II:522–523, II:526–527
- Par asset swap spreads, I:530, I:531
- Par CDS spread, I:531
- Par-coupon curve, III:561
- Pareto, Vilfredo, II:467, II:468–469, II:474
- Pareto(2) distribution, II:441
- Pareto distributions
 - density function of, II:738
 - generalized (GPD), II:745–746, II:747, III:230–231
 - in loss distributions, III:108–109
 - parameters for determining, II:738
 - stable, II:738–741
 - stable/varying density, II:739f
 - tails of, II:751
- Pareto law, II:469
- Pareto-Lévy stable distribution, III:242
- Partial differential equations (PDEs)
 - for American options, II:664–665
 - equations for option pricing, II:660–665
 - framework for, I:261, I:265, II:675, III:555
 - pricing European options with, II:665–674
 - usefulness of, II:659–660
 - use of, III:18–19
- Partitioning, binary recursive, II:376–377, II:376f
- Paths
 - in Brownian motion, III:501, III:502f
 - dependence, III:18–19
 - stochastic, II:297
- Payments, I:229, II:611–612
- Payment shock, III:72
- Payoff-rate process, I:121–122
- Payoffs, III:466, III:638–639
- PCA (principal components analysis). *See* principal component analysis (PCA)
- Pearson skewness, III:204–205
- Pension funds, constraints of, II:62
- Pension plans, II:541, III:132
- P/E (price/earnings) ratio, II:20–21, II:38
- Percentage rates, annual *vs.* effective, II:615–617
- Percolation models, III:276
- Performance attribution, II:57, II:58, II:104, II:188–189, II:252–253, II:253t
- Performance-seeking portfolios (PSPs), I:36, I:37
- Perpetuities, II:607–608
- Pharmaceutical companies, II:7–8, II:11, II:244

- Phillips-Perron statistic, *II:386, II:398*
- Pickand-Balkema-de Haan theorem, *II:746*
- Pickand estimator, *III:273*
- Pliska, Stankey, *II:476*
- Plot function, in MATLAB, *III:428–432*
- P-null sets, *III:197*
- Pochhammer symbol, *III:256*
- Poincaré, Henri, *II:469*
- POINT[®]
- features of, *II:193n, II:291n*
 - modeling with, *II:182*
 - screen shot of, *II:287f, II:288f*
 - use of, *II:179, II:189, II:286–287*
- Point processes, *III:270–272*
- Poisson-Merton jump process, distribution tails for, *III:540–541*
- Poisson-Merton jump variable, *III:540*
- Poisson processes
- compounded, *III:497*
 - homogeneous, *III:270–271*
 - and jumps, *I:93, III:498, III:540*
 - for modeling durations, *II:461*
 - as stochastic process, *III:496, III:497, III:506*
 - use of, *I:262, I:315–316*
- Poisson variables, distribution of, *III:271f*
- Policy iteration algorithm (Howard algorithm), *II:676–677*
- Polyhedral sets, *I:33, I:33f*
- Polynomial fitting of trend stationary process, *II:702–703, II:702f*
- Population profiles, in transition matrices, *III:32–34*
- Portfolio allocation, example using MATLAB, *III:436–441*
- Portfolio management
- approaches to, *II:108–110*
 - checklist for robust, *III:144*
 - for credit risk, *I:416–417*
 - of large portfolios, *III:325*
 - and mean-variance framework, *I:196*
 - real world, *I:190*
 - software for, *II:75 (See also Excel)*
 - tax-aware, *II:74–75*
 - using Bayesian techniques, *I:196*
- Portfolio managers, *III:444–445*
- approaches used by, *II:108–109*
 - enhanced indexers, *II:268*
 - example of, *III:436–441, III:437t*
 - questions considered by, *II:277*
 - specialization of, *II:48–49*
 - traditional vs. quantitative, *II:109, II:110t*
 - types of, *II:179, II:286*
- Portfolio optimization
- for American options, *II:678*
 - classical mean-variance problem, *III:441–444*
 - constraints on, *II:62*
 - defined, *I:36*
 - formulation of theory, *II:476*
 - max-min problem, *III:139*
 - models of, *II:84–85n*
 - robust, *III:146*
 - techniques of, *II:115–116*
 - uncertainty in, *I:192–193, II:82–83*
- Portfolios. *See* constraints, portfolio allocation of, *I:192–193, II:72*
- assessment of risk factors of, *III:637–638*
- benchmark, *I:41–42, II:180*
- building efficient, *II:115*
- bullet vs. barbell, *III:308t, III:309t*
- bullet vs. barbell (hypothetical), *III:308*
- cap-weighted, *I:38f*
- centering optimal, *I:199*
- considerations for rebalancing of, *II:75*
- construction of, *I:37–38, II:56–57, II:102–104, II:102f, II:114–116, II:179–184, II:261–264, II:286–287, II:301–303*
- cor-plus, and DTS, *I:398*
- credit bond, hedging of, *I:405*
- data on, *II:365t*
- diversification of, *I:10–12*
- efficient, *I:12, I:77, I:288f, I:289f, I:290f*
- efficient set of, *I:13*
- efficient vs. feasible, *I:13*
- efficient vs. optimal, *I:5*
- examples of, *II:261t, II:262t*
- expected returns from, *I:6–7, I:7, I:12t, I:69t, I:195*
- factor exposures in, *II:183t, II:184t, II:263t, II:264t*
- factor model approach to, *II:224*
- feasible and efficient, *I:12–14*
- feasible set of, *I:12–13, I:13f*
- index-tracking, *II:186*
- information content of, *I:192*
- long-short, *II:181–182, II:226f*
- management of fixed-income, *I:391*
- and market completeness, *I:50–52*
- mean-variance efficient, *I:66, I:69f*
- mean-variance optimization of, *II:79*
- momentum, *II:182f*
- monitoring of, *II:106*
- MSR (maximum Sharpe Ratio), *I:36–37*
- normalized, *II:157*
- optimal, *I:14–15, I:14f, I:15–17, II:181t*
- optimization-based approach to, *II:224–225*
- optimization of, *I:17–18, I:40, II:56–57, II:301–303*
- optimized, *II:116*
- performance-seeking, *I:36*
- quadratic approximation for value, *III:644–645*
- rebalancing of, *II:287–288*
- replication of, *II:476*
- resampling of, *I:189, II:78–80, II:84*
- returns of, *I:6–7*
- risk control in, *II:181–182*
- riskless, *I:509*
- with risky assets, *I:12–17*
- robust optimization of, *II:80–84*
- rule-based, *II:116*
- selection of, *I:3–19, III:351–353, III:356*
- self-financing, *II:660–661*
- stress tests for, *I:412*
- tangency, *I:36–37*
- tilting of, *II:263–264*
- tracking, *II:187t*
- weighting in, *I:50–51, II:64–65*
- weights of, *I:191–192*
- yield simulations of, *I:284–285*
- Portfolio sorts
- based on EBITDA/EV factor, *II:216–217, II:216f*
 - based on revisions factor, *II:217–218, II:217f*
 - based on share repurchase factor, *II:218, II:218f*
 - information ratios for, *II:219*
 - results from, *II:225f*
 - use of, *II:214–219*
- Portfolio trades, arbitrage, *I:440t*
- Position distribution and likelihood function, *I:142–143*
- Positive homogeneity property, *III:327–328*
- Posterior distribution, *I:159, I:165*
- Posterior odds ratio, *I:157*
- Posterior tradeoff, and normal mean, *I:158–159*
- Power conditional value at risk measure, *III:356*
- Power law, *III:234–235*
- Power plants/refineries, valuation and hedging of, *I:563*
- Power sets, *III:156, III:156t*
- Precision, *I:158, II:702*
- Predictability, *II:122–127*
- Predictions, *I:167, II:124*
- Predictive return modes, adoption of, *II:128–129*

- Preferred habitat hypothesis, III:569–570
- Prepayments
 burnout, III:19
 calculating speeds of, III:50–56
 in cash-flow yields, III:4
 conditional rate of (CPR), III:30, III:50–51, III:58–59
 defaults and involuntary, III:59, III:74–77
 defined, III:50
 disincentives for, III:7–8
 drivers of, III:77
 effect of time on rates of, III:73–74
 evaluation of, III:62
 factors influencing speeds of, III:69–74
 fundamentals of, III:66–69
 for home equity securities, III:55–56
 interactions with defaults, III:76–77
 interest rate path dependency of, III:6
 lag in, III:24–25
 levels of analysis, III:50
 lock-ins, III:73
 modeling of, I:258, I:267, I:268, III:63n, III:598–600
 practical interpretations of, III:20
 rates of, III:74
 reasons for, III:48
 risk of, II:281, II:281t
 S-curves for, III:67–68, III:67f
 sources of, III:23–24
 voluntary, III:38
 voluntary *vs.* involuntary, III:30, III:75–76
- Prepay modeling, III:19–20
 rational exercise, III:25
- Present value, I:268n, II:19, II:603–604, II:609, III:9–10
- Price/earnings (P/E) ratio, II:20–21, II:38
- Price patterns, scaling in, III:279
- Price processes, bonds, I:128
- Prices
 bid/ask, III:625
 Black-Scholes, II:673–674
 changes in, II:722f, II:723f, II:742, III:305–306, III:305t
 compression of, III:303
 computing clean, I:214–215
 dirty, I:382
 distribution of, I:510
 estimating changes in bond, I:373–374
 flexible and sticky in CPI basket, I:292
 formula for discounted, I:110
 marked-to-market, I:430
 modeling realistic, I:93–94
 natural logarithm of, I:85
 path-dependent, III:193n
 strike, I:484–485, I:486
 truncation of, III:304
vs. value, I:455n
- Price time series, autocorrelation in, III:274
- Pricing
 backward induction, III:18
 formulas for relationships, I:105–110
 grids for, III:18–19
 linear, I:52–55
 models for, II:127–128
 rational, I:53
 risk-neutral, I:533, I:544
 rule representation, I:260–261
 use of trees, III:22–23
- Principal component analysis (PCA)
 compared to factor analysis, II:166–168
 concept of, II:157
 defined, II:147, II:276
 discussed, II:157–164
 illustration of, II:158–163
 with stable distributions, II:163–164
 usefulness of, II:158
 use of, I:39–40, II:142, II:168–169
- Principal components, defined, II:148, II:159
- Principal components analysis (PCA), I:556
- Prior elicitation, informative, I:152–153, I:159
- Prior precision, I:158
- Priors, I:153, I:165–167, I:168, I:171–172
- Probabilistic decision theory, II:719–721, II:729
- Probabilities
 in Bayesian framework, I:140, I:144, I:146–148
 conditional, I:117, II:517–518, III:477
 formulas for conditional, I:108t
 interpretation of, II:123
 in models, II:299
 posterior, I:140, I:144
 prior, I:140, I:144
 prior beliefs about, I:147
 realistic, III:596–597
 as relative frequencies, III:152
 risk-adjusted, I:264
 risk neutral, I:58–59, I:59, I:102, I:104, I:111–114, I:115–116, I:117, III:594–596
- Probability density function (PDF), III:384–385
- Probability distributions
 binomial, III:186t
 continuous, III:578
 for drawing black balls, III:176–177
 inverting the cumulating, III:646
 for prepayment models, III:598
 for rate of return, I:7t, I:9t
 use of, III:638, III:645–646
- Probability-integral transformation (PIT), III:365
- Probability law, III:161
- Probability measures, III:157–159, III:594–597
- Probability of default (PD). *See* default probabilities
- Probability theory, II:133, II:700–701
- Probit regression models, II:348–349, II:350
- Processes
 absolute volatility of, III:474
 exponential, III:498
 martingale, I:119, I:262–263, III:509, III:517
 non-decreasing, III:503–505
 normal tempered stable, III:504–505
 predictable, II:132–133
 subordinated, III:387–388
 weakly stationary, II:360–361
- Process maps, III:94
- Proctor & Gamble, cash flows of, II:567–568, II:568t, II:571–573, II:573t
- Product transitions, III:66, III:71–73
- Profit, riskless, I:480
- Profitability ratios, II:555–557, II:563
- Profit margin ratios, II:555–556
- Profit opportunities, I:261
- Programming, linear, I:29, I:32–33
- Programming, stochastic
 defined, III:123–124
 in finance, III:125–126
 general multistage model for
 financial planning, III:128–132
 use of scenario trees in, III:131–132
vs. continuous-time models, III:127–128
vs. other methods in finance, III:126–128
- Projected successive over relaxation (PSOR) method, II:677
- Projections, as-was, usefulness of, II:38
- Propagation effect, III:351
- Prospectus prepayment curve (PPC), III:54–55, III:56
- Protection, buying/selling of, I:230–231
- 100 PSA (Public Securities Association prepayment benchmark), III:51–52, III:55
- Pseudo-random numbers, generation of, III:647

- PSPs (performance-seeking portfolios), *I:36, I:37*
- Public Securities Association (PSA) prepayment benchmark, *III:51–55, III:51f, III:62–63*
- Pull to par value, *I:216*
- Pure returns, *II:51*
- Put-call parity, *I:437*
for American-style options, *I:446–448, I:452–453, I:452t*
for European options, *I:499*
for European-style options, *I:444–446, I:445t, I:451, I:451t*
perfect substitutes in European-style, *I:445t*
relations of, *I:446*
- Put-call parity relationship, *I:445, I:446, I:485*
- Put options, *I:439*
- Puts, American-style
early exercise of, *I:444, I:450–451*
error on value of, *II:677t, II:678t*
lower price bound, *I:443–444, I:450*
numerical results for, *II:677–678*
- Puts, European-style
arbitrage trades, *I:443t*
lower price bound, *I:443, I:450*
- Pyrrho's lemma, *II:330, II:331*
- Q-statistic of squared residuals, *II:422*
- Quadratic objective, two-dimensional, *I:29f*
- Quadratic programming, *I:29, I:33–34*
- Quadratic variation, *III:474*
- Quantiles
development of regression, *II:356*
methods, *II:354–356*
plot (QQ-plot) of, *III:272*
use of regression, *II:353–354, II:356–357*
- Quantitative methods, *II:483*
- Quantitative portfolio allocation, use of, *I:17–18*
- Quantitative strategies, backtesting of, *I:201*
- Quintile returns, *II:97–98*
- Quotes
delayed, *II:454*
discrepancies in, *II:453–454*
histograms from simple returns, *II:458f*
methods for sampling, *II:457–460*
mid-quote closing, *II:460f*
mid-quote format, *II:456*
mid-quote time-interpolated, *II:460f*
quantile plots of, *II:459f, II:461f*
- R^2 , adjusted, *II:315–316*
- Radon-Nikodym derivative, *I:111, I:130, I:133–134, III:510–511, III:515*
- Ramp, loans on, *III:52*
- Randomized operational time, *III:521*
- Randomness, *I:164, III:534–537, III:580*
- Random numbers
clusters in, *III:649–650*
generation of, *III:645–647*
practicality of, *III:647*
reproducing series of, *III:646*
simulations of, *III:650f*
- Random walks
advanced models of, *I:92–94*
arithmetic, *I:82–84, I:97, II:125*
for Brownian motion, *III:478–479*
computation of, *I:83, I:85, I:87, I:90*
correlated, *I:92–93, II:502–503*
defined, *III:486*
in forecastability, *II:127*
generation of, *I:85*
geometric, *I:84–88, I:89, I:97*
and linear nonstationary models, *II:508*
multivariate, *I:93*
parameters of, *I:87–88*
polynomial fitting of, *II:704f*
simulation of, *I:87*
and standard deviation, *II:385*
500-step samples, *II:708f*
strict, *II:126*
use of, *II:132, III:474*
variables in, *I:83–84*
- Range notes, valuing, *I:252*
- RAS Asset Management, *III:624*
- Rate-and-term refinancing, *III:66*
- Rating agencies, *I:300, III:44*
effect of actions of, *I:367–369*
role of, *I:362*
- Rating migration, *I:362, I:367–369*
- Rating outlooks, *I:365–366*
- Ratings
maturity of, *I:301*
- Ratings-based step-ups, *I:352*
- Rating transitions, *I:368, I:368t, I:381*
- Ratios
analysis of, *II:575–576*
classification of, *II:545–546*
defined, *II:545*
quick (acid test), *II:554*
scales of, *II:487*
- Real estate prices, effect of, *III:44*
- Real yield duration, calculation of, *I:286*
- Receipts, depository, *II:36*
- Recoveries, in foreclosures, *III:75*
- Recovery percentages, *III:30–31*
- Recovery rates
calibration of assumption, *I:537–538*
for captive finance companies, *I:366–367*
and credit risk, *I:362*
dealing with, *I:334n*
on defaulted securities, *I:367t*
drivers of, *I:372*
modeling of, *I:316–317*
random, *I:383*
relationship to default process, *I:372, I:376*
time dimension to, *I:366–377*
- Rectangular distribution, *III:219–221*
- Recursive out-of-sample test, *II:236*
- Recursive valuation process, *I:244*
- Reduced form models, usefulness of, *I:412*
- Redundant assets/securities, *I:51*
- Reference entities, *I:526*
- Reference priors, *I:159–160n*
- Refinancing
and ARMs, *III:72*
categories of, *III:48*
discussion of, *III:68–69*
rate-and-term, *III:68*
speed of, *III:25–26*
threshold model, *III:18*
- Refinancing, paths of rates, *III:8t*
- RefiSMM(Price) function, *III:25–26*
- Regime switching, *I:173n*
- Regression
binary, *III:364*
properties of, *II:309–310*
spurious, *II:384, II:385*
stepwise, *II:331*
- Regression analysis
results for dummy variable regression, *II:348t*
usefulness of, *II:305*
use in finance, *II:316–328*
variables in, *II:330*
- Regression coefficients, testing of, *I:170*
- Regression disturbances, *I:164*
- Regression equations, *II:309–310*
- Regression function, *II:309*
- Regression models, *I:168–169, I:170–172, II:302*
- Regressions
estimation of linear, *II:311–314*
explanatory power of, *II:315–316*
linear, *II:310–311*
and linear models, *II:308–311*
pitfalls of, *II:329–330*
sampling distributions of, *II:314*
spurious, *II:329*
- Regression theory, classical, *II:237*
- Regressors, *II:308–310, II:311, II:330*

- Reg T (Treasury Regulation T), *I:67*
- Relative valuation analysis
 hypothetical example of, *II:40–45*
 hypothetical results, *II:40t*
 implications of hypothetical,
II:41–42
 low or negative numbers in, *II:42–43*
- Relative valuation methods
 choice of valuation multiples in,
II:38–39
 usefulness of, *II:45*
 use of, *II:33–34, II:45*
- Replication, *I:526*
- Reports, *II:200–201, II:283–286*
- Research, process of quantitative,
II:717f
- Residuals, *II:220, II:328–329*
- Restructuring, *I:528–530, I:529, I:529t, I:530, I:537*
- Return covariance matrix formula,
II:141
- Return distributions, *III:333f, III:388–392*
- Return effects, *II:47–48, II:51, II:51f*
- Return generating function, *II:256*
- Return on assets, *II:547–548, II:548–550*
- Return on equity (ROE), *II:37–38, II:41–42, II:548, II:550*
- Return on investment ratios,
II:547–551, II:548, II:563
- Returns
 active, *II:115*
 arithmetic *vs.* geometric average,
II:598
 defined, *II:598*
 estimated moments of, *II:204*
 estimates of expected, *I:190–191*
 ex ante, *I:7*
 excess, *I:66, I:67, I:74*
 expected, *I:71–72, II:13–14, II:112*
 ex post, *I:6*
 fat tails of conditional distribution,
II:753n
 finite variance of, *III:383–384*
 forecasting of, *II:111–112, II:362*
 historical, *II:285f, III:389t*
 monthly *vs.* size-related variables,
II:52t
 naïve, *II:51, II:53f*
 naïve *vs.* pure, *II:52f, II:53–54*
 Nasdaq, Dow Jones, bond, *II:365f*
 pure, *II:51, II:53f, II:54t*
 robust estimators for, *I:40–41*
 rolling 24-month, *II:229f*
 systematic *vs.* idiosyncratic, *II:173*
 time-series properties of, *II:733–734*
- Returns to factors, *II:248*
- Return to maturity expectations
 hypothesis, *III:569*
- Return volatility, excess and DTS,
I:396–397
- Reverse optimization, *I:203n*
- Riemann-Lebesgue integrals, *III:483*
- Riemann-Stieltjes integrals, *I:122, III:473–474, III:487*
- Riemann sum, *II:743–744*
- Risk. *See also* operational risk
 alternative definitions of, *III:350*
 analyzing with multifactor models,
II:184–188
 assessment of, *III:640–641*
 asymmetry of, *III:350–351*
 budgeting of, *II:115, II:286–287*
 of CAPM investors, *I:73–74*
 changes in, *II:368, III:351*
 coherent measures of, *III:327–329*
 collective, *II:470*
 common factor/specific, *II:258*
 controlling, *I:397*
 correlated, *II:271t*
 correlated *vs.* isolated, *II:271*
 counterparty, *I:478, I:479*
 decomposition of, *II:250–253, II:257–261, II:265*
 and descriptors, *II:140*
 downside, *III:382*
 effect of correlation of asset returns
 on portfolio, *I:11–12*
 effect of number of stocks on,
II:249f
 estimation of, *I:40*
 in financial assets, *I:369*
 forecasting of, *II:112–113*
 fundamental, *II:199*
 funding, *II:199*
 horizon, *II:199*
 idiosyncratic, *II:178, II:188, II:188t, II:283, II:285t, II:291*
 idiosyncratic *vs.* systematic, *I:40–41*
 implementation, *II:199*
 including spread in estimation of,
I:399
 indexes of, *II:140, II:256*
 interest rate, *I:521–522, III:4*
 issue specific, *II:283t*
 liquidity, *II:199*
 main sources of, *II:211*
 market price of, *III:579, III:588, III:591*
 model (*See* model risk)
 modeling, *III:11*
 momentum, *II:181t*
 as multidimensional phenomenon,
III:350
 noise trader, *II:199*
 perspective on, *II:91–92*
 portfolio, *I:7–10, I:9–10, I:11, II:180t*
 repayment, *III:48*
 price movement costs, *II:69*
 quantification of, *I:4, I:7–8*
 realized, *II:118*
 reinvestment, *III:4–5*
 relativity of, *III:350*
 residual, *II:258–259*
 by sector, *II:185t*
 in securities, *I:73*
 sources of, *II:173–174, II:251f, II:274, II:281–282*
 systematic, *II:186*
 tail, *I:384, I:385*
 true *vs.* uncertainty, *II:721*
 in a two-asset portfolio, *I:8*
 in wind farm investments, *I:563–564*
- Risk analysis, *II:268–286, II:273t, II:274t, II:275t*
- Risk aversion, *I:404*
 in analysis, *III:570*
 coefficient for, *I:59*
 functions, *III:339f*
 of investors, *I:191*
 and portfolio management, *I:37*
- Risk-based pricing, *III:70*
- Risk decomposition
 active, *II:259–260, II:259f*
 active systematic-active residual,
II:260, II:260f
 insights of, *II:252*
 overview of, *II:261f*
 summary of, *II:260–261*
 systematic-residual, *II:258–259, II:259f*
 total risk, *II:258, II:258f*
- Risk exposures, *I:394, I:521*
- Risk factors
 allocation of, *I:398*
 constraints on, *II:63–64*
 identification of, *II:256*
 macroeconomic, *I:415–416*
 missing, *II:693*
 systematic, *II:268, II:474*
 unsystematic, *II:474*
- Riskiness, determining, *I:145*
- Risk management
 internal models of, *III:289–290*
 in investment process, *II:104*
 portfolio, *III:643–644*
 in portfolio construction, *II:303*
 and quasi-convex functions, *I:28*
- Risk measures, safety-first, *III:352, III:354–356, III:357*
- RiskMetrics™ Group
 approach of, *III:322–323*
 comparison with FTSE100 volatility,
III:413f
 methodology of, *III:412–413*
 software of, *III:413*
 website of, *III:412*

- Risk models
 - applications of, *II:286–290*
 - comparisons among, *II:747–751*
 - defined, *II:692*
 - equity, *II:172–173, II:192–193, II:255, II:264*
 - indicator, *III:93–94*
 - and market volatility, *II:748*
 - multifactor, *II:257–258*
 - principal of, *II:292n*
 - and uncertainty, *II:724*
 - use of, *II:171–172, II:268, II:290*
- Risk neutral, use of term, *III:593–594*
- Risk neutral density (RND)
 - concept of, *II:521*
 - fitting data to models of, *II:526–527*
 - generally, *II:527*
 - parametric models for, *II:523–525*
- Risk oversight, *II:303*
- Risk premiums
 - for default, *III:599*
 - importance of, *III:587*
 - quantifying, *III:580–581*
 - of time value, *I:513*
 - as a variable in discount bond prices, *III:581*
 - variables, *I:403, I:405*
- Risk reports
 - credit risk, *II:278–281*
 - detailed, *II:272–286*
 - factor exposure, *II:275–283*
 - implied volatility, *II:282*
 - inflation, *II:282*
 - issue-level, *II:283–285*
 - liquidity, *II:282*
 - prepayment risk, *II:281*
 - risk source interaction, *II:281–282*
 - scenario analysis, *II:285–286*
 - summary, *II:272–275*
 - tax-policy, *II:282–283*
- Risk tolerance, *II:720–721, II:725, II:729f*
- Risky bonds, investment in, *II:726–729*
- Robot analogy, *III:594*
- Robust covariance matrix, *II:446*
- Robust optimization, *II:83, III:141–142*
- Robust portfolio optimization, *I:17–18, I:193, III:138–142*
 - effect on performance, *III:144*
 - need for research in, *III:145–146*
 - practical considerations for, *III:144–145*
 - in practice, *III:142–144*
- Rolling windows, use of, *II:371*
- Roots
 - complex, *II:632–634, II:636–637*
 - in homogenous difference equations, *II:642*
 - real, *II:630–632, II:635–636*
- Ross, Stephen, *II:468, II:475*
- Rounding, impact of, *III:306n*
- Roy CAPM, *I:67, I:69, I:70*
- Ruin problem, development of, *II:470–471*
- Runge-Kutta method, *II:650–652, II:651f, II:652f*
- Russell 1000, *II:213, II:236–237*
- Saddle points, *I:23, I:23f, I:30*
- Sales, net credit, *II:557–558*
- Samples
 - effect of size, *I:158–159, I:159f, III:407*
 - importance of size, *III:152*
 - and model complexity, *II:703–707*
 - in probability, *III:153*
 - selection of, *II:716*
- Sampling
 - antithetic, *I:383*
 - importance, *I:384, III:648–649*
 - stratified, *II:115, III:648*
- Sampling error, *III:396*
- Samuelson, Paul, *I:556, II:468, II:473–474*
- Sandmann-Sondermann model, *I:493*
- Sarbanes-Oxley Act (2002), *II:542*
- Scalar products, *II:625–626*
- Scale parameters, *I:160n*
- Scaling laws, use of, *III:280*
- Scaling vs. self-similarity, *III:278–280*
- Scenario analysis
 - constraints on, *III:130*
 - factor-based, *II:189–192, II:193*
 - for operational risk, *III:93*
 - usefulness of, *II:179*
 - use of, *II:288–290, III:378*
- Scenarios
 - defined, *III:128*
 - defining, *II:189*
 - generation of, *III:128–132*
 - network representation of, *III:129f*
 - number needed of, *III:640–641*
- Scholes, Myron, *II:468, II:476*
- Schönbucher-Schubert (SS) approach, *I:329–331*
- Schwarz criterion, *II:387, II:389*
- Scorecard Approach, *III:100n*
- Scott model, *II:681–682*
- SDMs (state dependent models), *I:342, I:351–352*
- Secrecy, in economics, *II:716*
- Sector views, implementation of, *II:182–184*
- Securities
 - alteration of cash flows of, *I:210*
 - arbitrage-free value of, *I:261*
 - baskets of, *I:483–484*
 - convertible, *I:462*
 - creating weights for, *II:102–104, II:103f*
 - evaluation of, *I:50*
 - fixed income, *I:209–210, II:268*
 - formula for prices, *I:107*
 - non-Treasury, *I:222–223, I:223t*
 - of other countries, *I:226*
 - payoffs of, *I:49–50, I:116–117, I:121–122*
 - pricing European-style, *III:642*
 - primary, *I:458*
 - primitive, *I:51*
 - private label (*See* MBS (mortgage-backed securities), nonagency)
 - ranking of, *I:200–201*
 - redundant, *I:124*
 - risk-free, *I:115*
 - selection of, *I:225–226*
 - structured, *I:564, I:565–566*
 - supply and demand schedule of, *III:626f*
 - valuing credit-risky, *III:645*
 - variables on losses, *I:370*
- Securities and Exchange Commission (SEC)
 - filings with, *II:532*
- Security levels, two-bond portfolio, *I:382t*
- Selection, adverse vs. favorable, *III:76–77*
- Self-exciting TAR (SETAR) model, *II:405*
- Self-similarity, *III:278–280*
- Selling price, expected future, *II:19–20*
- Semimartingales, settings in change of time, *III:520–521*
- Semi-parametric models
 - tail in, *II:744–747*
- Semiparametric/nonparametric methods, use of, *II:522*
- Semivariance, as alternative to variance, *III:352*
- Sensitivity, *III:643–644*
- Sensitivity analysis, *I:192, II:235*
- Sequences, *I:378, III:649–651, III:650*
- Series, *II:299, II:386, II:507–508, II:512*
- SETAR model, *II:425–426*
- Set of feasible points, *I:28, I:31*
- Set operations, defined, *III:153–154*
- Sets, *III:154*
- Settlement date, *I:478*
- Settlements, *I:526–528*
- Shareholders
 - common, *II:4*
 - equity of, *II:535*
 - negative equity of, *II:42*
 - preferred, *II:4–5*
 - statement of equity, *II:541*

- Shares, repurchases of, *II:207, II:210f, II:211, II:215–216, II:227*
- Sharpe, William, *I:75, II:468, II:474*
- Sharpe-Lintner CAPM (SL-CAPM), *I:66–67, I:75, I:78n*
- Sharpe ratios, *I:40, I:62, I:193*
- Sharpe's single-index model, *I:74–75*
- Shipping options, pricing of, *I:565*
- Shortfall, expected, *I:385–386*
- Short positions, *I:67*
- Short rate models, *III:543–545, III:545–550, III:552–554, III:557, III:604–610*
- Short rates, *III:212–213, III:541, III:549, III:595–596*
- Short selling
 constraints on, *I:67*
 effect of constraints on, *I:17, I:191–192, II:461*
 effect of on efficient frontiers, *I:17f*
 example, *I:480–481*
 as hedging route, *I:409*
 in inefficient markets, *I:71f*
 and market efficiency, *I:70–71*
 net portfolio value, *I:433t*
 and OAS, *I:259*
 and real estate, *II:396–397*
 in reverse cash-and-carry trade, *I:483*
 for terminal wealth positions, *I:460–461*
 using futures, *I:432–433*
- Shrinkage
 estimation of, *I:192, I:194–195, I:201–202, III:142*
 optimal intensity of, *I:202n–203n*
 use of estimators, *II:78*
- δ -algebra, *III:15, III:157*
- δ -fields
 defined, *III:508*
- Signals (forecasting variables), use of
 in forecasting returns, *II:111–112*
 evaluation of, *II:111–112*
- Similarity, selecting criteria for, *II:35*
- Simulated average life, *III:12*
- Simulations
 credit loss, *I:378–380*
 defined, *III:637*
 efficiency of, *I:384*
 financial applications of, *III:642–645*
 process of, *III:638*
 technique of, *III:444–445*
- Single firm models, *I:343–352*
- Single monthly mortality rate (SMM), *III:50–51, III:58*
- Skewness
 defined, *III:238–239*
 and density function, *III:204–205*
 indicating, *III:235*
 and the Student's *t*-distribution, *III:387*
 treatment of stocks with, *I:41*
- Sklar's theorem, *I:326, III:288*
- Skorokhod embedding problem, *III:504*
- Slackness conditions, complementary, *I:32*
- SL-CAPM (Sharpe-Lintner CAPM), *I:66–67, I:75, I:78n*
- Slope elasticity measure, *III:315, III:317*
- Smith, Adam, *II:468, II:472*
- Smoothing, in nonparametric methods, *II:411–412*
- Smoothing constant, *III:409–410*
- Smoothly truncated stable distribution (STS distribution), *III:245–246*
- Smooth transition AR (STAR) model, *II:408–409*
- Sobol sequences, pricing European call options with, *III:445–446*
- Software
 case sensitivity of, *III:434*
 comments in MATLAB code, *III:427*
 developments in, *II:481–482*
 macros in, *III:450–452, III:450f, III:460, III:466*
 pseudo-random number generation, *III:646–647*
 random number generation commands, *III:645–647*
- RiskMetrics Group, *III:413, III:644*
- simulation, *III:651f*
 for stable distributions, *III:344, III:383*
 stochastic programming applications, *III:126*
 use of third party, *II:481*
- Solutions, stability of, *II:652–653*
- Solvers, in MATLAB, *III:435*
- Space in probability, *III:156, III:157*
- Sparse tensor product, *II:673*
- S&P 60 Canada index, *I:550–552, I:550t, I:553f*
- Spearman, Charles, *II:153–154*
- Spearman model, *II:153–154*
- Spearman's rho, *I:327, I:332, I:336n*
- Splits, in recursive partitioning, *II:376–377*
- Spot curves, with key rate shifts, *III:313f, III:314f*
- Spot price models, energy commodities, *I:556–557*
- Spot rates
 arbitrage-free evolution of, *I:557–558*
 bootstrapping of curve, *I:217–220*
 calculation of, *III:581*
 and cash flows in OAS analysis, *I:259*
 changes in, *III:311, III:312f, III:312t*
 computing, *I:219–220*
 under continuous compounding, *III:571*
 defined, *III:595*
 effect of changes in, *I:514, III:313–314, III:314t*
 and forward rates, *III:572*
 models of, *III:579–581*
 paths of monthly, *III:9–10, III:10t*
 theoretical, *I:217*
 Treasury, *I:217*
 uses for, *I:222*
- Spot yields, *III:565, III:566, III:571*
- Spread analysis, *II:290t*
 table of, *II:290t*
- Spread duration, beta-adjusted, *I:394*
- Spreads
 absolute and relative change volatility, *I:396f*
 change in, *I:392, I:393, I:394f, I:399*
 determining for asset swaps, *I:227–228*
 level vs. volatility of, *I:397*
 measurement of, *II:336–337*
 measure of exposure to change in, *I:397*
 nominal, use of, *III:5*
 option-adjusted, *I:253–255, I:254f*
 reasons for, *I:210–211*
 relative vs. absolute modeling, *I:393*
 volatility vs. level, *I:394–396, I:395f*
 zero-volatility, *III:5*
- Squared Gaussian (SqG) model, *III:547–548*
- Square-root rule, *III:534*
- SR-SARV model class, *II:370*
- St. Petersburg paradox, *III:480*
- Stability
 notion of, *II:667*
 in Paretian distribution, *II:739–741*
 property of, *II:740–741, III:236–237, III:244–245*
- Stable density functions, *III:236f*
- Stable Paretian model, α -stable distribution in, *II:748*
- Standard Default Assumption (SDA) convention, *III:59–60, III:60f*
- Standard deviations
 and covariance, *I:9*
 defined, *III:168*
 mean, *III:353*
 posterior, *I:155*
 related to variance, *III:203–204*
 rolling, *II:362–363*

- Standard deviations (*Continued*)
 and scale of possible outcomes,
III:168f
 for tail, *III:341*
- Standard errors. *See also* errors
 for average estimators, *III:400–402*
 defined, *III:399*
 estimation of, *III:640*
 of the estimator, *III:400*
 for exponentially weighted moving
 averages (EWMA), *III:411–412*
 reduction of, *III:648*
- Standard normality, testing for,
III:366–367
- Standard North American contract
 (SNAC), *I:529*
- Standard & Poors 500
 auto correlation functions of, *II:389t*
 cointegration regression, *II:390t*
 daily close, *III:402f*
 daily returns (2003), *III:326f*
 distributions of, *III:384f*
 error correction model, *II:391t*
 historical distributions of, *III:390f*
 index and dividends (1962–2006),
II:388f
 parameter estimates of, *III:385t*,
III:387t, *III:388t*
 return and excess return data
 (2005), *II:316–317t*
 stationarity test for, *II:389t*
 time scaling of, *III:383f*
 worst returns for, *III:382t*
- State dependent models (SDMs), *I:342*,
I:351–352
- Statement of stockholders' equity,
II:541
- State price deflators
 defined, *I:103*, *I:129–130*
 determining, *I:118–119*, *I:124*
 formulas for, *I:107–108*, *I:109–110*
 in multiperiod settings, *I:105*
 and trading strategy, *I:106*
- State prices
 and arbitrage, *I:55–56*
 condition, *I:54*
 defined, *I:101–102*
 and equivalent martingale
 measures, *I:133–134*
 vectors, *I:53–55*, *I:58*, *I:119*
- States, probabilities of, *I:115*
- States of the world, *I:457–458*, *I:459*,
II:306, *II:308*, *II:720*
- State space, *I:269n*
- Static factor models, *II:150*
- Stationary series, trend *vs.* difference,
II:512–513
- Stationary univariate moving average,
II:506
- Statistical concepts, importance of,
II:126–127
- Statistical factors, *II:177*
- Statistical learning, *II:298*
- Statistical methodology, EWMA,
III:409
- Statistical tests, inconsistencies in,
II:335–336
- Statistics, *II:387*, *II:499*
- Stein paradox, *I:194*
- Stein-Stein model, *II:682*
- Step-up callable notes, valuing of,
I:251–252
- Stochastic, defined, *III:162*
- Stochastic control (SC), *III:124*
- Stochastic differential equations
 (SDEs)
 binomial/trinomial solutions to,
III:610–613
 with change of time methods,
III:523
 defined, *II:658*
 examples of, *III:523–524*
 generalization to several
 dimensions with, *III:490–491*
 intuition behind, *III:486–487*
 modeling states of the world with,
III:127
 for MRAM equation, *III:525–526*
 setting of change of time, *III:521*
 solution of, *III:491–493*
 steps to definition, *III:487*
 usefulness of, *III:493*
 use of, *II:295*, *III:485–486*,
III:489–490, *III:536*, *III:603*,
III:619
- Stochastic discount factor, *I:57–58*
- Stochastic integrals
 defined, *III:481–482*
 intuition behind, *III:473–475*
 in Ito processes, *III:487*
 properties of, *III:482–483*
 steps in defining, *III:474–475*
- Stochastic processes
 behavior of, *I:262*
 characteristic function of, *III:496*
 characteristics of, *II:360*
 continuous-time, *III:496*, *III:506*
 defined, *I:263–264*, *I:269n*, *II:518*,
III:476, *III:496*
 discrete time, *II:501*
 properties of, *II:515*
 representation of, *II:514–515*
 and scaling, *III:279*
 specification of, *II:692–693*
- Stochastic programs
 features of, *III:124*, *III:132*
- Stochastic time series, linear,
II:401–402
- Stochastic volatility models (SVMs)
 with change of time, *III:520*
 continuous-time, *III:656*
 discrete, *III:656–657*
 importance of, *III:658*
 for modeling derivatives,
III:655–656
 multifactor models for, *III:657–658*
 and subordinators, *III:521–522*
 use of, *III:653*, *III:656*
- Stock indexes
 interim cash flows in, *I:482*
 risk control against, *II:262–263*
- Stock markets
 bubbles in, *II:386*
 as complex system, *II:47–48*
 1987 crash, *II:521*, *III:585–586*
 dynamic relationships among,
II:393–396
 effects of crises, *III:233–234*
 variables effects on different sectors
 of, *II:55*
- Stock options, valuation of long-term,
I:449
- Stock price models
 binomial, *III:161*, *III:171–173*, *III:173f*
 multinomial, *III:180–182*, *III:181f*,
III:184
 probability distribution of
 two-period, *III:181t*
- Stock prices
 anomalies in, *II:111t*
 behavior of, *II:58*
 correlation of, *I:92–93*
 and dividends, *II:4–5*
 lognormal, *III:655–656*
 processes of, *I:125*
- Stock research, main areas of, *II:244t*
- Stock returns, *II:56*, *II:159f*
- Stocks
 batting average of, *II:99*, *II:99f*
 characteristics of, *II:204*
 common, *II:4*, *II:316–322*
 cross-sectional, *II:197*
 defined, *II:106*
 defining parameters of, *II:49*
 determinants of, *II:245f*
 execution price of, *III:626*
 fair value *vs.* expected return, *II:13f*
 finding value for XYZ, Inc., *II:31t*
 information coefficient of, *II:98f*
 information sources for, *II:90f*
 measures of consistency, *II:99–100*
 mispriced, *II:6–7*
 quantitative research metrics tests,
II:97–99
 quintile spread of, *II:97f*
 relative ranking of, *I:196–197*
 review of correlations, *II:101f*

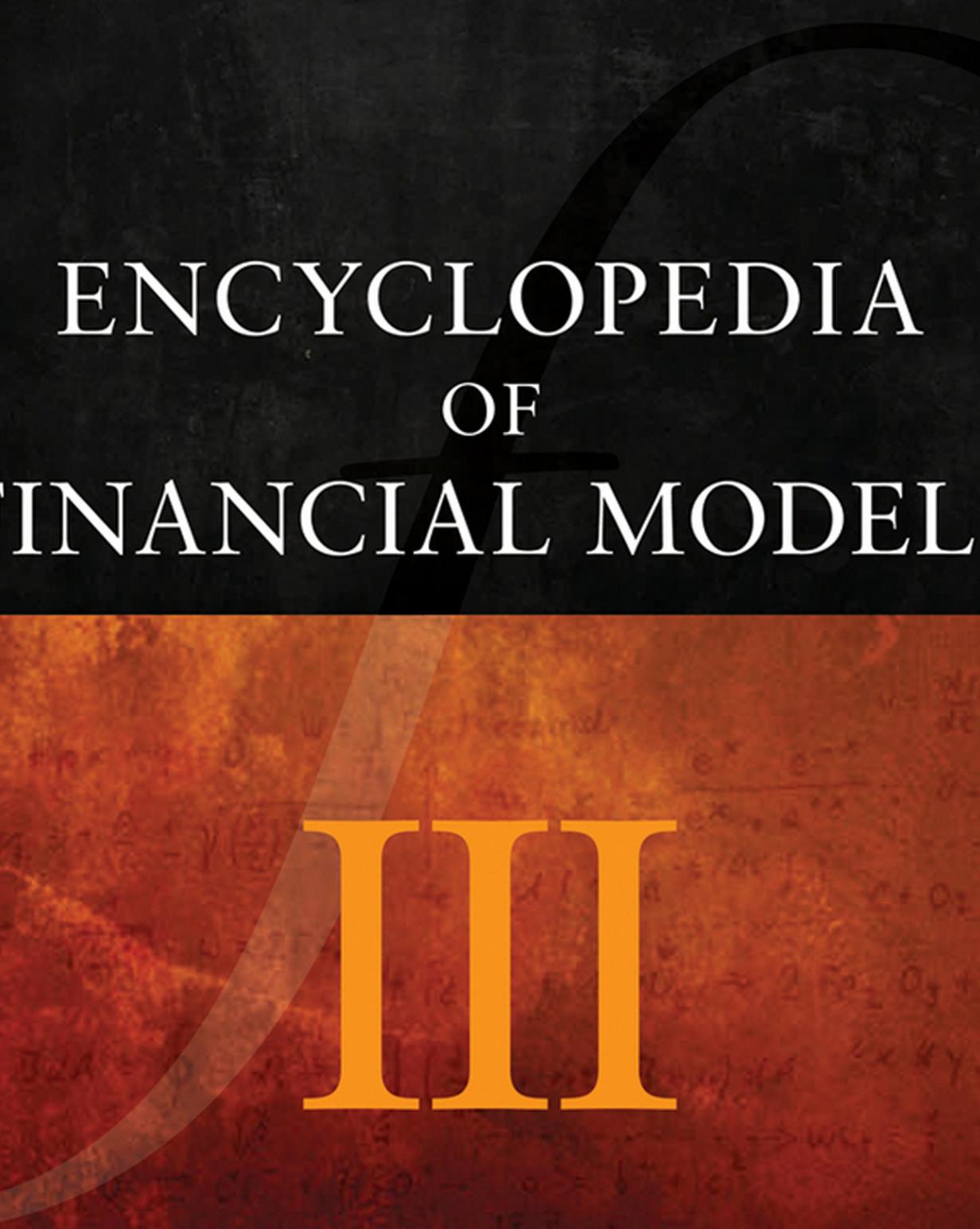
- sale/terminal price of, *II:5*
 short selling of, *I:432–433*
 similarities between, *II:245f*
 sorting of, *II:215*
 testing of, *II:95, II:96f*
 that pay no dividend, *II:17*
 use of, *II:90*
 valuation of, *II:6, II:8–9, II:14, II:18–19*
 weightings of, *II:101f*
- Stock selection**
 models for, *II:197*
 in quantitative equity investment process, *II:105*
 quantitative model, *II:94–95*
 for retail sector, *II:94f*
 strategies for, *II:195*
 tree for, *II:379–381, II:380f*
- Stopping times, *II:685***
Straontonovich, Ruslan, *II:470*
Strategies, backtesting of, *II:235–236*
Stress tests, *I:412, I:417, I:418, III:93, III:596–597*
Strike price, *I:509, I:514*
Strong Law of Large Numbers (SLLN), *I:270n, III:263–264*
Structural breaks, *I:167, III:274–275*
Student's *t* distribution
 applications to stock returns, *III:215–216*
 and AVaR, *III:334–335*
 classical, *II:734–738*
 density function of, *II:735*
 discussion of, *III:213–216*
 distribution function of, *III:215f*
 for downside risk estimation, *III:386–387*
 fitting and simulation of, *II:737–738*
 heavy tails of, *I:160n, I:176, II:747–748, II:751, III:227–228*
 limitations of, *II:736*
 in modeling credit risk, *I:387–388*
 normals representation in, *I:177–178*
 skewed, *II:736–737, II:753n*
 skewness of, *III:390*
 standard deviation of, *I:173n*
 symmetry of, *III:387*
 tails of, *III:392*
 use of, *I:153–154, I:172n, III:234*
- Student's *t*-test, *II:219***
Sturge's rule, *II:495*
Style analysis, *II:189*
Style factors, *II:247*
Style indexes, *II:48*
Stylized facts, *II:503–504*
Subadditivity property, *III:328*
Subordinated processes, *I:186n, III:277, III:521–522*
- Successive over relaxation (SOR) method, *II:677***
Summation stability property (Gaussian distribution), *II:732–733*
Supervisory Capital Assessment Program, *I:300, I:412*
Support, defined, *II:200*
Survey bias, *I:293*
Survival probability, *I:533–535*
Swap agreements, *I:434, I:435–436n*
Swap curves, *I:226, II:275–276*
Swap rates, *I:226, III:536f*
Swaps
 with change of time method, *III:522*
 covariance/correlation, *I:547–548, I:549–550, I:552*
 duration-matched, *I:285*
 freight rate, *I:558*
 modeling and pricing of, *I:548–550*
 summary of studies on, *I:546t*
 valuing of, *I:434–435*
- Swap spread (SS) risk, *II:278, II:278t***
Swaptions, *I:502–503, III:550*
Synergies, in conglomerates, *II:43–44*
Systematic risk, *II:290*
Systems
 homogenous, *II:624*
 linear, *II:624*
 types of, *II:47, II:58*
- Tailing the hedge, defined, *I:433***
Tail losses
 in loss functions, *III:369–370*
Tail probability, *III:320*
Tail risk, *I:377, I:385, II:752*
Tails
 across assets through time, *II:735–736*
 behavior of in operational losses, *III:111–112*
 in density functions, *III:203*
 dependence, *I:327–328, I:387*
 Gaussian, *III:98–99, III:260*
 heavy, *II:734–744, III:238*
 modeling heaviness of, *II:742–743*
 for normal and STS distributions, *III:246t*
 power tail decay property, *II:739, III:244*
 properties of, *III:261–262*
 tempering of, *II:741*
- Takeovers, probability of, *I:144–145***
Tangential contour lines, *I:29–30, I:30f, I:32f*
Tanker market, *I:565*
TAR-F test, *II:426*
TAR(1) series, simulated time plot of, *II:404f*
- Tatonnement, concept of, *II:468***
Taxes
 and bonds, *I:226*
 capital gains, *II:73*
 cash, *II:573*
 for cash/futures transactions, *I:484*
 complexity of, *II:73–74*
 deferred income, *II:535, II:538*
 effect on returns, *II:83–84, II:84, II:85n*
 in financial statements, *II:541*
 impact of, *I:286–287*
 incorporating expense of, *II:73–75*
 managing implications of, *III:146*
 and Treasury strips, *I:218*
- Tax policy risk, *II:282–283***
Technology, effect of on relative values, *II:37*
Telescoping futures strategy, *I:433*
Tempered stable distributions
 discussions of, *III:246–252, III:384–386*
 generalized (GTS), *III:249*
 Kim-Rachev (KRTS), *III:251–252*
 modified (MTS), *III:249–250*
 normal (NTS), *III:250–251*
 probability densities of, *III:247f, III:248f, III:250f, III:252f*
 rapidly decreasing (RDTS), *III:252*
 tempering function in, *III:254, III:258n*
- Tempered stable processes, *III:499–501, III:500t, III:512–517***
Tempering functions, *III:254, III:255t*
Templates, for data storage, *II:204*
Terminal profit, options and forwards, *I:438f, I:439f*
Terminal values, *II:45*
Terminology
 of delinquency, default and loss, *III:56*
 of prepayment, *III:49–50*
 standard, of tree models, *II:376*
- Term structure**
 in contiguous time, *III:572–573*
 continuous time models of, *III:570–571*
 defined, *III:560*
 eclectic theory of, *III:570*
 of forward rates, *III:586*
 mathematical relationships of, *III:562*
 modeling of, *I:490–494, III:560*
 of partial differential equations, *III:583–584*
 in real world, *III:568–570*
- Term structure modeling**
 applications of, *III:584–586*
 arbitrage-free, *III:594*

- Term structure modeling (*Continued*)
 calibration of, III:580–581
 discount function in, III:565
 discussion of, III:560–561
- Term structure models
 approaches to, III:603–604
 defined, I:262, I:263
 discrete time, III:562–563
 discussion of, III:561–562
 of interest rates, I:314
 internal consistency checks for, III:581
 with no mean reversion, III:613–616
 for OAS, I:265–267
 quantitative, III:563
 static *vs.* dynamic, III:561–562
- Term structures, III:567–568, III:570, III:579, III:587
- Tests
 Anderson-Darling (AD), III:112–113
 BDS statistic, II:423–424, II:427
 bispectral, II:422–423
 cointegration, II:708–710
 Kolmogorov-Smirnov (KS), III:112–113
 monotonic relation (MR), II:219
 nonlinearity, II:426–427, II:427^t
 nonparametric, II:422–424
 out-of-sample *vs.* in-sample, II:236
 parametric, II:424–426
 RESET, II:424–425
 run tests, III:364
 threshold, II:425–426
 for uniformity, III:366
- TEV (tracking error volatility), II:180, II:186, II:272–274, II:286–287
- Theil-Sen regression algorithm, II:440–442, II:443–446, II:444^t
- The Internal Measurement Approach (BIS), III:100n
- Theoretical value, determination of, III:10–11
- Théorie de la Spéculation (The Theory of Speculation)* (Bachelier), II:121–122, II:469
- Theory of point processes, II:470–471
- Three Mile Island power plant crisis, II:51–52
- Three-stage growth model, II:9–10
- Threshold autoregressive (TAR) models, II:404–408
- Thresholds, II:746–747
- Through the cycle, defined, I:302–303, I:309–310
- Thurstone, Louis Leon, II:154
- Tick data. *See* high-frequency data (HFD)
- Time
 in differential equations, II:643–644
 physical *vs.* intrinsic scales of, II:742
 use of for financial data, II:546–547
- Time aggregation, II:369
- Time decay, I:509, I:513, I:521^f
- Time dependency, capture of, II:362–363
- Time discretization, II:666, II:679
- Time increments
 models of, I:79
 in parameter estimation, I:83
- Time intervals, size of, II:300–301
- Time lags, II:299–300
- Time points, spacing of, II:501
- Time premiums, I:485
- Time series
 autocorrelation of, II:331
 causal, II:504
 concepts of, II:501–503
 continuity of, I:80
 defined, II:501–502, II:519
 fractal nature of, III:480
 importance of, II:360
 multivariate, II:502
 stationary, II:502
 stationary/nonstationary, II:299
 for stock prices, II:296
- Time to expiry, I:513
- Time value, I:513, I:513^f, II:595–596
- TIPS (Treasury inflation-protected securities)
 and after-tax inflation risk, I:287
 apparent real yield premium, I:293^f
 effect of inflation and flexible price CPI, I:292^f
 features of, I:277
 and flexible price CPI, I:291^f
 and inflation, I:290, I:294
 performance link with short-term inflation, I:291–292
 real yields on, I:278
 spread to nominal yield curve, I:281^f
 volatility of, I:288–290, I:294
vs. real yield, I:293–294
 10-year data, I:279–280
 yield of, I:284
 yields from, I:278
- TLF model, strengths of, III:388–389
- Total asset turnover ratio, II:558
- Total return reports, II:237^t
- Total return swaps, I:540–542, I:541–542
- Trace test statistic, II:392
- Tracking error
 actual *vs.* predicted, II:69
 alternate definitions of, II:67–68
 defined, II:115, II:119
 estimates of future, II:69
 as measure of consistency, II:99–100
 reduction of, II:262–263
 standard definition, II:67
 with TIPS, I:293
- Tracking error volatility (TEV). *See* TEV (tracking error volatility)
- Trade optimizers, role of, II:116–117
- Trades
 amount needed for market impact, III:624
 cash-and-carry, I:487
 crossing of, II:75
 importance of execution of, III:623, III:631
 measurement of size, III:628
 in portfolio construction, II:104, II:116–117
 round-trip time of, II:451
 size effects of, III:372, III:630
 speed of, II:105
 timing of, III:628–629
- Trading costs, II:118, III:627–628, III:631–632
- Trading gains, defined, I:122, I:123
- Trading horizons, extending, III:624
- Trading lists, II:289^t
- Trading strategies
 backtesting of, II:236–237
 categories of, II:195
 in continuous-state, continuous-time, I:122
 development of factor-based, II:197–198, II:211
 factor-based, II:195, II:232–235
 factor weights in, II:233^f
 in multiperiod settings, I:105
 risk to, II:198–200
 self-financing, I:126–127, I:136
- Trading venues, electronic, II:57
- Training windows, moving, II:713–714
- Tranches, III:38, III:39^t, III:45
- Transaction costs
 in backtesting, II:235
 in benchmarking, II:67
 components of, II:119
 consideration of, II:64, II:85–86n
 dimensions of, III:631
 effect of, I:483
 figuring, II:85n
 fixed, II:72–73
 forecasting of, II:113–114
 incorporation of, II:69–73, II:84
 international, III:629
 linear, II:70
 and liquidity, III:624–625
 managing, III:146
 measurement of, III:626
 piecewise-linear, II:70–72, II:71^f

- quadratic, *II:72*
in risk modeling, *II:693*
types of, *III:623*
- Transformations, nonlinear,
III:630–631
- Transition probabilities, *I:368, I:381t*
- Treasuries
correlations of, *III:405t*
covariance matrix of, *III:406t*
curve risk, *II:277t*
discount function for, *III:564–565*
futures, *I:482*
inflation-indexed, *I:286*
movements of, *III:403f*
on-the-run, *I:227, III:7, III:560*
par yield curve, *I:218t*
spot rates, *I:220*
3-month, *II:415–416, II:416f*
volatility of, *III:404–406, III:406t*
- Treasury bill rates, weekly data, *I:89f*
- Treasury inflation-protected securities (TIPS). *See* TIPS (Treasury inflation-protected securities)
- Treasury Regulation T (Reg T), *I:67*
- Treasury securities, *I:210–211*
comparable, defined, *III:5*
in futures contracts, *I:483*
hypothetical, illustration of duration/convexity, *III:308–310, III:308t*
maturities of, *I:226*
options on, *I:490*
par rates for, *I:217*
prediction of 10-year yield, *II:322–328*
valuation of, *I:216*
yield of, *II:324–327t*
- Treasury strips, *I:218t, I:220–221, I:286, III:560*
- Treasury yield curves, *I:226, III:561*
- Trees/lattices
adjusted to current market price, *I:496f*
bushy trees, *I:265, I:266f*
calibrated, *I:495*
convertible bond value, *I:274–275*
extended pricing tree, *III:23f*
from historical data, *III:131f*
pruning of, *II:377*
stock price, *I:274*
three-period scenario, *III:131f*
trinomial, *I:81, I:273, I:495–496*
use of in modeling, *I:494–496*
- Trees/lattices, binomial
building of, *I:273*
for convertible bonds, *I:275f*
discussion of, *I:80–81*
interest rate, *I:244*
model of, *I:273–275*
- stock price model, *III:173*
- term structure evolution, *I:495f*
use of, *I:114–115, I:114f*
- Trends
deterministic, *II:383*
in financial time series, *II:504*
and integrated series, *II:512–514*
stochastic, *II:383, II:384*
- Treynor-Black model, *I:203n*
- Trinomial stochastic models, *II:11–12*
- Truncated Lévy flight (TLF), *III:382, III:384–386*
IDD in, *III:386*
time scaling of, *III:385f*
- Truncation, *III:385–386*
- Truth in Savings Act, *II:615*
- T*-statistic, *II:240n, II:336, II:350, II:390*
- Tuple, defined, *III:157*
- Turnover
assessment of, *III:68*
defined, *III:66*
in MBSs, *III:48*
in portfolios, *II:234, II:235*
- Two beta trap, *I:74–77*
- Two-factor models, *III:553–554*
- Two-stage growth model, *II:9*
- U.K. index-linked gilts, tax treatment of, *I:287*
- Uncertainties
and Bayesian statistics, *I:140*
in measurement processes, *II:367*
modeling of, *II:306, III:124, III:131–132*
and model risk, *II:729*
quantification of, *I:101*
representation of, *III:128*
time behavior of, *II:359*
- Uncertainty sets
effect of size of, *III:143*
in portfolio allocation, *II:80*
selection of, *III:140–141*
structured, *III:143–144*
in three dimensions, *II:81f*
use of, *III:138, III:140*
- Uncertain volatility model, *II:673–674*
- Underperformance, finding reasons for, *II:118*
- Underwater, on homeowner's equity, *III:73*
- Unemployment rate
as an economic measure, *II:398*
application of TAR models to, *II:405–406*
characteristics of series, *II:430*
forecasts from, *II:433*
performance of forecasting, *II:432–433, II:432t*
and risk, *II:292n*
- test of nonlinearity, *II:431, II:431t*
- time plot of, *II:406f, II:430f*
- Uniqueness, theorem of, *III:490*
- Unit root series, *II:385*
- Univariate linear regression model, *I:163–170*
- Univariate stationary series, *II:504*
- U.S. Bankruptcy Code. *See also* bankruptcy
Chapter 7, *I:350*
Chapter 11, *I:342, I:350*
Utility, *I:56, II:469, II:471, II:719–720*
- Validation, out of sample, *II:711*
- Valuation
arbitrage-free, *I:216–217, I:220–222, I:221t*
and cash flows, *I:223*
defined, *I:209*
effect of business cycle on, *I:303–304*
fundamental principle of, *I:209*
with Monte Carlo simulation, *III:6–12*
of natural gas/oil storage, *I:560–561*
of non-Treasury securities, *I:222–223*
relative, *I:225, II:34–40, II:44–45*
risk-neutral, *I:557, III:595–596, III:601*
total firm, *II:21–23*
uncertainty in, *II:15*
use of lattices for, *I:240*
- Value
absolute *vs.* relative basis of, *I:259–260*
analysis of relative, *I:225*
arbitrage-free, *I:221*
book *vs.* market of firms, *II:559–560*
determining present, *II:600–601*
formulas for analysis of, *II:238–239*
identification of relative, *I:405*
intrinsic, *I:484–485*
present, discounted, *II:601f*
relative, *I:405, II:37–38*
vs. price, *I:455n*
- Value at risk (VaR). *See also* CVaR (credit value at risk)
in backtesting, *II:748*
backtesting of, *II:749f, III:325–327, III:365–367*
boxplot of, *III:325f*
and coherent risk measures, *III:329*
conditional, *III:332, III:355–356, III:382*
deficiencies in, *I:407, III:321, III:331–332, III:347*
defined, *II:754n, III:319–322*
density and distribution functions, *III:320f*

- Value at risk (VaR) (*Continued*)
determining from simulation, III:639f
distribution-free confidence intervals for, III:292–293
estimation of, II:366, III:289–290, III:373–376, III:644, III:644t
exceedances of, III:325–326
IDD in, III:290
interest rate covariance matrix in, III:403
levels of confidence with, III:290–291
liquidity-adjusted, III:374, III:376
in low market volatility, II:748
measurements by, II:354
methods of computation, III:323
modeling of, II:130–131, III:375–376
and model risk, II:695
normal against confidence level, III:294f
portfolio problem, I:193
in practice, III:321–325
relative spreads between predictions, II:750f, II:751f, II:752f
as safety-first risk measure, III:355
standard normal distribution of, III:324t
use of, II:365
vs. deviation measures, III:320–321
- Value of operations, process for finding, II:30t
- Values, lagged, II:130
- Van der Korput sequences, III:650
- Variables
antithetic, III:647–648
application of macro, II:193n
behavior of, III:152–153
categorical, II:333–334, II:350
classification, II:176
declaration of in VBL, III:457–458
dependence between, II:306–307
dependent categorical, II:348–350
dependent/independent in CAPM, I:67
dichotomous, II:350
dummy, II:334
exogenous *vs.* endogenous, II:692
fat-tailed, III:280
independent and identically distributed, II:125
independent categorical, II:333–348
interactions between, II:378
large numbers of, II:147
macroeconomic, II:54–55, II:177
in maximum likelihood calculations, II:312–313
mixing of categorical and quantitative, II:334–335
nonstationary, II:388–393
as observation or measurement, II:306
random, I:159n
in regression analysis, II:330
separable, II:647
slope, III:553
split formation of, III:130f
spread, II:336
standardization of, II:205
stationary, II:385, II:386
stationary/nonstationary, II:384–386
stochastic, III:159–164
use of dummy, II:335, II:343–344
- Variables, random, II:297
 α -stable, III:242–244, III:244–245
Bernoulli, III:169
continuous, III:200–201, III:205–206
on countable spaces, III:160–161, III:166
defined, III:162
discrete, III:165
infinitely divisible, III:253
in probability, III:159–164
sequences of, I:389
on uncountable spaces, III:161–162
use of, I:82
- Variance gamma process, III:499, III:504
- Variance matrix, II:370–371
- Variances
addressing inequality of, I:168
based on covariance matrix, II:161t, II:163t, II:164f
conditional, I:180
conditional/unconditional, II:361
in dispersion parameters, III:202–203
equal, I:164
as measure of risk, I:8
in probability, III:167–169
reduction in, III:647–651
unequal, I:167–168, I:172
- Variances/covariances, II:112–113, II:302–303, III:395–396
- Variance swaps, I:545–547, I:549, I:552
- Variational formulation, and finite element space, II:670–672
- Variation margins, I:478
- Vasicek model
with change of time, III:523–524
for coupon-bond call options, I:501–502
distribution of, I:493
in history, I:491
for short rates, III:545–546
use of, I:89, I:497
valuing zero-coupon bond calls with, I:499–500
- VBA (Visual Basic for Applications)
built-in numeric functions of, III:456
comments in, III:453
control flow statements, III:458–460
debugging in, III:461
debugging tools of, III:461, III:477
example programs, III:449–452, III:461–466
in Excel, III:449, III:450f
FactorialFun1, III:455–456
functions, user-defined, III:463f
functions in, III:477
generating Brownian motion paths in, III:463–465
If statements, III:459
For loops, III:458–459
methods (actions) in, III:452–453
modules, defined, III:455
as object-oriented language, III:452, III:466
objects in, III:452
operators in, III:459–460
Option Explicit command, III:458
pricing European call options, III:465–466
programing of input dialog boxes, III:460–461
programming tips for, III:454–461
properties in, III:453
random numbers in, III:464–465
subroutines and user-defined functions in, III:466–477
subroutines *vs.* user-defined functions in, III:455–457
use of Option Explicit command, III:458
user-defined functions, III:463f
user interaction with, III:460–461
variable declaration in, III:457–458
With/End structure in, III:453–454
writing code in, III:453–454
- Vech notation, II:371–372
- VEC model, II:372
- Vector autoregressive (VAR) model, II:393
- Vectors, II:621–622, II:625–626, II:628
- Vega, I:521
- Vichara Technology, III:41–42, III:43t
- Visual Basic for Applications (VBA). *See* VBA
- Volatilities
absolute *vs.* relative, III:404–405
actual, I:514
aim of models of, I:176
analysis of, II:270–272
and ARCH models, II:409

- assumptions about, *III:7*
 calculation of, *II:272, III:534t*
 calculation of daily, *III:533–534*
 calibration of local, *II:681–685*
 clustering of, *II:359, II:716, III:402*
 confidence intervals for, *III:399–400*
 constant, *III:653*
 decisions for measuring, *III:403–404*
 defined, *III:533, III:653*
 with different mean reversions,
III:538f
 of the diffusion, *I:125*
 effect of local, *III:609*
 effect on hedging, *I:517–518*
 of energy commodities, *I:556–557*
 estimation of, *II:368–369*
 in EWMA estimates, *III:410–411*
 exposure to, *II:252f, II:252t*
 forecasts of, *I:179–180, II:172,*
II:367–368
 in FTSE 100, *III:412–413*
 historical, *I:513, III:534, III:654*
 hypothetical modelers of, *III:408*
 implied, *I:513–514, II:282, II:662,*
III:654
 in interest rate structure models,
I:492
 jump-diffusion, *III:657*
 level-dependent, *III:654–655,*
III:656
 local, *II:681, II:682–683, III:655*
 as a measure, *I:545, II:373*
 measurement of, *I:393, III:403–406*
 minimization of, *II:179*
 in models, *II:302*
 models of, *II:428*
 in option pricing, *I:513–514*
 patterns in, *I:395*
 in random walks, *I:84*
 and risk, *II:270*
 in risk-neutral measures, *III:587*
 smile of, *III:557*
 and the smoothing constant,
III:409–410
 states of, *I:180–181*
 stochastic, *I:94, I:547, I:548,*
III:655–658, III:656, III:658
 stochastic models, *II:681*
 time increments of, *I:83*
 of time series, *I:80*
 time-varying, *II:733–734*
 types of, *III:658*
 vs. annual standard deviation,
III:534
 Volatility clustering, *III:242, III:388*
 Volatility curves, *III:534–535,*
III:535t
- Volatility measures, nonstochastic,
III:654–655
 Volatility multiples, use of,
III:536
 Volatility risk, *I:509*
 Volatility skew, *III:550, III:551f,*
III:555–556, III:654
 measuring, *III:550*
 Volatility smile, *II:681, III:555–557,*
III:556f, III:654, III:656
 Volatility swaps, *I:545–547, I:552*
 for S&P Canada index (example),
I:550–552
 valuing of, *I:549*
 Volume-weighted average price
 (VWAP), *II:117, III:626–627*
 VPRs (voluntary prepayment rates)
 calculation of, *III:76*
 in cash flow calculators, *III:34*
 defined, *III:30*
 impacts of, *III:38*
- W. T. Grant, cash flows of, *II:576*
 Waldrop, Mitchell, *II:699*
 Wal-Mart, *II:569, II:570f*
 Walras, Leon, *II:467, II:468–469,*
II:474
 Waterfalls, development of, *III:8*
 Weak laws of large numbers (WLLN),
III:263
 Wealth, *I:460t, III:130*
 Weather, as chaotic system, *II:653*
 Weibull density, *III:107f*
 Weibull distributions, *III:106–107,*
III:112, III:229, III:262, III:265,
III:267, III:268
 Weighting, efficient, *I:41–42*
 Weights, *II:115, II:185t, II:231–232,*
II:724
 Weirton Steel, cash flows of,
II:577f
 What's the hedge, *I:300, I:303, I:306,*
I:417. See also hedge test
 White noise. *See noise, white*
 Wiener processes, *I:95, I:491, I:497,*
III:534–535, III:579, III:581
 Wilson, Kenneth, *II:480*
 Wind farms, valuation of, *I:563–564*
 Wold representation, *II:506*
 Working capital, *II:551*
 concept of, *II:567*
- XML (eXtensible Markup Language),
 development of, *II:482*
- Yield and bond loss matrix, *III:41f*
 Yield curve risk, *III:307, III:316–317*
- Yield curves
 horizon, *III:585*
 initial consistency with, *III:544*
 issuer par, *I:238t, I:244t*
 nonparallel, *III:309–310*
 parallel shifts in, *III:308–309*
 par-coupon, *III:585*
 reshaping duration, *III:315–316*
 in scenario analysis, *II:290*
 SEDUR/LEDUR, *III:316, III:317*
 shifts in, *III:586*
 slope of, *III:315*
 in term structures, *III:560*
 in valuation, *I:235*
- Yields
 calculation of, *II:613–618*
 comparison across countries, *I:226*
 dividend, *II:4*
 on investments, *II:617–618, II:619*
 loss-adjusted, *III:36, III:40*
 and loss matrix analysis, *III:40–41*
 projected, *III:37f, III:38f*
 real, *I:278–280, I:280f*
 rolling, *I:258–259*
- Yield spreads
 computation of, *I:226*
 determining, *I:373–374*
 for different rating grades, *I:374t*
 in Merton model, *I:305–306*
 over swap and treasury curves,
I:226–227
- Zero-coupon bonds
 assumptions about, *I:261*
 calculations using CIR model, *I:502t*
 calculations using Vasicek model,
I:502t
 defaultable, *I:317, I:335n*
 default-free, *I:318*
 development of valuation model
 for, *III:582–583*
 equations for, *III:554*
 future market price for, *I:492–493*
 lattices for, *I:266f*
 market for, *I:264*
 and martingales, *I:262*
 PDEs of, *I:268–269n*
 pricing of, *I:316*
 term structure model for, *III:584*
 value of, *III:572–573*
 valuing, *I:213, I:499–501, I:499t*
 Zero coupon rates, *III:546–547*
 Zero coupon securities, *I:218*
 Zero one distribution, *III:169–170*
 Zero volatility spread, *III:11–12*
 Zipf's law, *III:263, III:269*
 Z-scores, *II:191, II:240n*



ENCYCLOPEDIA
OF
FINANCIAL MODELS

III

FRANK J. FABOZZI, EDITOR

ENCYCLOPEDIA
OF
FINANCIAL MODELS

Volume III

ENCYCLOPEDIA
OF
FINANCIAL MODELS

Volume III

FRANK J. FABOZZI, EDITOR



WILEY

John Wiley & Sons, Inc.

Copyright © 2013 by Frank J. Fabozzi. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books. For more information about Wiley products, visit our web site at www.wiley.com.

ISBN: 978-1-118-00673-3 (3 v. set : cloth)

ISBN: 978-1-118-010327 (v. 1)

ISBN: 978-1-118-010334 (v. 2)

ISBN: 978-1-118-010341 (v. 3)

ISBN: 978-1-118-182365 (ebk.)

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

About the Editor

Frank J. Fabozzi is Professor of Finance at EDHEC Business School and a member of the EDHEC Risk Institute. Prior to joining EDHEC in August 2011, he held various professorial positions in finance at Yale University's School of Management from 1994 to 2011 and from 1986 to 1992 was a visiting professor of finance and accounting at MIT's Sloan School of Management. From 2008 to 2011, he was an affiliated professor in the Institute of Statistics, Econometrics, and Mathematical Finance at the University of Karlsruhe in Germany. Prior to 1986 he held professorial positions at Lafayette College, Fordham University, Queens College (CUNY), and Hofstra University. From 2003 to 2011, he served on Princeton University's Advisory Council for the Department of Operations Research and Financial Engineering and since then has been a visiting fellow in that department.

Professor Fabozzi is the editor of the *Journal of Portfolio Management*, as well as on the editorial board of the *Journal of Fixed Income*, *Journal of Asset Management*, *Quantitative Finance*, *Review of Futures Markets*, *Journal of Mathematical Finance*, *Journal of Structured Finance*, *Annals of Financial Economics*, and *Theoretical Economic Letters*.

He has authored and edited a number of books in asset management and quantitative finance. His coauthored books in quantitative finance include *A Probability Metrics Approach to Financial Risk Measures* (2011), *Financial Modeling with Lévy Processes and Volatility Clustering* (2011), *Quantitative Equity Investing: Techniques and Strategies* (2010), *Probability and Statistics for Finance* (2010), *Simulation and Optimization Modeling in Finance* (2010), *Bayesian Methods in Finance* (2008), *Advanced Stochastic Models, Risk Assessment, and Portfolio Optimization: The Ideal Risk* (2008), *Financial Econometrics: From Basics to Advanced Modeling Techniques* (2007), *Robust Portfolio Optimization and Management* (2007), and *Mathematics of Financial Modeling and Investment Management* (2004). His books in applied mathematics include *The Methods of Distances in the Theory of Probability and Statistics* (2013) and *Robust and Non-Robust Models in Statistics* (2009). He coauthored three monographs for the Research Foundation of the CFA Institute: *The Impact of the Financial Crisis on the Asset Management Industry* (2010), *Challenges in Quantitative Equity Management* (2008), and *Trends in Quantitative Finance* (2006).

Professor Fabozzi's research papers have appeared in numerous journals, including *Journal of Finance*, *Journal of Finance and Quantitative Analysis*, *Econometric Theory*, *Operations Research*, *Journal of Banking and Finance*, *Journal of Economic Dynamics and Control*, *Studies in Nonlinear Dynamics and Econometrics*, *European Journal of Operational Research*, *Annals of Operations Research*, *Quantitative Finance*, *European Financial Management*, and *The Econometric Journal*. His 2010 article published in *European Financial Management* with Professors Robert Shiller, and Radu Tunaru, "Property Derivatives for Managing European Real-Estate Risk," received the Best Paper Award and his paper with the same coauthors entitled "A Pricing Framework for Real Estate Derivatives" was awarded

Best Research Paper at the 10th Research Conference Campus for Finance held annually at WHU Otto Beisheim School of Management, Vallendar, Germany. An article coauthored with Dr. Sergio Focardi, "An Autoregressive Conditional Duration Model of Credit Risk Contagion," published in 2005 in *Journal of Risk Finance* was the winner of the 2006 Outstanding Paper by Emerald Literati Network.

He has received several awards and honors for his body of work. In 1994 he was awarded an Honorary Doctorate of Humane Letters from Nova Southeastern University. In 2002 he was inducted into the Fixed Income Analysts Society's Hall of Fame, established by the society "to recognize the lifetime achievements of outstanding practitioners in the advancement of the analysis of fixed-income securities and portfolios." In 2007 he was the recipient of the C. Stewart Sheppard Award given by the CFA Institute "in recognition of outstanding contribution to continuing education in the CFA profession." He was the cover story in the July 1999 issue of *Bloomberg Magazine* entitled "The Boswell of Bonds."

Professor Fabozzi was the co-founder of Information Management Network (now a subsidiary of Euromoney), a conference company specializing in financial topics. He is a trustee for the BlackRock family of closed-end funds where he is the chair of the performance committee and a member of the audit committee. He was a director of Guardian Mutual Funds and Guardian Annuity Funds.

He earned both an M.A. and B.A. in economics and statistics in June 1970 from the City College of New York and elected to Phi Beta Kappa in 1969. He earned a Ph.D. in Economics in September 1972 from the City University of New York. Professor Fabozzi holds two professional designations: Chartered Financial Analyst (1977) and Certified Public Accountant (1982).

Contents

Contributors	xi		
Preface	xvii		
Guide to the <i>Encyclopedia of Financial Models</i>	xxxiii		
Index	661		
Volume I			
Asset Allocation	1		
Mean-Variance Model for Portfolio Selection	3		
Principles of Optimization for Portfolio Selection	21		
Asset Allocation and Portfolio Construction Techniques in Designing the Performance-Seeking Portfolio	35		
Asset Pricing Models	47		
General Principles of Asset Pricing	49		
Capital Asset Pricing Models	65		
Modeling Asset Price Dynamics	79		
Arbitrage Pricing: Finite-State Models	99		
Arbitrage Pricing: Continuous-State, Continuous-Time Models	121		
Bayesian Analysis and Financial Modeling Applications	137		
Basic Principles of Bayesian Analysis	139		
Introduction to Bayesian Inference	151		
Bayesian Linear Regression Model	163		
Bayesian Estimation of ARCH-Type Volatility Models	175		
Bayesian Techniques and the Black-Litterman Model	189		
Bond Valuation		207	
Basics of Bond Valuation		209	
Relative Value Analysis of Fixed-Income Products		225	
Yield Curves and Valuation Lattices		235	
Using the Lattice Model to Value Bonds with Embedded Options, Floaters, Option, and Caps/Floors		243	
Understanding the Building Blocks for OAS Models		257	
Quantitative Models to Value Convertible Bonds		271	
Quantitative Approaches to Inflation-Indexed Bonds		277	
Credit Risk Modeling		297	
An Introduction to Credit Risk Models		299	
Default Correlation in Intensity Models for Credit Risk Modeling		313	
Structural Models in Credit Risk Modeling		341	
Modeling Portfolio Credit Risk		361	
Simulating the Credit Loss Distribution		377	
Managing Credit Spread Risk Using Duration Times Spread (DTS)		391	
Credit Spread Decomposition		401	
Credit Derivatives and Hedging Credit Risk		407	
Derivatives Valuation		421	
No-Arbitrage Price Relations for Forwards, Futures, and Swaps		423	
No-Arbitrage Price Relations for Options		437	
Introduction to Contingent Claims Analysis		457	
Black-Scholes Option Pricing Model		465	

Pricing of Futures/Forwards and Options	477	Classification and Regression Trees and Their Use in Financial Modeling	375
Pricing Options on Interest Rate Instruments	489	Applying Cointegration to Problems in Finance	383
Basics of Currency Option Pricing Models	507	Nonlinearity and Nonlinear Econometric Models in Finance	401
Credit Default Swap Valuation	525	Robust Estimates of Betas and Correlations	437
Valuation of Fixed Income Total Return Swaps	541	Working with High-Frequency Data	449
Pricing of Variance, Volatility, Covariance, and Correlation Swaps	545	Financial Modeling Principles	465
Modeling, Pricing, and Risk Management of Assets and Derivatives in Energy and Shipping	555	Milestones in Financial Modeling	467
		From Art to Financial Modeling	479
		Basic Data Description for Financial Modeling and Analysis	485
		Time Series Concepts, Representations, and Models	501
		Extracting Risk-Neutral Density Information from Options Market Prices	521
		Financial Statement Analysis	529
		Financial Statements	531
		Financial Ratio Analysis	545
		Cash-Flow Analysis	565
		Finite Mathematics for Financial Modeling	579
		Important Functions and Their Features	581
		Time Value of Money	595
		Fundamentals of Matrix Algebra	621
		Difference Equations	629
		Differential Equations	643
		Partial Differential Equations in Finance	659
		Model Risk and Selection	689
		Model Risk	691
		Model Selection and Its Pitfalls	699
		Managing the Model Risk with the Methods of the Probabilistic Decision Theory	719
		Fat-Tailed Models for Risk Estimation	731
		Volume III	
		Mortgage-Backed Securities Analysis and Valuation	1
		Valuing Mortgage-Backed and Asset-Backed Securities	3
		The Active-Passive Decomposition Model for MBS	17
		Analysis of Nonagency Mortgage-Backed Securities	29
Volume II			
Equity Models and Valuation	1		
Dividend Discount Models	3		
Discounted Cash Flow Methods for Equity Valuation	15		
Relative Valuation Methods for Equity Analysis	33		
Equity Analysis in a Complex Market	47		
Equity Portfolio Selection Models in Practice	61		
Basics of Quantitative Equity Investing	89		
Quantitative Equity Portfolio Management	107		
Forecasting Stock Returns	121		
Factor Models for Portfolio Construction	135		
Factor Models	137		
Principal Components Analysis and Factor Analysis	153		
Multifactor Equity Risk Models and Their Applications	171		
Factor-Based Equity Portfolio Construction and Analysis	195		
Cross-Sectional Factor-Based Models and Trading Strategies	213		
The Fundamentals of Fundamental Factor Models	243		
Multifactor Equity Risk Models and Their Applications	255		
Multifactor Fixed Income Risk Models and Their Applications	267		
Financial Econometrics	293		
Scope and Methods of Financial Econometrics	295		
Regression Analysis: Theory and Estimation	305		
Categorical and Dummy Variables in Regression Models	333		
Quantile Regression	353		
ARCH/GARCH Models in Applied Financial Econometrics	359		

Measurements of Prepayments for Residential Mortgage-Backed Securities	47	Back-Testing Market Risk Models	361
Prepayments and Factors Influencing the Return of Principal for Residential Mortgage-Backed Securities	65	Estimating Liquidity Risks	371
Operational Risk	79	Estimate of Downside Risk with Fat-Tailed and Skewed Models	381
Operational Risk	81	Moving Average Models for Volatility and Correlation, and Covariance Matrices	395
Operational Risk Models	91	Software for Financial Modeling	415
Modeling Operational Loss Distributions	103	Introduction to Financial Model Building with MATLAB	417
Optimization Tools	121	Introduction to Visual Basic for Applications	449
Introduction to Stochastic Programming and Its Applications to Finance	123	Stochastic Processes and Tools	469
Robust Portfolio Optimization	137	Stochastic Integrals	471
Probability Theory	149	Stochastic Differential Equations	485
Concepts of Probability Theory	151	Stochastic Processes in Continuous Time	495
Discrete Probability Distributions	165	Conditional Expectation and Change of Measure	507
Continuous Probability Distributions	195	Change of Time Methods	519
Continuous Probability Distributions with Appealing Statistical Properties	207	Term Structure Modeling	531
Continuous Probability Distributions Dealing with Extreme Events	227	The Concept and Measures of Interest Rate Volatility	533
Stable and Tempered Stable Distributions	241	Short-Rate Term Structure Models	543
Fat Tails, Scaling, and Stable Laws	259	Static Term Structure Modeling in Discrete and Continuous Time	559
Copulas	283	The Dynamic Term Structure Model	575
Applications of Order Statistics to Risk Management Problems	289	Essential Classes of Interest Rate Models and Their Use	593
Risk Measures	297	A Review of No Arbitrage Interest Rate Models	603
Measuring Interest Rate Risk: Effective Duration and Convexity	299	Trading Cost Models	621
Yield Curve Risk Measures	307	Modeling Market Impact Costs	623
Value-at-Risk	319	Volatility	635
Average Value-at-Risk	331	Monte Carlo Simulation in Finance	637
Risk Measures and Portfolio Selection	349	Stochastic Volatility	653

Contributors

Yves Achdou, PhD

Professor, Lab. Jacques-Louis Lions, University Paris-Diderot, Paris, France

Irene Aldridge

Managing Partner, Able Alpha Trading

Carol Alexander, PhD

Professor of Finance, University of Sussex

Andrew Alford, PhD

Managing Director, Quantitative Investment Strategies, Goldman Sachs Asset Management

Noël Amenc, PhD

Professor of Finance, EDHEC Business School, Director, EDHEC-Risk Institute

Bala Arshanapalli, PhD

Professor of Finance, Indiana University Northwest

David Audley, PhD

Senior Lecturer, The Johns Hopkins University

Jennifer Bender, PhD

Vice President, MSCI

William S. Berliner

Executive Vice President, Manhattan Advisory Services Inc.

Anand K. Bhattacharya, PhD

Professor of Finance Practice, Department of Finance, W. P. Carey School of Business, Arizona State University

Michele Leonardo Bianchi, PhD

Research Analyst, Specialized Intermediaries Supervision Department, Bank of Italy

Olivier Bokanowski

Associate Professor, Lab. Jacques-Louis Lions, University Paris-Diderot, Paris, France

Gerald W. Buetow Jr., PhD, CFA

President and Founder, BFRC Services, LLC

Paul Bukowski, CFA, FCAS

Executive President, Head of Equities, Hartford Investment Management

Joseph A. Cerniglia

Visiting Researcher, Courant Institute of Mathematical Sciences, New York University

Ren-Raw Chen

Professor of Finance, Graduate School of Business, Fordham University

Anna Chernobai, PhD

Assistant Professor of Finance, M. J. Whitman School of Management, Syracuse University

Richard Chin

Investment Manager, New York Life Investments

António Baldaque da Silva
Managing Director, Barclays

Siddhartha G. Dastidar, PhD, CFA
Vice President, Barclays

Arik Ben Dor, PhD
Managing Director, Barclays

Michael Dorigan, PhD
Senior Quantitative Analyst, PNC Capital
Advisors

Kevin Dowd, PhD
Partner, Cobden Partners, London

Pamela P. Drake, PhD, CFA
J. Gray Ferguson Professor of Finance, College
of Business, James Madison University

Lev Dynkin, PhD
Managing Director, Barclays

Brian Eales
Academic Leader (Retired), London Metropolitan
University

Abel Elizalde, PhD
Credit Derivatives Strategy, J.P. Morgan

Robert F. Engle, PhD
Michael Armellino Professorship in the Man-
agement of Financial Services and Director of
the Volatility Institute, Leonard N. Stern School
of Business, New York University

Frank J. Fabozzi, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

Peter Fitton
Manager, Scientific Development, CreditXpert
Inc.

Sergio M. Focardi, PhD
Partner, The Intertek Group

Radu Găbudean, PhD
Vice President, Barclays

Vacslav S. Glukhov, PhD
Head of Quantitative Strategies and Data Ana-
lytics, Liquidnet Europe Ltd, London, United
Kingdom

Felix Goltz, PhD
Head of Applied Research, EDHEC-Risk
Institute

Chris Gowlland, CFA
Senior Quantitative Analyst, Delaware Invest-
ments

Biliana S. Güner
Assistant Professor of Statistics and Economet-
rics, Özyeğin University, Turkey

Francis Gupta, PhD
Director, Index Research & Design, Dow Jones
Indexes

Markus Höchstötter, PhD
Assistant Professor, University of Karlsruhe

John S. J. Hsu, PhD
Professor of Statistics and Applied Probability,
University of California, Santa Barbara

Jay Hyman, PhD
Managing Director, Barclays, Tel Aviv

Bruce I. Jacobs, PhD
Principal, Jacobs Levy Equity Management

Robert R. Johnson, PhD, CFA
Independent Financial Consultant,
Charlottesville, VA

Frank J. Jones, PhD
Professor, Accounting and Finance Depart-
ment, San Jose State University and Chairman,
Investment Committee, Private Ocean Wealth
Management

Robert Jones, CFA
Chairman, Arwen Advisors, and Chairman and
CIO, Systems Two Advisors

Andrew Kalotay, PhD

President, Andrew Kalotay Associates

Young Shin Kim, PhD

Research Assistant Professor, School of Economics and Business Engineering, University of Karlsruhe and KIT

Petter N. Kolm, PhD

Director of the Mathematics in Finance Masters Program and Clinical Associate Professor, Courant Institute of Mathematical Sciences, New York University

Glen A. Larsen Jr., PhD CFA

Professor of Finance, Indiana University Kelley School of Business—Indianapolis

Anthony Lazanas

Managing Director, Barclays

Arturo Leccadito, PhD

Business Administration Department, Università della Calabria

Tony Lelièvre, PhD

Professor, CERMICS, Ecole des Ponts Paristech, Marne-la-Vallée, France

Alexander Levin, PhD

Director, Financial Engineering, Andrew Davidson & Co., Inc.

Kenneth N. Levy, CFA

Principal, Jacobs Levy Equity Management

Terence Lim, PhD, CFA

CEO, Arwen Advisors

Peter C. L. Lin

PhD Candidate, The Johns Hopkins University

Steven V. Mann, PhD

Professor of Finance, Moore School of Business, University of South Carolina

Harry M. Markowitz, PhD

Consultant and Nobel Prize Winner, Economics, 1990

Lionel Martellini, PhD

Professor of Finance, EDHEC Business School, Scientific Director, EDHEC-Risk Institute

James F. McNatt, CFA

Executive Vice President, ValueWealth Services

Christian Menn, Dr Rer Pol

Managing Partner, RIVACON

Ivan Mitov

Head of Quantitative Research, FinAnalytica

Edwin H. Neave

Professor Emeritus, School of Business, Queen's University, Kingston, Ontario

William Nelson, PhD

Professor of Finance, Indiana University Northwest

Frank Nielsen

Managing Director of Quantitative Research, Fidelity Investments - Global Asset Allocation

Philip O. Obazee

Senior Vice President and Head of Derivatives, Delaware Investments

Dominic O'Kane, PhD

Affiliated Professor of Finance, EDHEC Business School, Nice, France

Dessislava A. Pachamanova

Associate Professor of Operations Research, Babson College

Bruce D. Phelps

Managing Director, Barclays

Thomas K. Philips, PhD

Regional Head of Investment Risk and Performance, BNP Paribas Investment Partners

David Philpotts

QEP Global Equities, Schroder Investment Management, Sydney, Australia

Wesley Phoa

Senior Vice President, Capital International Research, Inc.

Svetlozar T. Rachev, PhD Dr Sci

Frey Family Foundation Chair Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

Boryana Racheva-Yotova, PhD

President, FinAnalytica

Shrikant Ramanmurthy

Consultant, New York, NY

Srichander Ramaswamy, PhD

Senior Economist, Bank for International Settlements, Basel, Switzerland

Patrice Retkowsky

Senior Research Engineer, EDHEC-Risk Institute

Paul Sclavounos

Department of Mechanical Engineering, Massachusetts Institute of Technology

Shani Shamah

Consultant, RBC Capital Markets

Koray D. Simsek, PhD

Associate Professor, Sabanci School of Management, Sabanci University

James Sochacki

Professor of Applied Mathematics, James Madison University

Arne D. Staal

Director, Barclays

Maxwell J. Stevenson, PhD

Discipline of Finance, Business School, University of Sydney, Australia

Filippo Stefanini

Head of Hedge Funds and Manager Selection, Eurizon Capital SGR

Stoyan V. Stoyanov, PhD

Professor of Finance at EDHEC Business School and Head of Research for EDHEC Risk Institute-Asia

Anatoliy Swishchuk, PhD

Professor of Mathematics and Statistics, University of Calgary

Ruey S. Tsay, PhD

H.G.B. Alexander Professor of Econometrics and Statistics, University of Chicago Booth School of Business

Radu S. Tunaru

Professor of Quantitative Finance, Business School, University of Kent

Cenk Ural, PhD

Vice President, Barclays

Donald R. Van Deventer, PhD

Chairman and Chief Executive Officer, Kamakura Corporation

Raman Vardharaj

Vice President, Oppenheimer Funds

Robert E. Whaley, PhD

Valere Blair Potter Professor of Management and Co-Director of the Financial Markets Research Center, Owen Graduate School of Management, Vanderbilt University

Mark B. Wickard

Senior Vice President/Corporate Cash
Investment Advisor, Morgan Stanley Smith
Bamey

James X. Xiong, PhD, CFA

Senior Research Consultant, Ibbotson
Associates, A Morningstar Company

Guofu Zhou

Frederick Bierman and James E. Spears Profes-
sor of Finance, Olin Business School, Washing-
ton University in St. Louis

Min Zhu

Business School, Queensland University of
Technology, Australia

Preface

It is often said that investment management is an art, not a science. However, since the early 1990s the market has witnessed a progressive shift toward a more industrial view of the investment management process. There are several reasons for this change. First, with globalization the universe of investable assets has grown many times over. Asset managers might have to choose from among several thousand possible investments from around the globe. Second, institutional investors, often together with their consultants, have encouraged asset management firms to adopt an increasingly structured process with documented steps and measurable results. Pressure from regulators and the media is another factor. Finally, the sheer size of the markets makes it imperative to adopt safe and repeatable methodologies.

In its modern sense, financial modeling is the design (or engineering) of financial instruments and portfolios of financial instruments that result in predetermined cash flows contingent upon different events. Broadly speaking, financial models are employed to manage investment portfolios and risk. The objective is the transfer of risk from one entity to another via appropriate financial arrangements. Though the aggregate risk is a quantity that cannot be altered, risk can be transferred if there is a willing counterparty.

Financial modeling came to the forefront of finance in the 1980s, with the broad diffusion

of derivative instruments. However, the concept and practice of financial modeling are quite old. The notion of the diversification of risk (central to modern risk management) and the quantification of insurance risk (a requisite for pricing insurance policies) were already understood, at least in practical terms, in the 14th century. The rich epistolary of Francesco Datini, a 14th-century merchant, banker, and insurer from Prato (Tuscany, Italy), contains detailed instructions to his agents on how to diversify risk and insure cargo.

What is specific to modern financial modeling is the quantitative management of risk. Both the pricing of contracts and the optimization of investments require some basic capabilities of statistical modeling of financial contingencies. It is the size, diversity, and efficiency of modern competitive markets that makes the use of financial modeling imperative.

This three-volume encyclopedia offers not only coverage of the fundamentals and advances in financial modeling but provides the mathematical and statistical techniques needed to develop and test financial models, as well as the practical issues associated with implementation. The encyclopedia offers the following unique features:

- The entries for the encyclopedia were written by experts from around the world. This diverse collection of expertise has created the most definitive coverage of established and

cutting-edge financial models, applications, and tools in this ever-evolving field.

- The series emphasizes both technical and managerial issues. This approach provides researchers, educators, students, and practitioners with a balanced understanding of the topics and the necessary background to deal with issues related to financial modeling.
- Each entry follows a format that includes the author, entry abstract, introduction, body, listing of key points, notes, and references. This enables readers to pick and choose among various sections of an entry, and creates consistency throughout the entire encyclopedia.
- The numerous illustrations and tables throughout the work highlight complex topics and assist further understanding.
- Each volume includes a complete table of contents and index for easy access to various parts of the encyclopedia.

TOPIC CATEGORIES

As is the practice in the creation of an encyclopedia, the topic categories are presented alphabetically. The topic categories and a brief description of each topic follow.

VOLUME I

Asset Allocation

A major activity in the investment management process is establishing policy guidelines to satisfy the investment objectives. Setting policy begins with the asset allocation decision. That is, a decision must be made as to how the funds to be invested should be distributed among the major asset classes (e.g., equities, fixed income, and alternative asset classes). The term “asset allocation” includes (1) policy asset allocation, (2) dynamic asset allocation, and (3) tactical asset allocation. Policy asset allocation decisions can loosely be characterized as long-term asset allocation decisions, in which the investor seeks to assess an appropriate long-term “normal” asset mix that represents an ideal blend of controlled risk and enhanced return. In dynamic asset allocation the asset mix (i.e., the

allocation among the asset classes) is mechanically shifted in response to changing market conditions. Once the policy asset allocation has been established, the investor can turn his or her attention to the possibility of active departures from the normal asset mix established by policy. If a decision to deviate from this mix is based upon rigorous objective measures of value, it is often called tactical asset allocation. The fundamental model used in establishing the policy asset allocation is the mean-variance portfolio model formulated by Harry Markowitz in 1952, popularly referred to as the theory of portfolio selection and modern portfolio theory.

Asset Pricing Models

Asset pricing models seek to formalize the relationship that should exist between asset returns and risk if investors behave in a hypothesized manner. At its most basic level, asset pricing is mainly about transforming asset payoffs into prices. The two most well-known asset pricing models are the arbitrage pricing theory and the capital asset pricing model. The fundamental theorem of asset pricing asserts the equivalence of three key issues in finance: (1) absence of arbitrage; (2) existence of a positive linear pricing rule; and (3) existence of an investor who prefers more to less and who has maximized his or her utility. There are two types of arbitrage opportunities. The first is paying nothing today and obtaining something in the future, and the second is obtaining something today and with no future obligations. Although the principle of absence of arbitrage is fundamental for understanding asset valuation in a competitive market, there are well-known limits to arbitrage resulting from restrictions imposed on rational traders, and, as a result, pricing inefficiencies may exist for a period of time.

Bayesian Analysis and Financial Modeling Applications

Financial models describe in mathematical terms the relationships between financial random variables through time and/or across assets. The fundamental assumption is that the

model relationship is valid independent of the time period or the asset class under consideration. Financial data contain both meaningful information and random noise. An adequate financial model not only extracts optimally the relevant information from the historical data but also performs well when tested with new data. The uncertainty brought about by the presence of data noise makes imperative the use of statistical analysis as part of the process of financial model building, model evaluation, and model testing. Statistical analysis is employed from the vantage point of either of the two main statistical philosophical traditions—frequentist and Bayesian. An important difference between the two lies with the interpretation of the concept of probability. As the name suggests, advocates of the frequentist approach interpret the probability of an event as the limit of its long-run relative frequency (i.e., the frequency with which it occurs as the amount of data increases without bound). Since the time financial models became a mainstream tool to aid in understanding financial markets and formulating investment strategies, the framework applied in finance has been the frequentist approach. However, strict adherence to this interpretation is not always possible in practice. When studying rare events, for instance, large samples of data may not be available, and in such cases proponents of frequentist statistics resort to theoretical results. The Bayesian view of the world is based on the subjectivist interpretation of probability: Probability is subjective, a degree of belief that is updated as information or data are acquired. Only in the last two decades has Bayesian statistics started to gain greater acceptance in financial modeling, despite its introduction about 250 years ago. It has been the advancements of computing power and the development of new computational methods that have fostered the growing use of Bayesian statistics in financial modeling.

Bond Valuation

The value of any financial asset is the present value of its expected future cash flows. To value

a bond (also referred to as a fixed-income security), one must be able to estimate the bond's remaining cash flows and identify the appropriate discount rate(s) at which to discount the cash flows. The traditional approach to bond valuation is to discount every cash flow with the same discount rate. Simply put, the relevant term structure of interest rate used in valuation is assumed to be flat. This approach, however, permits opportunities for arbitrage. Alternatively, the arbitrage-free valuation approach starts with the premise that a bond should be viewed as a portfolio or package of zero-coupon bonds. Moreover, each of the bond's cash flows is valued using a unique discount rate that depends on the term structure of interest rates and when in time the cash flow is. The relevant set of discount rates (that is, spot rates) is derived from an appropriate term structure of interest rates and when used to value risky bonds augmented with a suitable risk spread or premium. Rather than modeling to calculate the fair value of its price, the market price can be taken as given so as to compute a yield measure or a spread measure. Popular yield measures are the yield to maturity, yield to call, yield to put, and cash flow yield. Nominal spread, static (or zero-volatility) spread, and option-adjusted spread are popular relative value measures quoted in the bond market. Complications in bond valuation arise when a bond has one or more embedded options such as call, put, or conversion features. For bonds with embedded options, the financial modeling draws from options theory, more specifically, the use of the lattice model to value a bond with embedded options.

Credit Risk Modeling

Credit risk is a broad term used to refer to three types of risk: default risk, credit spread risk, and downgrade risk. Default risk is the risk that the counterparty to a transaction will fail to satisfy the terms of the obligation with respect to the timely payment of interest and repayment of the amount borrowed. The counterparty could be the issuer of a debt obligation or an entity on

the other side of a private transaction such as a derivative trade or a collateralized loan agreement (i.e., a repurchase agreement or a securities lending agreement). The default risk of a counterparty is often initially gauged by the credit rating assigned by one of the three rating companies—Standard & Poor’s, Moody’s Investors Service, and Fitch Ratings. Although default risk is the one that most market participants think of when reference is made to credit risk, even in the absence of default, investors are concerned about the decline in the market value of their portfolio bond holdings due to a change in credit spread or the price performance of their holdings relative to a bond index. This risk is due to an adverse change in credit spreads, referred to as credit spread risk, or when it is attributed solely to the downgrade of the credit rating of an entity, it is called downgrade risk. Financial modeling of credit risk is used (1) to measure, monitor, and control a portfolio’s credit risk, and (2) to price credit risky debt instruments. There are two general categories of credit risk models: structural models and reduced-form models. There is considerable debate as to which type of model is the best to employ.

Derivatives Valuation

A derivative instrument is a contract whose value depends on some underlying asset. The term “derivative” is used to describe this product because its value is derived from the value of the underlying asset. The underlying asset, simply referred to as the “underlying,” can be either a commodity, a financial instrument, or some reference entity such as an interest rate or stock index, leading to the classification of commodity derivatives and financial derivatives. Although there are close conceptual relations between derivative instruments and cash market instruments such as debt and equity, the two classes of instruments are used differently: Debt and equity are used primarily for raising funds from investors, while derivatives are primarily

used for dividing up and trading risks. Moreover, debt and equity are direct claims against a firm’s assets, while derivative instruments are usually claims on a third party. A derivative’s value depends on the value of the underlying, but the derivative instrument itself represents a claim on the “counterparty” to the trade. Derivatives instruments are classified in terms of their payoff characteristics: linear and nonlinear payoffs. The former, also referred to as symmetric payoff derivatives, includes forward, futures, and swap contracts while the latter include options. Basically, a linear payoff derivative is a risk-sharing arrangement between the counterparties since both are sharing the risk regarding the price of the underlying. In contrast, nonlinear payoff derivative instruments (also referred to as asymmetric payoff derivatives) are insurance arrangements because one party to the trade is willing to insure the counterparty of a minimum or maximum (depending on the contract) price. The amount received by the insuring party is referred to as the contract price or premium. Derivative instruments are used for controlling risk exposure with respect to the underlying. Hedging is a special case of risk control where a party seeks to eliminate the risk exposure. Derivative valuation or pricing is developed based on no-arbitrage price relations, relying on the assumption that two perfect substitutes must have the same price.

VOLUME II

Difference Equations and Differential Equations

The tools of linear difference equations and differential equations have found many applications in finance. A difference equation is an equation that involves differences between successive values of a function of a discrete variable. A function of such a variable is one that provides a rule for assigning values in sequences to it. The theory of linear difference equations covers three areas: solving difference equations, describing the behavior

of difference equations, and identifying the equilibrium (or critical value) and stability of difference equations. Linear difference equations are important in the context of dynamic econometric models. Stochastic models in finance are expressed as linear difference equations with random disturbances added. Understanding the behavior of solutions of linear difference equations helps develop intuition for the behavior of these models. In nontechnical terms, differential equations are equations that express a relationship between a function and one or more derivatives (or differentials) of that function. The relationship between difference equations and differential equations is that the latter are invaluable for modeling situations in finance where there is a continually changing value. The problem is that not all changes in value occur continuously. If the change in value occurs incrementally rather than continuously, then differential equations have their limitations. Instead, a financial modeler can use difference equations, which are recursively defined sequences. It would be difficult to overemphasize the importance of differential equations in financial modeling where they are used to express laws that govern the evolution of price probability distributions, the solution of economic variational problems (such as intertemporal optimization), and conditions for continuous hedging (such as in the Black-Scholes option pricing model). The two broad types of differential equations are ordinary differential equations and partial differential equations. The former are equations or systems of equations involving only one independent variable. Another way of saying this is that ordinary differential equations involve only total derivatives. Partial differential equations are differential equations or systems of equations involving partial derivatives. When one or more of the variables is a stochastic process, we have the case of stochastic differential equations and the solution is also a stochastic process. An assumption must be made about what is driving noise in a stochastic differential

equation. In most applications, it is assumed that the noise term follows a Gaussian random variable, although other types of random variables can be assumed.

Equity Models and Valuation

Traditional fundamental equity analysis involves the analysis of a company's operations for the purpose of assessing its economic prospects. The analysis begins with the financial statements of the company in order to investigate the earnings, cash flow, profitability, and debt burden. The fundamental analyst will look at the major product lines, the economic outlook for the products (including existing and potential competitors), and the industries in which the company operates. The result of this analysis will be the growth prospects of earnings. Based on the growth prospects of earnings, a fundamental analyst attempts to determine the fair value of the stock using one or more equity valuation models. The two most commonly used approaches for valuing a firm's equity are based on discounted cash flow and relative valuation models. The principal idea underlying discounted cash flow models is that what an investor pays for a share of stock should reflect what is expected to be received from it—return on the investor's investment. What an investor receives are cash dividends in the future. Therefore, the value of a share of stock should be equal to the present value of all the future cash flows an investor expects to receive from that share. To value stock, therefore, an investor must project future cash flows, which, in turn, means projecting future dividends. Popular discounted cash flow models include the basic dividend discount model, which assumes a constant dividend growth, and the multiple-phase models, which include the two-stage dividend growth model and the stochastic dividend discount models. Relative valuation methods use multiples or ratios—such as price/earnings, price/book, or price/free cash flow—to determine whether a stock is trading at higher or lower multiples than its peers.

There are two critical assumptions in using relative valuation: (1) the universe of firms selected to be included in the peer group are in fact comparable, and (2) the average multiple across the universe of firms can be treated as a reasonable approximation of “fair value” for those firms. This second assumption may be problematic during periods of market panic or euphoria. Managers of quantitative equity firms employ techniques that allow them to identify attractive stock candidates, focusing not on a single stock as is done with traditional fundamental analysis but rather on stock characteristics in order to explain why one stock outperforms another stock. They do so by statistically identifying a group of characteristics to create a quantitative selection model. In contrast to the traditional fundamental stock selection, quantitative equity managers create a repeatable process that utilizes the stock selection model to identify attractive stocks. Equity portfolio managers have used various statistical models for forecasting returns and risk. These models, referred to as predictive return models, make conditional forecasts of expected returns using the current information set. Predictive return models include regressive models, linear autoregressive models, dynamic factor models, and hidden-variable models.

Factor Models and Portfolio Construction

Quantitative asset managers typically employ multifactor risk models for the purpose of constructing and rebalancing portfolios and analyzing portfolio performance. A multifactor risk model, or simply factor model, attempts to estimate and characterize the risk of a portfolio, either relative to a benchmark such as a market index or in absolute value. The model allows the decomposition of risk factors into a systematic and an idiosyncratic component. The portfolio’s risk exposure to broad risk factors is captured by the systematic risk. For equity portfolios these are typically fundamental factors (e.g., market capitalization and value

vs. growth), technical (e.g., momentum), and industry/sector/country. For fixed-income portfolios, systematic risk captures a portfolio’s exposure to broad risk factors such as the term structure of interest rates, credit spreads, optionality (call and prepayment), credit, and sectors. The portfolio’s systematic risk depends not only on its exposure to these risk factors but also the volatility of the risk factors and how they correlate with each other. In contrast to systematic risk, idiosyncratic risk captures the uncertainty associated with news affecting the holdings of individual issuers in the portfolio. In equity portfolios, idiosyncratic risk can be easily diversified by reducing the importance of individual issuers in the portfolio. Because of the larger number of issuers in bond indexes, however, this is a difficult task. There are different types of factor models depending on the factors. Factors can be exogenous variables or abstract variables formed by portfolios. Exogenous factors (or known factors) can be identified from traditional fundamental analysis or from economic theory that suggests macroeconomic factors. Abstract factors, also called unidentified or latent factors, can be determined with the statistical tool of factor analysis or principal component analysis. The simplest type of factor models is where the factors are assumed to be known or observable, so that time-series data are those factors that can be used to estimate the model. The four most commonly used approaches for the evaluation of return premiums and risk characteristics to factors are portfolio sorts, factor models, factor portfolios, and information coefficients. Despite its use by quantitative asset managers, the basic building blocks of factor models used by model builders and by traditional fundamental analysts are the same: They both seek to identify the drivers of returns for the asset class being analyzed.

Financial Econometrics

Econometrics is the branch of economics that draws heavily on statistics for testing and

analyzing economic relationships. The economic equivalent of the laws of physics, econometrics represents the quantitative, mathematical laws of economics. Financial econometrics is the econometrics of financial markets. It is a quest for models that describe financial time series such as prices, returns, interest rates, financial ratios, defaults, and so on. Although there are similarities between financial econometric models and models of the physical sciences, there are two important differences. First, the physical sciences aim at finding immutable laws of nature; econometric models model the economy or financial markets—artifacts subject to change. Because the economy and financial markets are artifacts subject to change, econometric models are not unique representations valid throughout time; they must adapt to the changing environment. Second, while basic physical laws are expressed as differential equations, financial econometrics uses both continuous-time and discrete-time models.

Financial Modeling Principles

The origins of financial modeling can be traced back to the development of mathematical equilibrium at the end of the nineteenth century, followed in the beginning of the twentieth century with the introduction of sophisticated mathematical tools for dealing with the uncertainty of prices and returns. In the 1950s and 1960s, financial modelers had tools for dealing with probabilistic models for describing markets, the principles of contingent claims analysis, an optimization framework for portfolio selection based on mean and variance of asset returns, and an equilibrium model for pricing capital assets. The 1970s ushered in models for pricing contingent claims and a new model for pricing capital assets based on arbitrage pricing. Consequently, by the end of the 1970s, the frameworks for financial modeling were well known. It was the advancement of computing power and refinements of the theories to take into account real-world market imperfections and

conventions starting in the 1980s that facilitated implementation and broader acceptance of mathematical modeling of financial decisions. The diffusion of low-cost high-performance computers has allowed the broad use of numerical methods, the landscape of financial modeling. The importance of finding closed-form solutions and the consequent search for simple models has been dramatically reduced. Computationally intensive methods such as Monte Carlo simulations and the numerical solution of differential equations are now widely used. As a consequence, it has become feasible to represent prices and returns with relatively complex models. Nonnormal probability distributions have become commonplace in many sectors of financial modeling. It is fair to say that the key limitation of financial modeling is now the size of available data samples or training sets, not the computations; it is the data that limit the complexity of estimates. Mathematical modeling has also undergone major changes. Techniques such as equivalent martingale methods are being used in derivative pricing, and cointegration, the theory of fat-tailed processes, and state-space modeling (including ARCH/GARCH and stochastic volatility models) are being used in financial modeling.

Financial Statement Analysis

Much of the financial data that are used in constructing financial models for forecasting and valuation purposes draw from the financial statements that companies are required to provide to investors. The four basic financial statements are the balance sheet, the income statement, the statement of cash flows, and the statement of shareholders' equity. It is important to understand these data so that the information conveyed by them is interpreted properly in financial modeling. The financial statements are created using several assumptions that affect how to use and interpret the financial data. The analysis of financial statements involves the selection, evaluation, and

interpretation of financial data and other pertinent information to assist in evaluating the operating performance and financial condition of a company. The operating performance of a company is a measure of how well a company has used its resources—its assets, both tangible and intangible—to produce a return on its investment. The financial condition of a company is a measure of its ability to satisfy its obligations, such as the payment of interest on its debt in a timely manner. There are many tools available in the analysis of financial information. These tools include financial ratio analysis and cash flow analysis. Cash flows are essential ingredients in valuation. Therefore, understanding past and current cash flows may help in forecasting future cash flows and, hence, determine the value of the company. Moreover, understanding cash flow allows the assessment of the ability of a firm to maintain current dividends and its current capital expenditure policy without relying on external financing. Financial modelers must understand how to use these financial ratios and cash flow information in the most effective manner in building models.

Finite Mathematics and Basic Functions for Financial Modeling

The collection of mathematical tools that does not include calculus is often referred to as “finite mathematics.” This includes matrix algebra, probability theory, and statistical analysis. Ordinary algebra deals with operations such as addition and multiplication performed on individual numbers. In financial modeling, it is useful to consider operations performed on ordered arrays of numbers. Ordered arrays of numbers are called vectors and matrices while individual numbers are called scalars. Probability theory is the mathematical approach to formalize the uncertainty of events. Even though a decision maker may not know which one of the set of possible events may finally occur, with probability theory a decision maker has the means of providing each event with

a certain probability. Furthermore, it provides the decision maker with the axioms to compute the probability of a composed event in a unique way. The rather formal environment of probability theory translates in a reasonable manner to the problems related to risk and uncertainty in finance such as, for example, the future price of a financial asset. Today, investors may be aware of the price of a certain asset, but they cannot say for sure what value it might have tomorrow. To make a prudent decision, investors need to assess the possible scenarios for tomorrow’s price and assign to each scenario a probability of occurrence. Only then can investors reasonably determine whether the financial asset satisfies an investment objective included within a portfolio. Probability models are theoretical models of the occurrence of uncertain events. In contrast, statistics is about empirical data and can be broadly defined as a set of methods used to make inferences from a known sample to a larger population that is in general unknown. In finance, a particular important example is making inferences from the past (the known sample) to the future (the unknown population). There are important mathematical functions with which the financial modeler should be acquainted. These include the continuous function, the indicator function, the derivative of a function, the monotonic function, and the integral, as well as special functions such as the characteristic function of random variables and the factorial, the gamma, beta, and Bessel functions.

Liquidity and Trading Costs

In broad terms, liquidity refers to the ability to execute a trade or liquidate a position with little or no cost or inconvenience. Liquidity depends on the market where a financial instrument is traded, the type of position traded, and sometimes the size and trading strategy of an individual trade. Liquidity risks are those associated with the prospect of imperfect market liquidity and can relate to risk of loss or

risk to cash flows. There are two main aspects to liquidity risk measurement: the measurement of liquidity-adjusted measures of market risk and the measurement of liquidity risks per se. Market practitioners often assume that markets are liquid—that is, that they can liquidate or unwind positions at going market prices—usually taken to be the mean of bid and ask prices—without too much difficulty or cost. This assumption is very convenient and provides a justification for the practice of marking positions to market prices. However, it is often empirically questionable, and the failure to allow for liquidity can undermine the measurement of market risk. Because liquidity risk is a major risk factor in its own right, portfolio managers and traders will need to measure this risk in order to formulate effective portfolio and trading strategies. A considerable amount of work has been done in the equity market in estimating liquidity risk. Because transaction costs are incurred when buying or selling stocks, poorly executed trades can adversely impact portfolio returns and therefore relative performance. Transaction costs are classified as explicit costs such as brokerage and taxes, and implicit costs, which include market impact cost, price movement risk, and opportunity cost. Broadly speaking, market impact cost is the price that a trader has to pay for obtaining liquidity in the market and is a key component of trading costs that must be modeled so that effective trading programs for executing trades can be developed. Typical forecasting models for market impact costs are based on a statistical factor approach where the independent variables are trade-based factors or asset-based factors.

VOLUME III

Model Risk and Selection

Model risk is the risk of error in pricing or risk-forecasting models. In practice, model risk arises because (1) any model involves simpli-

fication and calibration, and both of these require subjective judgments that are prone to error, and/or (2) a model is used inappropriately. Although model risk cannot be avoided, there are many ways in which financial modelers can manage this risk. These include (1) recognizing model risk, (2) identifying, evaluating, and checking the model's key assumption, (3) selecting the simplest reasonable model, (4) resisting the temptation to ignore small discrepancies in results, (5) testing the model against known problems, (6) plotting results and employing nonparametric statistics, (7) back-testing and stress-testing the model, (8) estimating model risk quantitatively, and (9) reevaluating models periodically. In financial modeling, model selection requires a blend of theory, creativity, and machine learning. The machine-learning approach starts with a set of empirical data that the financial modeler wants to explain. Data are explained by a family of models that include an unbounded number of parameters and are able to fit data with arbitrary precision. There is a trade-off between model complexity and the size of the data sample. To implement this trade-off, ensuring that models have forecasting power, the fitting of sample data is constrained to avoid fitting noise. Constraints are embodied in criteria such as the Akaike information criterion or the Bayesian information criterion. Economic and financial data are generally scarce given the complexity of their patterns. This scarcity introduces uncertainty as regards statistical estimates obtained by the financial modeler. It means that the data might be compatible with many different models with the same level of statistical confidence. Methods of probabilistic decision theory can be used to deal with model risk due to uncertainty regarding the model's parameters. Probabilistic decision making starts from the Bayesian inference process and involves computer simulations in all realistic situations. Since a risk model is typically a combination of a probability distribution model and a risk measure, a critical assumption is the probability distribution assumed for

the random variable of interest. Too often, the Gaussian distribution is the model of choice. Empirical evidence supports the use of probability distributions that exhibit fat tails such as the Student's t distribution and its asymmetric version and the Pareto stable class of distributions and their tempered extensions. Extreme value theory offers another approach for risk modeling.

Mortgage-Backed Securities Analysis and Valuation

Mortgage-backed securities are fixed-income securities backed by a pool of mortgage loans. Residential mortgage-backed securities (RMBS) are backed by a pool of residential mortgage loans (one-to-four family dwellings). The RMBS market includes agency RMBS and nonagency RMBS. The former are securities issued by the Government National Mortgage Association (Ginnie Mae), Fannie Mae, and Freddie Mac. Agency RMBS include passthrough securities, collateralized mortgage obligations, and stripped mortgage-backed securities (interest-only and principal-only securities). The valuation of RMBS is complicated due to prepayment risk, a form of call risk. In contrast, nonagency RMBS are issued by private entities, have no implicit or explicit government guarantee, and therefore require one or more forms of credit enhancement in order to be assigned a credit rating. The analysis of nonagency RMBS must take into account both prepayment risk and credit risk. The most commonly used method for valuing RMBS is the Monte Carlo method, although other methods have garnered favor, in particular the decomposition method. The analysis of RMBS requires an understanding of the factors that impact prepayments.

Operational Risk

Operational risk has been regarded as a mere part of a financial institution's "other" risks. However, failures of major financial entities

have made regulators and investors aware of the importance of this risk. In general terms, operational risk is the risk of loss resulting from inadequate or failed internal processes, people, or systems or from external events. This risk encompasses legal risks, which includes, but is not limited to, exposure to fines, penalties, or punitive damages resulting from supervisory actions, as well as private settlements. Operational risk can be classified according to several principles: nature of the loss (internally inflicted or externally inflicted), direct losses or indirect losses, degree of expectancy (expected or unexpected), risk type, event type or loss type, and by the magnitude (or severity) of loss and the frequency of loss. Operational risk can be the cause of reputational risk, a risk that can occur when the market reaction to an operational loss event results in reduction in the market value of a financial institution that is greater than the amount of the initial loss. The two principal approaches in modeling operational loss distributions are the nonparametric approach and the parametric approach. It is important to employ a model that captures tail events, and for this reason in operational risk modeling, distributions that are characterized as light-tailed distributions should be used with caution. The models that have been proposed for assessing operational risk can be broadly classified into top-down models and bottom-up models. Top-down models quantify operational risk without attempting to identify the events or causes of losses. Bottom-up models quantify operational risk on a micro level, being based on identified internal events. The obstacle hindering the implementation of these models is the scarcity of available historical operational loss data.

Optimization Tools

Optimization is an area in applied mathematics that, most generally, deals with efficient algorithms for finding an optimal solution among a set of solutions that satisfy given constraints. Mathematical programming, a management

science tool that uses mathematical optimization models to assist in decision making, includes linear programming, integer programming, mixed-integer programming, nonlinear programming, stochastic programming, and goal programming. Unlike other mathematical tools that are available to decision makers such as statistical models (which tell the decision maker what occurred in the past), forecasting models (which tell the decision maker what might happen in the future), and simulation models (which tell the decision maker what will happen under different conditions), mathematical programming models allow the decision maker to identify the “best” solution. Markowitz’s mean-variance model for portfolio selection is an example of an application of one type of mathematical programming (quadratic programming). Traditional optimization modeling assumes that the inputs to the algorithms are certain, but there are also branches of optimization such as robust optimization that study the optimal decision under uncertainty about the parameters of the problem. Stochastic programming deals with both the uncertainty about the parameters and a multiperiod decision-making framework.

Probability Distributions

In financial models where the outcome of interest is a random variable, an assumption must be made about the random variable’s probability distribution. There are two types of probability distributions: discrete and continuous. Discrete probability distributions are needed whenever the random variable is to describe a quantity that can assume values from a countable set, either finite or infinite. A discrete probability distribution (or law) is quite intuitive in that it assigns certain values, positive probabilities, adding up to one, while any other value automatically has zero probability. Continuous probability distributions are needed when the random variable of interest can assume any value inside of one or more

intervals of real numbers such as, for example, any number greater than zero. Asset returns, for example, whether measured monthly, weekly, daily, or at an even higher frequency are commonly modeled as continuous random variables. In contrast to discrete probability distributions that assign positive probability to certain discrete values, continuous probability distributions assign zero probability to any single real number. Instead, only entire intervals of real numbers can have positive probability such as, for example, the event that some asset return is not negative. For each continuous probability distribution, this necessitates the so-called probability density, a function that determines how the entire probability mass of one is distributed. The density often serves as the proxy for the respective probability distribution. To model the behavior of certain financial assets in a stochastic environment, a financial modeler can usually resort to a variety of theoretical distributions. Most commonly, probability distributions are selected that are analytically well known. For example, the normal distribution (a continuous distribution)—also called the Gaussian distribution—is often the distribution of choice when asset returns are modeled. Or the exponential distribution is applied to characterize the randomness of the time between two successive defaults of firms in a bond portfolio. Many other distributions are related to them or built on them in a well-known manner. These distributions often display pleasant features such as stability under summation—meaning that the return of a portfolio of assets whose returns follow a certain distribution again follows the same distribution. However, one has to be careful using these distributions since their advantage of mathematical tractability is often outweighed by the fact that the stochastic behavior of the true asset returns is not well captured by these distributions. For example, although the normal distribution generally renders modeling easy because all moments of the distribution exist, it fails to reflect stylized facts commonly encountered in

asset returns—namely, the possibility of very extreme movements and skewness. To remedy this shortcoming, probability distributions accounting for such extreme price changes have become increasingly popular. Some of these distributions concentrate exclusively on the extreme values while others permit any real number, but in a way capable of reflecting market behavior. Consequently, a financial modeler has available a great selection of probability distributions to realistically reproduce asset price changes. Their common shortcoming is generally that they are mathematically difficult to handle.

Risk Measures

The standard assumption in financial models is that the distribution for the return on financial assets follows a normal (or Gaussian) distribution and therefore the standard deviation (or variance) is an appropriate measure of risk in the portfolio selection process. This is the risk measure that is used in the well-known Markowitz portfolio selection model (that is, mean-variance model), which is the foundation for modern portfolio theory. Mounting evidence since the early 1960s strongly suggests that return distributions do not follow a normal distribution, but instead exhibit heavy tails and, possibly, skewness. The “tails” of the distribution are where the extreme values occur, and these extreme values are more likely than would be predicted by the normal distribution. This means that between periods where the market exhibits relatively modest changes in prices and returns, there will be periods where there are changes that are much higher (that is, crashes and booms) than predicted by the normal distribution. This is of major concern to financial modelers in seeking to generate probability estimates for financial risk assessment. To more effectively implement portfolio selection, researchers have proposed alternative risk measures. These risk measures fall into

two disjointed categories: dispersion measures and safety-first measures. Dispersion measures include mean standard deviation, mean absolute deviation, mean absolute moment, index of dissimilarity, mean entropy, and mean colog. Safety-first risk measures include classical safety first, value-at-risk, average value-at-risk, expected tail loss, MiniMax, lower partial moment, downside risk, probability-weighted function of deviations below a specified target return, and power conditional value-at-risk. Despite these alternative risk measures, the most popular risk measure used in financial modeling is volatility as measured by the standard deviation. There are different types of volatility: historical, implied volatility, level-dependent volatility, local volatility, and stochastic volatility (e.g., jump-diffusion volatility). There are risk measures commonly used for bond portfolio management. These measures include duration, convexity, key rate duration, and spread duration.

Software for Financial Modeling

The development of financial models requires the modeler to be familiar with spreadsheets such as Microsoft Excel and/or a platform to implement concepts and algorithms such as the Palisade Decision Tools Suite and other Excel-based software (mostly @RISK1, Solver2, VBA3), and MATLAB. Financial modelers can choose one or the other, depending on their level of familiarity and comfort with spreadsheet programs and their add-ins versus programming environments such as MATLAB. Some tasks and implementations are easier in one environment than in the other. MATLAB is a modeling environment that allows for input and output processing, statistical analysis, simulation, and other types of model building for the purpose of analysis of a situation. MATLAB uses a number-array-oriented programming language, that is, a programming language in which vectors and matrices

are the basic data structures. Reliable built-in functions, a wide range of specialized toolboxes, easy interface with widespread software like Microsoft Excel, and beautiful graphing capabilities for data visualization make implementation with MATLAB efficient and useful for the financial modeler. Visual Basic for Applications (VBA) is a programming language environment that allows Microsoft Excel users to automate tasks, create their own functions, perform complex calculations, and interact with spreadsheets. VBA shares many of the same concepts as object-oriented programming languages. Despite some important limitations, VBA does add useful capabilities to spreadsheet modeling, and it is a good tool to know because Excel is the platform of choice for many finance professionals.

Stochastic Processes and Tools

Stochastic integration provides a coherent way to represent that instantaneous uncertainty (or volatility) cumulates over time. It is thus fundamental to the representation of financial processes such as interest rates, security prices, or cash flows. Stochastic integration operates on stochastic processes and produces random variables or other stochastic processes. Stochastic integration is a process defined on each path as the limit of a sum. However, these sums are different from the sums of the Riemann-Lebesgue integrals because the paths of stochastic processes are generally not of bounded variation. Stochastic integrals in the sense of Itô are defined through a process of approximation by (1) defining Brownian motion, which is the continuous limit of a random walk, (2) defining stochastic integrals for elementary functions as the sums of the products of the elementary functions multiplied by the increments of the Brownian motion, and (3) extending this definition to any function through approximating sequences. The major application of integration to financial modeling involves stochastic

integrals. An understanding of stochastic integrals is needed to understand an important tool in contingent claims valuation: stochastic differential equations. The dynamic of financial asset returns and prices can be expressed using a deterministic process if there is no uncertainty about its future behavior, or, with a stochastic process, in the more likely case when the value is uncertain. Stochastic processes in continuous time are the most used tool to explain the dynamic of financial assets returns and prices. They are the building blocks to construct financial models for portfolio optimization, derivatives pricing, and risk management. Continuous-time processes allow for more elegant theoretical modeling compared to discrete time models, and many results proven in probability theory can be applied to obtain a simple evaluation method.

Statistics

Probability models are theoretical models of the occurrence of uncertain events. In contrast, statistics is about empirical data and can be broadly defined as a set of methods used to make inferences from a known sample to a larger population that is in general unknown. In finance, a particular important example is making inferences from the past (the known sample) to the future (the unknown population). In statistics, probabilistic models are applied using data so as to estimate the parameters of these models. It is not assumed that all parameter values in the model are known. Instead, the data for the variables in the model to estimate the value of the parameters are used and then applied to test hypotheses or make inferences about their estimated values. In financial modeling, the statistical technique of regression models is the workhorse. However, because regression models are part of the field of financial econometrics, this topic is covered in that topic category. Understanding dependences or functional links between variables is a key theme in

financial modeling. In general terms, functional dependencies are represented by dynamic models. Many important models are linear models whose coefficients are correlation coefficients. In many instances in financial modeling, it is important to arrive at a quantitative measure of the strength of dependencies. The correlation coefficient provides such a measure. In many instances, however, the correlation coefficient might be misleading. In particular, there are cases of nonlinear dependencies that result in a zero correlation coefficient. From the point of view of financial modeling, this situation is particularly dangerous as it leads to substantially underestimated risk. Different measures of dependence have been proposed, in particular copula functions. The copula overcomes the drawbacks of the correlation as a measure of dependency by allowing for a more general measure than linear dependence, allowing for the modeling of dependence for extreme events, and being indifferent to continuously increasing transformations. Another essential tool in financial modeling, because it allows the incorporation of uncertainty in financial models and consideration of additional layers of complexity that are difficult to incorporate in analytical models, is Monte Carlo simulation. The main idea of Monte Carlo simulation is to represent the uncertainty in market variables through scenarios, and to evaluate parameters of interest that depend on these market variables in complex ways. The advantage of such an approach is that it can easily capture the dynamics of underlying processes and the otherwise complex effects of interactions among market variables. A substantial amount of research in recent years has been dedicated to making scenario generation more accurate and efficient, and a number of sophisticated computational techniques are now available to the financial modeler.

Term Structure Modeling

The arbitrage-free valuation approach to the valuation of option-free bonds, bonds with em-

bedded options, and option-type derivative instruments requires that a financial instrument be viewed as a package of zero-coupon bonds. Consequently, in financial modeling, it is essential to be able to discount each expected cash flow by the appropriate interest rate. That rate is referred to as the spot rate. The term structure of interest rates provides the relationship between spot rates and maturity. Because of its role in valuation of cash bonds and option-type derivatives, the estimation of the term structure of interest rates is of critical importance as an input into a financial model. In addition to its role in valuation modeling, term structure models are fundamental to expressing value, risk, and establishing relative value across the spectrum of instruments found in the various interest-rate or bond markets. The term structure is most often specified for a specific market such as the U.S. Treasury market, the bond market for double-A rated financial institutions, the interest rate market for LIBOR, and swaps. Static models of the term structure are characterizations that are devoted to relationships based on a given market and do not serve future scenarios where there is uncertainty. Standard static models include those known as the spot yield curve, discount function, par yield curve, and the implied forward curve. Instantiations of these models may be found in both a discrete- and continuous-time framework. An important consideration is establishing how these term structure models are constructed and how to transform one model into another. In modeling the behavior of interest rates, stochastic differential equations (SDEs) are commonly used. The SDEs used to model interest rates must capture the market properties of interest rates such as mean reversion and/or a volatility that depends on the level of interest rates. For a one-factor model, the SDE is used to model the behavior of the short-term rate, referred to as simply the "short rate." The addition of another factor (i.e., a two-factor model) involves extending the SDE to represent the behavior of the short rate and a long-term rate (i.e., long rate).

The entries can serve as material for a wide spectrum of courses, such as the following:

- Financial engineering
- Financial mathematics
- Financial econometrics
- Statistics with applications in finance
- Quantitative asset management
- Asset and derivative pricing
- Risk management

Frank J. Fabozzi
Editor, *Encyclopedia of Financial Models*

Guide to the *Encyclopedia of Financial Models*

The *Encyclopedia of Financial Models* provides comprehensive coverage of the field of financial modeling. This reference work consists of three separate volumes and 127 entries. Each entry provides coverage of the selected topic intended to inform a broad spectrum of readers ranging from finance professionals to academicians to students to fiduciaries. To derive the greatest possible benefit from the *Encyclopedia of Financial Models*, we have provided this guide. It explains how the information within the encyclopedia can be located.

ORGANIZATION

The *Encyclopedia of Financial Models* is organized to provide maximum ease of use for its readers.

Table of Contents

A complete table of contents for the entire encyclopedia appears in the front of each volume. This list of titles represents topics that have been carefully selected by the editor, Frank J. Fabozzi. The Preface includes a more detailed description of the volumes and the topic categories that the entries are grouped under.

Index

A Subject Index for the entire encyclopedia is located at the end of each volume. The sub-

jects in the index are listed alphabetically and indicate the volume and page number where information on this topic can be found.

Entries

Each entry in the *Encyclopedia of Financial Models* begins on a new page, so that the reader may quickly locate it. The author's name and affiliation are displayed at the beginning of the entry. All entries in the encyclopedia are organized according to a standard format, as follows:

- Title and author
- Abstract
- Introduction
- Body
- Key points
- Notes
- References

Abstract

The abstract for each entry gives an overview of the topic, but not necessarily the content of the entry. This is designed to put the topic in the context of the entire *Encyclopedia*, rather than give an overview of the specific entry content.

Introduction

The text of each entry begins with an introductory section that defines the topic under

discussion and summarizes the content. By reading this section, the reader gets a general idea about the content of a specific entry.

Body

The body of each entry explains the purpose, theory, and math behind each model.

Key Points

The key points section provides in bullet point format a review of the materials discussed in

each entry. It imparts to the reader the most important issues and concepts discussed.

Notes

The notes provide more detailed information and citations of further readings.

References

The references section lists the publications cited in the entry.

ENCYCLOPEDIA
OF
FINANCIAL MODELS

Volume III

Mortgage-Backed Securities Analysis and Valuation

Valuing Mortgage-Backed and Asset-Backed Securities

FRANK J. FABOZZI, PhD, CFA, CPA
Professor of Finance, EDHEC School of Business

MARK B. WICKARD
Senior Vice President/Corporate Cash Investment Advisor, Morgan Stanley Smith Bamey

Abstract: The valuing (or pricing) of a bond without an embedded option (that is, an option-free bond) is straightforward. The value is equal to the present value of the expected cash flows. Ignoring defaults, for an option-free bond the cash flows are known and consist of the periodic interest payments and principal at the maturity date. The interest or discount rates for computing the present value of the cash flows begin with the spot rates for a benchmark security and to those rates an appropriate spread is added. Moving from valuing option-free bonds to corporate bonds and agency debentures with embedded options is not simple. The interest rate-sensitive options that can be embedded into these bonds are call options, put options, accelerated sinking provisions, and, for floating-rate securities, caps on the interest rate. The reason valuation is complicated is that the embedded options must be taken into account and the theoretical option-free value of the bond must be adjusted accordingly. The technique typically used for valuing corporate bonds and agency debentures with embedded options is the lattice method. Mortgage-backed securities also have embedded options: the right of the borrowers in a loan pool to prepay their mortgage loan. However, because future cash flows for a loan pool are sensitive to not only the current interest rate but the history of rates since the loans were originated, the lattice method which is solved using backward induction cannot be employed. Instead, the most common methodology used for valuing mortgage-backed securities and mortgage-related asset-backed securities is the Monte Carlo simulation model. Other types of asset-backed securities are straightforward to value. In addition to the complications in valuing mortgage-backed securities and mortgage-related asset-backed securities, there is the difficulty in estimating their price sensitivity to changes in interest rates (that is, duration and convexity). The Monte Carlo simulation model can be used to compute the effective duration of these securities. This duration measure takes into consideration how a change in interest rates can impact a security's cash flow.

In this entry we will explain the methodology for valuing asset-backed securities (ABS) and mortgage-backed securities (MBS) and measures of relative value.¹ We begin by reviewing cash-flow yield analysis and the limitations

of the spread measure that is a result of that analysis—the nominal spread. We then look at a better spread measure called the zero-volatility spread, but point out its limitation as a measure of relative value for MBS products because

of the borrower's prepayment option and for ABS products where the prepayment option has value. Finally, we look at the methodology for valuing MBS and for ABS products where the prepayment option has value—the *Monte Carlo simulation model*. A by-product of this model is a spread measure called the option-adjusted spread (OAS). This measure is superior to the nominal spread and the zero-volatility spread for ABS products where the prepayment option has a value because it takes into account how cash flows may change when interest rates change. That is, it recognizes the borrower's prepayment option and how that affects prepayments when interest rates may change in the future. While the OAS is superior to the two other spread measures, it is based on assumptions that must be understood by an investor and the sensitivity of the security's value and OAS to changes in those assumptions must be investigated.

CASH-FLOW YIELD ANALYSIS

The yield on any financial instrument is the interest rate that makes the present value of the expected cash flow equal to its market price plus accrued interest. For ABS and MBS, the yield calculated is called a *cash-flow yield*. The problem in calculating the cash-flow yield of MBS and ABS is that because of prepayments the cash flow is unknown. A prepayment is the amount of the payment made by the obligor in the loan pool that is in excess of the scheduled principal payment. Prepayments can be voluntary such as for refinancing the loan or involuntary such as for a default by the obligor. Consequently, to determine a cash-flow yield some assumption about the prepayment rate and recovery rate in the case of defaults must be made.²

The cash flow for MBS and ABS is typically monthly. The convention is to compare the yield on MBS and ABS to that of a Treasury coupon security by calculating the security's bond-

equivalent yield. The bond-equivalent yield for a Treasury coupon security is found by doubling the semiannual yield. However, it is incorrect to do this for MBS and ABS because the investor has the opportunity to generate greater interest by reinvesting the more frequent cash flows. The market practice is to calculate a yield so as to make it comparable to the yield to maturity on a bond-equivalent yield basis. The formula for annualizing the monthly cash-flow yield for MBS and ABS is as follows:

$$\text{Bond-equivalent yield} = 2[(1 + i_M)^6 - 1]$$

where i_M is the monthly interest rate that will equate the present value of the projected monthly cash flow to the market price (plus accrued interest) of the security.

All yield measures suffer from problems that limit their use in assessing a security's potential return. The yield to maturity for a Treasury, agency, or corporate bond has two major shortcomings as a measure of a bond's potential return. To realize the stated yield to maturity, the investor must: (1) reinvest the coupon payments at a rate equal to the yield to maturity and (2) hold the bond to the maturity date. The reinvestment of the coupon payments is critical and for long-term bonds can comprise as much as 80% of the bond's return. The risk of having to reinvest the interest payments at less than the computed yield is called *reinvestment risk*. The risk associated with a decline in the value of a security due to a rise in interest rates is called *interest rate risk* and in practice is quantified by computing the security's duration and convexity.

These shortcomings are equally applicable to the cash-flow yield measure for ABS and MBS: (1) the projected cash flows are assumed to be reinvested at the computed cash-flow yield and (2) the security is assumed to be held until the final payout based on some prepayment assumption. The importance of reinvestment risk—the risk that the cash flow will be reinvested at a rate less than the calculated cash-flow yield—is particularly important for amortizing MBS and

ABS products, because payments are monthly and both interest and principal must be reinvested. Moreover, an additional assumption is that the projected cash flow is actually realized. If the prepayment experience and the recovery rate realized differ from that assumed, the cash-flow yield will not be realized.

Given the computed cash-flow yield and the average life for a security based on some prepayment assumption and default/recovery assumption, the next step is to compare the yield to the yield for a comparable Treasury security. "Comparable" is typically defined as a Treasury security with the same maturity as the (weighted) average life or the duration of the security. The difference between the cash-flow yield and the yield on a comparable Treasury security is called the *nominal spread*.

Unfortunately, it is the nominal spread that investors will too often use as a measure of relative value for ABS and MBS. However, this spread masks the fact that a portion of the nominal spread may be compensation for accepting prepayment risk. Instead of nominal spread, investors need a measure that indicates the compensation after adjusting for prepayment risk for all MBS and for ABS where the prepayment option has value. This measure is called the option-adjusted spread. Before discussing this measure, we describe another spread measure commonly quoted for MBS and ABS called the zero-volatility spread. This measure takes into account another problem with the nominal spread. Specifically, the nominal spread is computed assuming that all the cash flows for a security should be discounted at only one interest rate. That is, it fails to recognize the term structure of interest rates.

ZERO-VOLATILITY SPREAD

The proper procedure to compare ABS and MBS to a Treasury is to compare it to a portfolio of Treasury securities that have the same cash flow. The value of the security is then equal to the present value of all of the cash flows. The secu-

rity's value, assuming the cash flows are default free, will equal the present value of the replicating portfolio of Treasury securities. In turn, these cash flows are valued at the Treasury spot rates.

The *zero-volatility spread* is a measure of the spread that the investor would realize over the entire Treasury spot rate curve if the non-Treasury security being analyzed is held to maturity. It is not a spread off one point on the Treasury yield curve, as is the nominal spread. The zero-volatility spread (also called the *Z-spread* and the *static spread*) is the spread that will make the present value of the cash flows from the non-Treasury security when discounted at the Treasury spot rate plus the spread equal to the market price plus accrued interest of the non-Treasury security. A trial-and-error procedure (or search algorithm) is required to determine the zero-volatility spread.

In general, the shorter the average life of the ABS/MBS, the less the zero-volatility spread will deviate from the nominal spread. The magnitude of the difference between the nominal spread and the zero-volatility spread also depends on the shape of the yield curve. The steeper the yield curve, the greater the difference.

If borrowers in the underlying loan pool have the right to prepay but do not typically take advantage of a decline in interest rates below the loan's rate to refinance, then the zero-volatility spread is the appropriate measure of relative value and it should be used in valuing cash flows to determine the value of ABS. This is the case, for example, for automobile loan ABS. While borrowers have the right to refinance when rates decline below the loan rate, they typically do not. In contrast, for standard residential mortgage loans, home equity loan ABS, and manufactured housing ABS, the borrowers in the underlying pool do refinance when interest rates decline below the loan rate. The next methodology and spread measure are used for products with this characteristic. Basically, they are used for all residential MBS and mortgage-related ABS.

VALUATION USING MONTE CARLO SIMULATION AND OAS ANALYSIS

In fixed income valuation modeling, there are two methodologies commonly used to value securities with embedded options—the Monte Carlo simulation model and the lattice model. The Monte Carlo simulation model involves simulating a large number of potential interest rate paths in order to assess the value of a security on those different paths. This model is the most flexible of the two valuation methodologies for valuing interest rate-sensitive instruments where the history of interest rates is important. MBS and mortgage-related ABS are commonly valued using this model. As explained below, a by-product of this valuation model is the OAS. (An alternative model for valuing agency passthrough securities that does not require a prepayment model is provided in Kalotay, Yang, and Fabozzi, 2004.)

A lattice model is used to value callable agency debentures and corporate bonds. This valuation model accommodates securities in which the decision to exercise a call option is not dependent on how interest rates evolved over time. That is, the decision of an issuer to call a bond will depend on the prevailing interest rate at which the issue can be refunded relative to the issue's coupon rate and the costs associated with refunding, and not the path interest rates took to get to that rate. MBS and mortgage-related ABS which allow prepayments have periodic cash flows that are interest rate path dependent. This means that the cash flow received in one period is determined not only by the current interest rate level, but also by the path that interest rates took to get to the current level.

Prepayments for MBS and mortgage-related ABS are interest rate path dependent because this month's prepayment rate depends on whether there have been prior opportunities to refinance since the underlying loans were originated. Moreover, the cash flows to be received

in the current month by investors in a bond class of MBS and mortgage-related ABS transaction depend on the outstanding balances of the other bond classes in the transaction. For example, in the case of a planned amortization class (PAC) bond in a collateralized mortgage obligation structure, all prepayments from the time the security was issued up to the valuation date affect the amount of support bonds outstanding and therefore the cash flow at the valuation date for the PAC bond.³ Thus, we need the history of prepayments to calculate the balances of bond classes in a structure.

Conceptually, valuation using the Monte Carlo simulation model is simple. In practice, however, it is very complex. The simulation involves generating a set of cash flows based on simulated future refinancing rates, which in turn imply simulated prepayment and default/recovery rates. The objective is to figure out how the value of the collateral gets transmitted to the bond classes in the structure. More specifically, modeling is used to identify where the value in a transaction has been allocated and where the risk (prepayment risk and credit risk) has been distributed in order to identify the bond classes with low risk and high value.

Simulating Interest Rate Paths and Cash Flows

Monte Carlo simulation is a management science/operations research technique that is commonly employed in finance.⁴ The purpose of Monte Carlo simulation is to generate a probability distribution for the outcome of some random variable of interest. In its application to valuing securities, it is used to generate interest rate paths so that potential cash flows on those paths can be determined and then each path is valued. (In the parlance of simulation, an interest rate path is referred to as a trial.) The value for the security on each of those interest rate paths is then one value in determining the estimated probability distribution for the security's value.

The procedure for generating the interest rate paths begins with a benchmark term structure of interest rates and associated with this benchmark are market prices for benchmark securities. Given the benchmark term structure of interest rates, the interest rate paths are adjusted (that is, calibrated) so that the average price produced by the model for each benchmark security will equal the market price for the benchmark security.

Most models use the on-the-run Treasury issues in this calibration process. Other model developers use off-the-run Treasury issues as well. The argument for using off-the-run Treasury issues is that the price/yield of on-the-run Treasury issues will not reflect their true economic value because the market price reflects their value for financing purposes (that is, an issue may be on special in the repo market). Some models use the London Interbank Offered Rate (LIBOR) curve instead of the Treasury curve. The reason is that some investors are interested in spreads that they can earn relative to their funding costs and LIBOR, for many investors, is a better proxy for that cost than Treasury rates.

To generate the interest rate paths, an assumption about the evolution of future interest rates is required. Most Monte Carlo simulation models use some form of one-factor interest rate model. The one factor used is the short-term interest rate. When using a particular one-factor interest rate model, several further assumptions must be made. The first, and the most important, is the assumption about the volatility of the short-term interest rate. The volatility assumption determines the dispersion of future interest rates in the simulation. Many model developers do not use one volatility number for the yield volatility of all maturities for the benchmark curve. Instead, they use either a short/long yield volatility or a term structure of yield volatility. A short/long yield volatility means that volatility is specified for maturities up to a certain number of years (short yield volatility) and a different yield volatility for greater maturities (long yield volatility). The

short yield volatility is assumed to be greater than the long yield volatility. A term structure of yield volatilities means that a yield volatility is assumed for each maturity. (In practice, interest rate volatility is extracted from interest rate cap market prices.) From these prices, a term structure of yield volatility is obtained. Differences in the assumption about volatility of short-term interest rates can have a material impact on the resulting value derived for the security.

Another assumption relates to the speed of mean reversion of the short-term interest rate. Mean reversion in an interest rate model has to do with not allowing interest rates to fall below a lower barrier and not exceed an upper barrier before rates revert back to some average interest rate specified by the model developer or user.

The random paths of interest rates should be generated from an arbitrage-free model of the future term structure of interest rates. By arbitrage free it is meant that the model replicates today's term structure of interest rates, an input of the model, and that for all future dates there is no possible arbitrage within the model.

The simulation works by generating many scenarios of future interest rate paths. In each month of a given scenario (that is, path), a monthly interest rate and a refinancing rate are generated. The monthly interest rates are used to discount the projected cash flows in the scenario. The refinancing rate is needed to determine the cash flows because it represents the opportunity cost the borrower is facing at that time.

If the refinancing rates are high relative to the borrower's loan rate, the borrower will have no incentive to refinance. For MBS and mortgage-related ABS, there is a disincentive to prepay (that is, the homeowner may avoid moving in order to avoid refinancing). If the refinancing rate is low relative to the borrower's loan rate, the borrower has an incentive to refinance.

Prepayments (voluntary and involuntary) and recoveries are projected by feeding the refinancing rate and loan characteristics into a

Table 1 Simulated Paths of One-Month Future Interest Rates

Month	Interest Rate Path Number						
	1	2	3	...	n	...	N
1	$f_1(1)$	$f_1(2)$	$f_1(3)$...	$f_1(n)$...	$f_1(N)$
2	$f_2(1)$	$f_2(2)$	$f_2(3)$...	$f_2(n)$...	$f_2(N)$
3	$f_3(1)$	$f_3(2)$	$f_3(3)$...	$f_3(n)$...	$f_3(N)$
...
t	$f_t(1)$	$f_t(2)$	$f_t(3)$...	$f_t(n)$...	$f_t(N)$
...
$M-2$	$f_{M-2}(1)$	$f_{M-2}(2)$	$f_{M-2}(3)$...	$f_{M-2}(n)$...	$f_{M-2}(N)$
$M-1$	$f_{M-1}(1)$	$f_{M-1}(2)$	$f_{M-1}(3)$...	$f_{M-1}(n)$...	$f_{M-1}(N)$
M	$f_M(1)$	$f_M(2)$	$f_M(3)$...	$f_M(n)$...	$f_M(N)$

Notation: $f_t(n)$ = one-month future interest rate for month t on path n , N = total number of interest rate paths; M = number of months for the loan pool.

prepayment model and default model. (In the case of agency MBS [Ginnie Mae, Fannie Mae, and Freddie Mac] no assumption about defaults is required.) Given the projected prepayments, the cash flows along an interest rate path can be determined. To be able to do this, the entire deal must be reverse engineered. That is, the deal's waterfall (that is, the rules for distribution of interest, principal repayment, and loss allocation) must be specified so that the cash flow for the bond class being valued can be determined. Model developers do not reverse engineer the deals. Rather, there are vendors who provide the waterfall for deals that are used in conjunction with the Monte Carlo simulation model.

To make this more concrete, consider a newly issued loan pool with a maturity of M months

that is the collateral for an MBS or mortgage-related ABS. Table 1 shows N simulated interest rate path scenarios. Each scenario consists of a path of M simulated 1-month future interest rates. (The determination of the number of paths generated is based on a variance-reduction method.⁵) So, the first assumption made to generate the short-term interest rate paths in Table 1 is the volatility of short-term interest rates.

Table 2 shows the paths of simulated refinancing rates corresponding to the scenarios shown in Table 1. In going from Table 1 to Table 2, an assumption must be made about the relationship between the benchmark short-term interest rate and the refinancing rate. The assumption is that there is a constant spread relationship between the refinancing rate and the interest rate for a

Table 2 Simulated Paths of Refinancing Rates

Month	Interest Rate Path Number						
	1	2	3	...	n	...	N
1	$r_1(1)$	$r_1(2)$	$r_1(3)$...	$r_1(n)$...	$r_1(N)$
2	$r_2(1)$	$r_2(2)$	$r_2(3)$...	$r_2(n)$...	$r_2(N)$
3	$r_3(1)$	$r_3(2)$	$r_3(3)$...	$r_3(n)$...	$r_3(N)$
...
t	$r_t(1)$	$r_t(2)$	$r_t(3)$...	$r_t(n)$...	$r_t(N)$
...
$M-2$	$r_{M-2}(1)$	$r_{M-2}(2)$	$r_{M-2}(3)$...	$r_{M-2}(n)$...	$r_{M-2}(N)$
$M-1$	$r_{M-1}(1)$	$r_{M-1}(2)$	$r_{M-1}(3)$...	$r_{M-1}(n)$...	$r_{M-1}(N)$
M	$r_M(1)$	$r_M(2)$	$r_M(3)$...	$r_M(n)$...	$r_M(N)$

Notation: $r_t(n)$ = refinancing rate for month t on path n ; N = total number of interest rate paths; M = number of months for the loan pool.

Table 3 Simulated Cash Flows for the Loan Pool

Month	Interest Rate Path Number						
	1	2	3	...	n	...	N
1	$C_1(1)$	$C_1(2)$	$C_1(3)$...	$C_1(n)$...	$C_1(N)$
2	$C_2(1)$	$C_2(2)$	$C_2(3)$...	$C_2(n)$...	$C_2(N)$
3	$C_3(1)$	$C_3(2)$	$C_3(3)$...	$C_3(n)$...	$C_3(N)$
...
t	$C_t(1)$	$C_t(2)$	$C_t(3)$...	$C_t(n)$...	$C_t(N)$
...
$M-2$	$C_{M-2}(1)$	$C_{M-2}(2)$	$C_{M-2}(3)$...	$C_{M-2}(n)$...	$C_{M-2}(N)$
$M-1$	$C_{M-1}(1)$	$C_{M-1}(2)$	$C_{M-1}(3)$...	$C_{M-1}(n)$...	$C_{M-1}(N)$
M	$C_M(1)$	$C_M(2)$	$C_M(3)$...	$C_M(n)$...	$C_M(N)$

Notation: $C_t(n)$ = loan pool's cash flow for month t on path n ; N = total number of interest rate paths; M = number of months for the loan pool.

maturity that is the best proxy for the borrowing rate. Typically, it is the 10-year rate that is used as a proxy.

Given the refinancing rates, the collateral's cash flows on each interest rate path can be generated. This requires a prepayment and default/recovery model. So our next assumption is that the prepayment and default/recovery models used to generate the loan pool's cash flows are correct. The resulting cash flows are depicted in Table 3.

Given the loan pool's cash flow for each month on each interest rate path, the next step is to use the waterfall for the structure to determine how the cash flow is distributed to the bond class being valued. Let us use BCC to denote the cash flow for that bond class. Table 4 shows the simulated cash flows on each of

the interest rate paths for the bond class being valued.

Calculating the Present Value of a Bond Class for a Scenario Interest Rate Path

Given the cash flows for the bond class on an interest rate path, the path's present value can be calculated. The discount rate for determining the present value is the simulated spot rate for each month on the interest rate path plus an appropriate spread. The spot rate on a path can be determined from the simulated future monthly rates. The relationship that holds between the simulated spot rate for month t on path n and the simulated future one-month rates is:

$$z_t(n) = \{[1 + f_1(n)][1 + f_2(n)] \cdots [1 + f_t(n)]\}^{1/t} - 1$$

Table 4 Simulated Cash Flows for the Bond Class Being Valued

Month	Interest Rate Path Number						
	1	2	3	...	n	...	N
1	$BCC_1(1)$	$BCC_1(2)$	$BCC_1(3)$...	$BCC_1(n)$...	$BCC_1(N)$
2	$BCC_2(1)$	$BCC_2(2)$	$BCC_2(3)$...	$BCC_2(n)$...	$BCC_2(N)$
3	$BCC_3(1)$	$BCC_3(2)$	$BCC_3(3)$...	$BCC_3(n)$...	$BCC_3(N)$
...
t	$BCC_t(1)$	$BCC_t(2)$	$BCC_t(3)$...	$BCC_t(n)$...	$BCC_t(N)$
...
$M-2$	$BCC_{M-2}(1)$	$BCC_{M-2}(2)$	$BCC_{M-2}(3)$...	$BCC_{M-2}(n)$...	$BCC_{M-2}(N)$
$M-1$	$BCC_{M-1}(1)$	$BCC_{M-1}(2)$	$BCC_{M-1}(3)$...	$BCC_{M-1}(n)$...	$BCC_{M-1}(N)$
M	$BCC_M(1)$	$BCC_M(2)$	$BCC_M(3)$...	$BCC_M(n)$...	$BCC_M(N)$

Notation: $BCC_t(n)$ = bond class's cash flow for month t on path n ; N = total number of interest rate paths; M = number of months for the loan pool.

Table 5 Simulated Paths of Monthly Spot Rates

Month	Interest Rate Path Number						
	1	2	3	...	n	...	N
1	$z_1(1)$	$z_1(2)$	$z_1(3)$...	$z_1(n)$...	$z_1(N)$
2	$z_2(1)$	$z_2(2)$	$z_2(3)$...	$z_2(n)$...	$z_2(N)$
3	$z_3(1)$	$z_3(2)$	$z_3(3)$...	$z_3(n)$...	$z_3(N)$
...
t	$z_t(1)$	$z_t(2)$	$z_t(3)$...	$z_t(n)$...	$z_t(N)$
...
$M-2$	$z_{M-2}(1)$	$z_{M-2}(2)$	$z_{M-2}(3)$...	$z_{M-2}(n)$...	$z_{M-2}(N)$
$M-1$	$z_{M-1}(1)$	$z_{M-1}(2)$	$z_{M-1}(3)$...	$z_{M-1}(n)$...	$z_{M-1}(N)$
M	$z_M(1)$	$z_M(2)$	$z_M(3)$...	$z_M(n)$...	$z_M(N)$

Notation: $z_t(n)$ = spot rate for month t on path n ; N = total number of interest rate paths; M = number of months for the loan pool.

where

$$z_t(n) = \text{simulated spot rate for month } t \text{ on path } n$$

$$f_j(n) = \text{simulated future one-month rate for month } j \text{ on path } n$$

Consequently, the interest rate path for the simulated future one-month rates can be converted to the interest rate path for the simulated monthly spot rates as shown in Table 5. Therefore, the present value of the cash flows for month t on interest rate path n discounted at the simulated spot rate for month t plus some spread is:

$$PV[BCC_t(n)] = \frac{BCC_t(n)}{[1 + z_t(n) + K]^t} \quad (1)$$

where

$$PV[BCC_t(n)] = \text{present value of the cash flow for the bond class for month } t \text{ on path } n$$

$$BCC_t(n) = \text{cash flow for the bond class for month } t \text{ on path } n$$

$$z_t(n) = \text{spot rate for month } t \text{ on path } n$$

$$K = \text{spread}$$

The present value for path n is the sum of the present value of the cash flows for each month

on path n . That is,

$$PV[\text{Path}(n)] = PV[BCC_1(n)] + PV[BCC_2(n)] + \dots + PV[BCC_M(n)] \quad (2)$$

where $PV[\text{Path}(n)]$ is the present value of interest rate path n .

Determining the Theoretical Value

The present value of a given interest rate path is treated as the *theoretical value* of a bond class if that path is realized. The theoretical value of the bond class using the Monte Carlo simulation model is determined by calculating the average of the theoretical values of all the interest rate paths. That is, the theoretical value is equal to

$$\text{Theoretical value} = \frac{PV[\text{Path}(1)] + \dots + PV[\text{Path}(N)]}{N} \quad (3)$$

where N is the number of interest rate paths.

Notice that the results of the Monte Carlo simulation model produce one value, the average value, and that value is taken as the theoretical value. However, as noted earlier, the purpose of a Monte Carlo simulation model is to estimate the probability distribution for the variable of interest. While a probability distribution can easily be obtained from the values for each path and summary information in addition to the mean such as dispersion and skewness

measures can be computed, it is rare if that information is provided. Basically, the reason is that investors rarely seek that information because too often they do not understand the Monte Carlo simulation process.

Moreover, it should be apparent how the Monte Carlo simulation model is driven by assumptions. Hence, a user of a model such as the one described here is subject to *modeling risk*. To mitigate modeling risk, an investor can test the sensitivity of the value produced by the model to alternative assumptions. For example, regarding the volatility assumption, the model can be rerun assuming both proportionality lower and higher volatility than initially assumed. The sensitivity to prepayments can be analyzed in the same way. From the sensitivity analysis, an investor can determine which assumptions appear to be more important for the security being considered for purchase.⁶

Option-Adjusted Spread

Thus far we have seen how the theoretical value of a security can be determined using the Monte Carlo simulation model. Recall that in the model, a spread (K) is added to the monthly spot rates on all the interest rate paths in Table 5 in order to determine the discount rate used for calculating the present value of the cash flows. The spread should reflect the risk associated with the security as required by the market. However, the reverse can be done. Given (1) the cash flows in Table 4 for the bond class being valued, (2) the spot rates in Table 5, and (3) the market price of the security being valued, one can determine the spread that will make the average value for the interest rate paths equal to the market price (plus accrued interest). That spread is what is referred to as the *option-adjusted spread* (OAS). Mathematically, OAS is the spread that will make

$$\text{Market price} + \text{Accrued interest} = \frac{\text{PV}[\text{Path}(1)] + \dots + \text{PV}[\text{Path}(N)]}{N} \quad (4)$$

where N is the number of interest rate paths.

Basically, the OAS is used to reconcile the model's value [that is, the value determined by the Monte Carlo simulation model given by equation (3)] with the market price. On the left-hand side of equation (4) is the market's valuation of the security as represented by the market price. On the right-hand side of the equation is the model's evaluation of the security (that is, the theoretical value), which is the average present value over all the interest rate paths. Basically, the OAS was developed as a measure of the spread that can be used to convert dollar differences between model value and market price. But what is it a "spread" over? In describing the model above, we can see that the OAS is measuring the average spread over the benchmark spot rate. It is an average spread since the OAS is found by averaging over the interest rate paths for the possible future benchmark spot rate curves.

This spread measure is superior to the nominal spread, which gives no recognition to the prepayment risk. The OAS is "option adjusted" because the cash flows on the interest rate paths are adjusted for the option of the borrowers to prepay.

Option Cost

The implied cost of the option embedded in a security can be obtained by calculating the difference between the OAS and the zero-volatility spread. That is,

$$\text{Option cost} = \text{Zero-volatility spread} - \text{OAS}$$

The *option cost* measures the prepayment (or option) risk embedded in MBS and ABS. Note that the cost of the option is a by-product of the OAS analysis, not valued explicitly with some option pricing model.

When the option cost is zero because the borrower tends not to exercise the prepayment option when interest rates decline below the loan rate or when there is no prepayment option, then substituting zero for the OAS in the previous equation and solving for the zero-volatility

spread, we get:

$$\text{Zero-volatility spread} = \text{OAS}$$

Consequently, when the value of the option is zero (that is, the option cost is zero) for a particular ABS, simply computing the zero-volatility spread for relative value purposes or for valuing that ABS is sufficient. Even if there is a small value for the option, the zero-volatility spread should be adequate rather than calculating an OAS using the Monte Carlo simulation model.

Simulated Average Life

The average life of a security when using the Monte Carlo simulation model is the weighted average time to receipt of principal payments (scheduled payments and projected prepayments). The average life reported in a Monte Carlo model is the average of the average lives along the interest rate paths. That is, for each interest rate path, there is an average life. The average of these average lives is the average life reported by the model.

Additional information is conveyed by the distribution of the average life. The greater the range and standard deviation of the average life, the more uncertainty there is about the security's average life.

MEASURING INTEREST RISK

There are two measures of interest rate risk that are commonly used: duration and convexity.⁷ Duration is a first approximation as to how the value of an individual security or the value of a portfolio will change when interest rates change. Convexity measures the change in the value of a security or portfolio that is not explained by duration. How these measures are computed when using the Monte Carlo simulation model is described in this section.

Duration

The most obvious way to measure a bond's price sensitivity as a percentage of its current

price to changes in interest rates is to change rates by a small number of basis points and calculate how its price will change. To do this, we introduce the following notation. Let

V_0 = initial value or price of the security

Δy = change in the yield of the security (in decimal)

V_- = the estimated value of the security if the yield is decreased by Δy

V_+ = the estimated value of the security if the yield is increased by Δy

There are two key points to keep in mind in the foregoing discussion. First, the change in yield referred to above is the same change in yield for all maturities. This assumption is commonly referred to as a "parallel yield curve shift assumption." Thus, the foregoing discussion about the price sensitivity of a security to interest rate changes is limited to parallel shifts in the yield curve. Second, the notation refers to the estimated value of the security. This value is obtained from a valuation model. Consequently, the resulting measure of the price sensitivity of a security to interest rate changes is only as good as the valuation model employed to obtain the estimated value of the security.

Now let's focus on the measure of interest. We are interested in the percentage change in the price of a security when interest rates change. This measure is referred to as duration. It can be demonstrated that duration can be estimated using the following formula:

$$\text{Duration} = \frac{V_- - V_+}{2V_0(\Delta y)} \quad (5)$$

The duration of a security can be interpreted as the approximate percentage change in price for a 100 basis point parallel shift in the yield curve. Thus, a bond with a duration of 5 will change by approximately 5% for a 100 basis point parallel shift in the yield curve. For a 50 basis point parallel shift in the yield curve, the bond's price will change by approximately 2.5%; for a 25 basis point parallel shift in the yield curve, 1.25%, and so on.

What this means is that in calculating the values of V_- and V_+ in the duration formula, the same cash flows used to calculate V_0 are used. Therefore, the change in the bond's price when the yield curve is shifted by a small number of basis points is due solely to discounting at the new yields. This assumption makes sense for option-free bonds such as Treasury securities and nonmortgage ABS such as credit card ABS and auto loan-backed ABS. However, the same cannot be said for MBS and mortgage-related ABS because for these products the cash flows are sensitive to changes in interest rates. Rather, for these products a change in yield will alter the expected cash flows because it will change expected payments.

The Monte Carlo simulation model takes into account how parallel shifts in the yield curve will affect the cash flows. Thus, when V_- and V_+ are the values produced from the valuation model, the resulting duration takes into account both the discounting at different interest rates and how the cash flows can change. When duration is calculated in this manner, it is referred to as *effective duration* or *option-adjusted duration*.

To calculate effective duration, the value of a security must be estimated when interest rates are shocked (that is, changed) up and down a given number of basis points. In terms of the Monte Carlo simulation model, the yield curve used is shocked up and down and the new curve is used to generate the values to be used in equation (5) to obtain effective duration.

There are two important aspects of this process of generating the values when the rates are shocked that are critical to understand. First, the assumption is that the relationships assumed do not change when rates are shocked up and down. Specifically, (1) the interest rate volatility is assumed to be unchanged to derive the new interest rate paths for a given shock (that is, the new Table 1), as well as the other assumptions made to generate the new Table 2 from the newly constructed Table 1, and (2) the OAS is assumed to be constant. The constancy of the OAS comes into play because when discount-

ing the new cash flows (that is, the cash flows in the new Table 4), the current OAS that was computed is assumed to be the same and is added to the new rates in the new Table 1.

Convexity

The duration measure indicates that regardless of whether interest rates increase or decrease, the approximate percentage price change is the same. However, this does not agree with the price volatility property of a bond. Specifically, while for small changes in yield the percentage price change will be the same for an increase or decrease in yield, for large changes in yield this is not true. This suggests that duration is only a good approximation of the percentage price change for a small change in yield.

The reason for this result is that duration is in fact a first approximation for a small change in yield. The approximation can be improved by using a second approximation. This approximation is referred to as "convexity." (The use of this term in the industry is unfortunate since the term "convexity" is also used to describe the shape or curvature of the price/yield relationship.) The convexity measure of a security can be used to approximate the change in price that is not explained by duration.

The convexity measure of a bond can be approximated using the following formula:

$$\text{Convexity measure} = \frac{V_+ + V_- - 2V_0}{2V_0(\Delta y)^2} \quad (6)$$

where the notation is the same as used earlier for duration. When the values for the inputs in the convexity measure as given in equation (6) are obtained from a Monte Carlo simulation model, the resulting convexity is referred to as *effective convexity*. Note that dealers often quote convexity by dividing the convexity measure by 100.

When the convexity measure is positive, we have the situation where the gain is greater than the loss for a given large change in rates. That is, the security exhibits *positive convexity*. Most nonmortgage ABS have positive convexity.

However, if the convexity measure is negative, we have the situation where the loss will be greater than the gain. A security with this characteristic is said to have *negative convexity* and it occurs with MBS and mortgage-related ABS.

KEY POINTS

- Valuing securities with interest rate-sensitive options requires the employment of a model that recognizes how future interest rates can change and how that impacts the expected cash flows.
- For bonds with embedded options such as callable bonds and puttable bonds, as well as bonds that have an accelerated sinking fund provision, the lattice method can be used. Unfortunately, the lattice model cannot be used for MBS and mortgage-related ABS because these securities have path-dependent cash flows and thus how interest rates have evolved prevents solving a lattice model.
- Instead of the lattice model, the Monte Carlo simulation model is used to value MBS and mortgage-related ABS. There are many assumptions in the model and therefore, sensitivity analysis should be used to test the sensitivity of the model's value to changes in the major assumptions.
- For ABS that do not have an embedded option (that is, no prepayment option) or where there is a prepayment option but for all intents and purposes the prepayment option is unlikely to be exercised, valuation is fairly straightforward—assuming a good model for estimating defaults and recoveries. It is simply the present value of the expected cash flow discounted at the benchmark spot rates plus an appropriate spread.
- The cash-flow yield measure for MBS and ABS is a flawed measure of value. The corresponding nominal spread is therefore similarly flawed. A better measure for ABS where the prepayment option has little value is the zero-volatility spread. For MBS and mortgage-related ABS, the commonly used

measure is the OAS. This measure adjusts the spread for the embedded option by adjusting the cash flows in the Monte Carlo simulation model (as well as in the lattice model).

- Because the OAS is derived from the Monte Carlo simulation model, it is also an assumption-driven product and therefore subject to modeling risk.
- The appropriate interest risk measures for MBS and mortgage-related ABS are effective duration and effective convexity. These measures require, as inputs, the estimated value of the security obtained by shocking the Monte Carlo simulation model.

NOTES

1. For a discussion of MBS, see Fabozzi, Bhattacharya, and Berliner (2011). Asset-backed securities are described in Fabozzi (2012).
2. For a discussion of prepayment models for MBS, see Fabozzi, Bhattacharya, and Berliner (2011).
3. PACs are described in Fabozzi, Bhattacharya, and Berliner (2011).
4. For applications of Monte Carlo simulation to finance, see Pachamanova and Fabozzi (2010).
5. Variance-reduction methods in Monte Carlo simulation are explained in Pachamanova and Fabozzi (2010).
6. For an illustration applied to an actual CMO transaction, see Fabozzi, Richard, and Horowitz (2006).
7. For an explanation of duration and convexity, see Fabozzi (1999, 2011).

REFERENCES

- Fabozzi, F. J. (1999). *Duration, Convexity, and Other Bond Risk Measures*. Hoboken, NJ: John Wiley & Sons.
- Fabozzi, F. J. (2012). *Bond Markets, Analysis, and Strategies*, 8th ed. Upper Saddle River, NJ: Pearson.
- Fabozzi, F. J., Bhattacharya, A. K., and Berliner, W. S. (2011). *Mortgage-Backed Securities:*

- Products, Structuring, and Analytics Techniques*, 2nd ed. Hoboken, NJ: John Wiley & Sons.
- Fabozzi, F. J., Richard, S. F., and Horowitz, D. S. (2006). Valuation of mortgage-backed securities. In F. J. Fabozzi (ed.), *The Handbook of Mortgage-Backed Securities*, 6th ed. (pp. 759–781). New York: McGraw Hill.
- Kalotay, A., Yang, D., and Fabozzi, F. J. (2004). An option-theoretic prepayment model for mortgages and mortgage-backed securities. *International Journal of Theoretical and Applied Finance* 7, 8: 949–978.
- Pachamanova, D., and Fabozzi, F. J. (2010). *Simulation and Optimization Modeling in Finance*. Hoboken, NJ: John Wiley & Sons.

The Active-Passive Decomposition Model for MBS

ALEXANDER LEVIN, PhD

Director, Financial Engineering, Andrew Davidson & Co., Inc.

Abstract: Even a simple mortgage pass-through is a path-dependent financial instrument, valuation of which depends on prepayment “burnout.” The burnout is caused by observed or unobserved heterogeneity of borrowers; as a result, a mortgage pool’s composition changes in the presence of refinancing incentives. An attractive modeling approach for dealing with this is to split a mortgage pool into mutually exclusive “active” and “passive” groups. Not only does such a method explain the burnout, it effectively decomposes the path-dependent valuation problem into two easy-to-solve path-independent ones. The method is faster than the traditional Monte Carlo sampling approach while delivering the full set of interest rate risk measures at no additional cost of computing time. The method can be applied to an attractive prepayment model specification where the speed is a function of the pool’s objective price, and not an interest rate. This makes universal refinancing modeling feasible as the same curve or curves can apply to both fixed- and adjustable-rate mortgages.

The *active-passive decomposition (APD) method* of mortgage-backed securities (MBS) modeling and valuation was introduced in Levin (2001, 2002, 2003). An efficient alternative to brute-force Monte Carlo simulation, the APD method splits a mortgage pass-through into two path-independent components, the active (refinanceable) and the passive (nonrefinanceable). Once this is done, the most time-efficient pricing structures operating backwards on probability trees or finite-difference grids could be employed. This valuation method runs faster than Monte Carlo simulation while deliver-

ing a much richer outcome—all stressed values required by mandatory risk assessments—at no additional cost. Risk managers and traders of unstructured mortgage instruments such as agency pass-through MBS, whole loans, stripped (IO/PO) derivatives, and mortgage servicing rights (MSRs) are immediate beneficiaries of the method.

The APD approach simulates the *burnout* effect in a natural and explicit way through modeling the heterogeneity of the collateral. Hence, it presents an analytical advantage over any other approach that requires ad hoc judgments

The extended APD model and its implementation presented here has greatly benefited from joint work with Andrew Davidson and Dan Szakallas.

about the achieved degree of burnout. Structured instruments—such as a collateralized mortgage obligation (CMO) and asset-backed security (ABS)—though they retain heavy sources of path-dependence (other than the burnout) and still rely on Monte Carlo pricing, can benefit from better, more robust prepay modeling.

The multi-population view of mortgage collateral is a known approach used to explain the burnout effect. In one of the earliest modeling attempts, Davidson (1987) and Davidson et al. (1988) proposed the *refinancing* threshold model, in which collateral is split into three or more American option bonds having differing strikes. A conceptually similar approach proposed by Kalotay and Young (2002) divides collateral into bonds differing by their exercise timing. Such structures naturally call for the *backward induction* pricing, but they fall short in replicating actually observed, probabilistically smoothed, prepayment behavior—even if many constituent bonds are used. On the other hand, analytical systems used in practice often employ multi-population mortgage models (see Hayre, 1994, 2000), but do not seek any computational benefits as they rely heavily on Monte Carlo simulation pricing anyway.

The APD is a “mortgage-like” model with refinancing S-curve, aging, and other ad hoc features, which are meant to capture nonefficient, empirical option exercise. Therefore, the APD model is capable of generating realistic prepayment behavior with only two constituent components, the active and the passive. This entry introduces an extended APD model and its applications.

PATH-DEPENDENCE AND PRICING PARTIAL DIFFERENTIAL EQUATION

Let us consider a hypothetical dynamic asset (“mortgage”) market price of which $P(t, x)$ depends on time t and one market factor x . The latter can be formally anything and does not

necessarily have to be the short market rate or the yield on the security analyzed. We treat $x(t)$ as a random process having a (generally, variable) drift rate μ and a volatility rate σ , and being disturbed by a standard Brownian motion $z(t)$, that is,

$$dx = \mu dt + \sigma dz \quad (1)$$

We assume further that the asset continuously pays the $c(t, x)$ coupon rate and its balance B is amortized at the $\lambda(t, x)$ rate, that is, $\partial B/\partial t = -\lambda B$. Then one can prove that the price function $P(t, x)$ should solve the following *partial differential equation* (PDE):

$$\underbrace{\frac{r + OAS}{P}}_{\text{expected return}} = \underbrace{\frac{1}{P} \frac{\partial P}{\partial t} + \frac{1}{P}(c + \lambda) - \lambda}_{\text{time return}} + \underbrace{\frac{1}{P} \frac{\partial P}{\partial x} \mu}_{\text{return}} + \underbrace{\frac{1}{2P} \frac{\partial^2 P}{\partial x^2} \sigma^2}_{\text{return}} \quad (2)$$

A derivation of this PDE can be found in Levin (1998), but it goes back at least to Fabozzi and Fong (1994). A notable feature of the above written PDE is that it does not contain the balance variable, B . The entire effect of possibly random prepayments is represented by the amortization rate function, $\lambda(t, x)$. Although the total cash flow observed for each accrual period does depend on the beginning-period balance, construction of a finite-difference scheme and the backward induction will require the knowledge of $\lambda(t, x)$, not the balance. This observation agrees with a trivial practical rule stating that the relative price is generally independent of the investment size.

Another interesting observation comes as follows. If we transform the economy having shifted all the rates, $r(t, x)$ and $c(t, x)$, by amortization rate $\lambda(t, x)$, then PDE (2) will be reduced to the constant-par asset’s pricing PDE. It means that a probability tree or finite difference *pricing grid* built in the “ λ -shifted” economy should, in principle, have as many dimensions as the total number of factors or state variables that affect r , c , and λ . In particular, if the coupon rate is fixed, and the amortization

rate λ depends only on current time (loan age) and the immediate market factor x , the entire valuation problem can be solved backwards on a two-dimensional (x, t) lattice. To implement this method, we would start our valuation process from maturity T when we surely know that the price is par, $P(T, x) = 1$, regardless the value of factor x .

Working backwards, we derive prices at age $t - 1$ from prices already found at age t . In doing so, we replace derivatives in PDE (2) by finite difference approximations, or weigh branches of the lattice by explicitly computed probabilities. If the market is multifactor, then x should be considered a vector; the lattice will require more dimensions. Generally, the efficiency of finite-difference methods deteriorates quickly on high-dimensional grids because the number of nodes and cash flows grows geometrically; probability trees may maintain their speed, but at the cost of accuracy, if the same number of emanating nodes is used to capture multifactor dynamics. If we decide to operate on a probability tree instead of employing a finite-difference grid, then, for every branch,

$$P_k = \frac{c_k + P_{k+1} + \lambda_k(1 - P_{k+1})}{1 + r_k + OAS} \quad (3)$$

where P_k is the previous-node value deduced from the next-node value P_{k+1} . Of course, probability weighting of thus obtained values applies to all emanating branches.

EXTENDED ACTIVE-PASSIVE DECOMPOSITION MODEL

Even for a simple fixed-rate mortgage pass-through, total amortization speed λ cannot be modeled as a function of time and the immediate market. Prepayment burnout is a strong source of path-dependence because the future refinancing activity is affected by the past incentives. One can think of a mortgage pool as of a heterogeneous population of participants having different refinancing propensities. Some borrowers have higher rate, better credit,

larger loans, or perhaps they face smaller state-enforced transaction costs. Once they leave the pool, the future prepayment activity gradually declines.

Instead of considering pricing PDE for the entire collateral, we propose decomposing it first into two components, "active" and "passive," differing in refinancing activity. Under the following two conditions, mortgage path-dependent collateral can be deemed a simple portfolio of two path-independent instruments:

1. Active and passive components prepay differently, but follow the immediate market and loan age.
2. Any migration between components is prohibited.

The Details

Here is a permissible example:

$$\begin{aligned} \text{ActiveSMM} &= \text{RefiSMM} + \text{TurnoverSMM} \\ \text{PassiveSMM} &= \beta^* \text{RefiSMM} + \text{TurnoverSMM} \end{aligned} \quad (4)$$

where RefiSMM denotes refinancing speed measured in terms of the single monthly mortality rate (SMM), TurnoverSMM is the turnover speed, and both are assumed to depend on market rates and loan age only. Parameter β quantifies relative refinancing activity for the passive component; it takes values between 0 and 1.

In order to find the total speed, we have to know the collateral composition. Denote ψ the ratio of active group to total, then

$$\begin{aligned} \lambda \equiv \text{TotalSMM} &= \psi^* \text{ActiveSMM} \\ &+ (1 - \psi)^* \text{PassiveSMM} \end{aligned} \quad (5)$$

All variables are time-dependent, but we omitted subscript t for simplicity. The initial value of ψ describes the composition of collateral at origination; both ψ_0 and β are parameters for the particular prepay model. The dynamic evolution of ψ from one time moment (t) to the next ($t + 1$) is as follows

$$\psi_{t+1} = \psi_t \frac{1 - \text{ActiveSMM}_t}{1 - \text{TotalSMM}_t} \quad (6)$$

It is worth considering a few trivial special cases. First, if ψ is zero at any instance of time, it will remain zero for life. Second, if ψ is 1 at any time, then it will retain this value as well because TotalSMM is identical to ActiveSMM from equation (5). Indeed, if the mortgage pool is either totally passive ($\psi = 0$) or totally active ($\psi = 1$), it will retain its status due to the complete absence of migration. In either of these two special cases, variables ψ and TotalSMM are path-independent, leading us to a key conclusion: The separate consideration of active and passive components avoids the problem of path-dependence altogether.

How the Model Works Forward

If $0 < \psi < 1$, then TotalSMM < ActiveSMM, the fraction in the right-hand side of formula (6) is less than 1, and ψ gradually falls. If we employed the APD model for prepay modeling while using Monte Carlo simulation for valuation, we could innovate compositional variable ψ month after month. First, we would compute refinancing and turnover speeds at time t from their respective models. Then, we would produce active, passive, and total speeds, all still at time t , from formulas (4) and (5). This information is not only sufficient to generate the t -month cash flow, but it also allows for finding the next-month composition, ψ_{t+1} , from formula (6), and proceeding forward.

Note that prepay speeds RefiSMM and TurnoverSMM depend only on current market rates and time, that is, they are path-independent. Naturally, ActiveSMM and PassiveSMM found from (4) will be path-independent as well. In contrast, variables ψ and TotalSMM are generally path-dependent except when ψ is either 0 or 1.

Let us visualize how the APD model works. Suppose we have a pool with $\psi_0 = 0.8$, that is, the active part constitutes 80% of the total at origination. Consider two possible scenarios:

Scenario A: Rates drop and remain low, inducing refinancing activity.

Scenario B: Rates rise and remain high.

Figures 1A and 1B show how the pool composition will evolve in these two cases. For scenario A, pool balance is amortized quickly due to the refinancing wave, but, more importantly, the active group (darker bars) evaporates much faster than the passive group (lighter bars). As the result, variable ψ drops from the original 80% to under 30% and, correspondingly, the total speed (as measured by conditional prepayment rate and denoted by CPR) declines—in the complete absence of any rate dynamics. A sizable speed reduction from 45 CPR to 30 CPR is caused exclusively by the burnout effect and reflected by ψ . This effect is not seen in scenario B where the active and the passive groups retire at similar rates. Pool composition barely changes, as does the total prepayment speed.

We could give prepayment behaviors depicted in Figures 1A and 1B another interesting practical interpretation. Let us assume that we wish to compare a regular fixed-rate pool (Figure 1A) with a prepayment-penalty pool (Figure 1B) under the same low-rate market conditions. The regular pool burns out—unlike the prepay-penalty one, which faces additional refinancing barriers. At the end of its penalty window (assume 60 months), this pool retains a relatively high level of ψ (71.7%). Looking at a matching speed level in Figure 1A, we conclude that, once the penalty window is over, the prepay speed will jump above 40 CPR (compared to 29 CPR of the regular pool). Therefore, the APD model naturally explains the “catch-up” effect actually known for prepay-penalty mortgages.

Above, we assumed a newly originated pool, the population of which is determined by parameter ψ_0 . In practice, a pool may be already seasoned, and today’s value of ψ , denote it $\psi(t_0)$, needs to be determined first. We will cover this task shortly.

How the Model Works in Backward Induction

If we decide to employ the APD model for backward valuation, we do not need to innovate path-dependent variables, ψ , and TotalSMM,

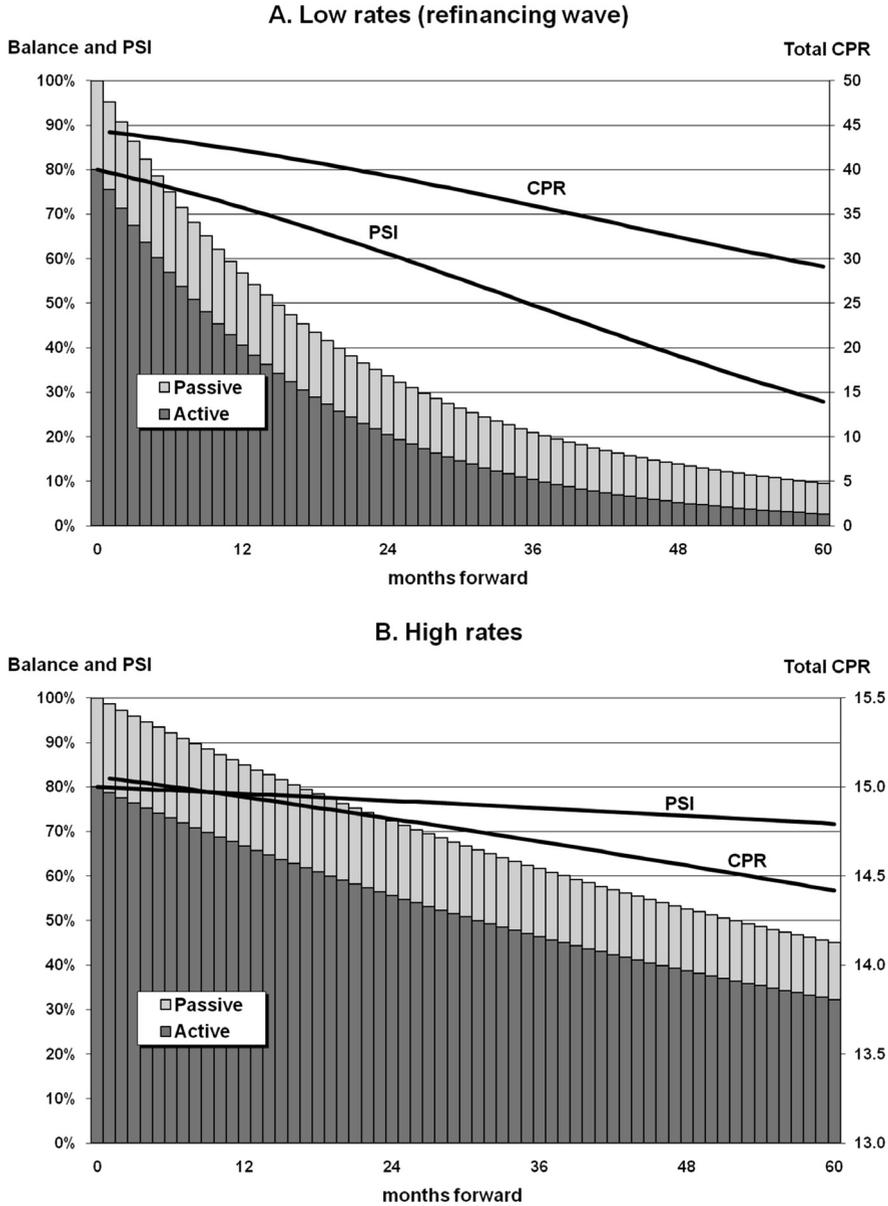


Figure 1 Simple APD Model: How It Works Forward

or keep track of their dynamics. Here are few simple steps to perform:

Step 1: Recover today’s value of the population variable, $\psi(t_0)$.

Step 2 Active: Generate cash flows on each node of a pricing grid (tree) for the active part

only and value it using a backward inducting scheme that solves pricing equation (2).

Step 2 Passive: Do the same for the passive part.

Step 3: Combine thus obtained values as

$$P = \psi(t_0)P_{active} + [1 - \psi(t_0)]P_{passive} \quad (7)$$

Interestingly enough, formula (7) applies to today’s prices obtained for all interest rate

levels of the pricing grid. As we mentioned above, computing prices on the entire grid is an inseparable part of backward valuation. Therefore, the total price can be also found on the grid, at no additional cost. In particular, the measures representing the sensitivities of an MBS price to the interest rates are found immediately, without any repetitive efforts with a stressed market (compare to Monte Carlo simulation). However, we can't apply formula (7) for future nodes because we know only $\psi(t_0)$ —today's value of ψ .

Initializing the Burnout Factor

If the pool is already seasoned, we have to assess $\psi(t_0)$ first before we can employ the APD model either for forward simulation or backward induction. There exist two main approaches to solve this problem: an analytical closed-form method and historical simulation.

Suppose that we know the pool's age, t_0 , factor, $F(t_0)$, and a constant turnover rate,¹ $\lambda_{turnover}$. Then, we can assess the turnover factor $F_{turnover}(t_0) = \exp(-\lambda_{turnover}t_0)$ along with the scheduled factor, $F_{scheduled}(t_0)$. Since the entire pool's amortization is driven by refinancing, turnover, and the scheduled payoff, the knowledge of two out of three factors along with the total pool's factor is enough to restore the entire time t_0 composition. It is easy to show that unknown $\psi(t_0)$ satisfies the following, generally transcendental, algebraic equation:

$$x + \alpha x^\beta = 1 \quad (8)$$

where α is a known parameter:

$$\alpha = \frac{1 - \psi_0}{\psi_0^\beta} \left[\frac{F_{turnover}(t_0)F_{scheduled}(t_0)}{F(t_0)} \right]^{1-\beta}$$

and β is the same speed-reducing multiplier that enters the APD model (4).

Of course, no numerical iterations are needed if β is 0, 1, or 0.5. For instance, $\beta = 1$ is a trivial case when the pool is homogeneous and is not subject to burnout, $\psi(t_0) \equiv \psi_0$. Case $\beta = 0$ was considered in Levin (2001, 2002); it leads to

$\psi(t_0) = 1 - (1 - \psi_0) \frac{F_{turnover}(t_0)F_{scheduled}(t_0)}{F(t_0)}$. A simple quadratic equation for $\psi(t_0)$ arises when $\beta = 0.5$, with only one meaningful positive solution. For all other values of β , numerical methods will suffice.

Solving equation (8) is an attractive way to initialize the burnout stage, as it does not require historical simulation of past refinancing incentives. However, it is valid only for very specific forms of the APD model, presented by formulas (4) and (5). Any possible extension of the model (such as discussed below) will make it impossible to recover the burnout stage using the pool's factor and age information only. An alternative method to estimate $\psi(t_0)$ would be a historical simulation of all prepayment components, that is, running the APD model forward from a pool origination until today. A relevant historical interest rate dataset will be required to facilitate this process.

EXTENSIONS AND NUANCES

In this section, we discuss several possibilities of exploring and extending the APD framework. We complete the section by disclosing its expected accuracy and limitations.

Computing Interest Rate Sensitivities Directly Off a Pricing Tree

Let us illustrate how interest rate exposures can be efficiently computed using prices produced on a pricing tree. The idea is to augment the tree with "ghost nodes" as shown in Figure 2; for simplicity and clarity, we illustrate the idea with a recombining binomial tree.²

The tree contains the usual nodes and links (solid lines) that refer to market conditions (interest rates) and their changes. The root node refers to today's market. We assume application of the pricing formula (3) for every transition. We carry this process from maturity backward until we reach the root. This process is carried out separately for Active and Passive

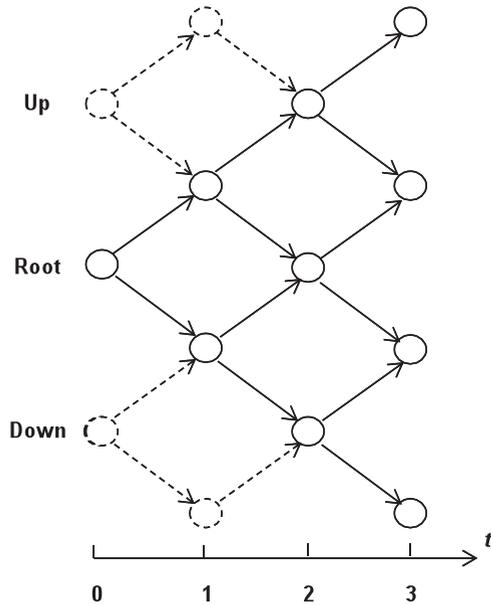


Figure 2 Extended Pricing Tree

components of the mortgage pool; at the root, we combine prices using formula (7).

Let us augment the actual tree with some nodes marked by “Up” and “Down” in Figure 2. Those nodes cannot be reached from the root, but can be perceived theoretically as results of immediate market shocks. We can add as many nodes at time $t = 0$ as we would like. These nodes and the emanating transitions are marked by dashed lines in Figure 2. If we assign transitional probabilities according to the law of our interest rate models and carry out the backward valuation process, we will end up with prices of Active, Passive, and Total prices at time $t = 0$. We can now measure duration and convexity using up and down shifts in the interest rate factor; we can also compile a risk report covering a substantial range on interest rate moves. These calculations will require carrying out the backward induction algorithm on a somewhat expanded tree, but otherwise, no extra computing efforts.

One practical question a user may have is whether interest rate shocks that are reflected in the up, the down, and other nodes are, in fact, parallel moves. In most cases, they are not. Each

node of the valuation tree represents the full set of market conditions altered by a single factor (e.g., the short rate). The entire yield curve becomes known via the relevant law of the term structure model. For example, long rates move less than the short rate if the single-factor model is mean reverting; the rate’s move may be comparable on a relative, not absolute, basis if the model is lognormal, and so on. These examples illustrate nonparallel moves in the yield curve. In these cases, it would be practically advisable to measure the Greeks with respect to the “most important” rate, such as the MBS current coupon rate or the 10-year reference rate.

Among a vast family of known short-rate models, there exists one special model whose internal law is consistent with the notion of parallel shocks. This is the Hull-White model with a zero mean reversion, also known as the Ho-Lee model³ (see, for example, Hull, 2005). When the short rate moves by x basis point, every zero-coupon rate will move by the same amount, regardless of its maturity.

If the Ho-Lee model is not employed and the sensitivity to parallel shocks of interest rates is a must (no approximation accepted), the tree-based valuation will have to be repeated using user-defined parallel moves of the yield curve. Whereas some advantages of the backward induction’s superior speed will be forfeited, the method will still stand as a viable alternative to the Monte Carlo method.

More Components, More Prepay Sources

The APD model given by (4), (5), and (6) is a two-component pool model exposed to two sources of prepayment, refinancing and turnover. Each of these features can be generalized. A mortgage pool can be thought of as a blend of many prepayment patterns (super-active, active, moderately active, and so on). On the other hand, there may exist prepayment sources that contribute to each of the groups, but are distinctly different from refinancing and

turnover. Let us briefly discuss both ways to extend the model.

As we already pointed out, even a two-component model ensures smooth prepayment behavior if each component does so. Within the APD framework, a refinancing model may include the traditional S-like curve, aging, and perhaps some other known empirical mortgage effects that can be attributed to a nonoptimal option exercise. The total prepayment speed is proven to be between RefiSMM and TurnoverSMM, being continuously weighted as controlled by variable $\psi(t)$. Adding more components into the model does not alter this fact nor does it add any smoothness in the prepay model. It is also more difficult to fit a three- or four-component model than the APD model presented here. We believe that even a simple, but dynamic, APD model captures the main prepayment factors, including burnout.

The APD model (4) assumes that the active and passive components share the same turnover rate, and their refinancing speeds relate to one another as 1 to β . We can consider some other prepayment source that is not propagated to the active and passive components identically, or with the 1 to β ratio. For example, we may introduce both default termination and credit cure prepay sources, additive to refinancing and turnover, but likely having a higher effect on the passive part than on the active part.⁴

Of course, additional prepayment sources can be formally included in the refinancing without assuming any more that active and passive refinancing models relate to one another. We will not be able to initialize $\psi(t_0)$ by solving equation (8), and we must use historical simulation for this purpose as discussed above. Principally, we may assume unrelated refinancing models built for the active and passive components, gaining generality with little sacrifice of convenience.

Residual Sources of Path-Dependence

The APD model takes care of the burnout effect, the major source of path-dependence for

fixed-rate mortgages. After the decomposition is done, we need to review residual sources of path-dependence and arrange the numerical valuation procedure to reduce or eliminate potential pricing errors.

Prepayment lag, a lookback option feature, is such a source. Applications to obtain a new mortgage replacing an old one enter the origination pipeline 30 to 90 days before the loan is actually closed and the existing debt is paid off. Even if the prepayment model features a lag, but the backward valuation scheme is unaware of its existence, the pricing results can be somewhat inaccurate. This ignorance of the lag by the backward induction scheme usually causes small errors for pass-through securities. However, mortgage strip derivatives are highly prepayment sensitive, and the lag may change their values in a sizable way.

It is generally known that lookbacks with fairly short lag periods can be accounted for in the course of a backward induction process. Let us assume, for example, that, on a trinomial monthly tree, speed λ_k actually depends on market rates lagging one month. Hence, the MBS value will also depend on both the current market and 1-month lagged market. This is to say that each valuation node of the tree should be “sliced” into three subnodes keeping track of prices matching three possible historical nodes, one month back. Of course, this costs computational time; efficiency may deteriorate quickly for deeper lags and more complex trees.

Approximate alternatives do exist and it is feasible to reduce pricing errors without much trouble. AD&Co employs a progressively sparse recombining pentagonal tree, which does not branch off every month. Branches of the tree are made from two to 12 months long so that the lagged market rates are explicitly known for most monthly steps. The lookback correction can also be adapted for the “fractional” prepayment lag that almost always exists due to the net payment delay between the accrued-month-end and the actual cash flow date. In such a case, λ_k could be interpolated

between the current-month and the previous-month values. Thus, the total lag processing should account for both prepay lookback and payment delay.

Another example of path-dependence not cured by pool decomposition is the coupon reset for adjustable-rate mortgages (ARMs). Both reset caps and nonlinear relationships between prepayments and coupons make it difficult for a backward induction scheme to account for this feature. One possible solution is to extend the state space and create an additional dimension that would keep track of the coupon rate for an ARM (Dorigan et al., 2001). This state-space extension will come at a cost of both computational efficiency and memory consumption. Levin (2002) suggests that the reset provisions found in typical ARMs allow for backward valuation with a practically acceptable accuracy, without any special measures on curing this path-dependence.

Modeling Prepayments Universally: Refinancing Speed as a Function of Price

We finish the entry with a rather interesting, if not unique, application of the APD idea where backward valuation of MBS is not an option, but a necessity. The academic literature contains quite a few works on the rational prepayment exercise models.⁵ Our APD model is not of that sort as it is a “mortgage-like” approach that can accommodate empirical features such as an S-curve or aging. Yet, it can address some shortcomings typically known for purely ad hoc empirical models. As we have already asserted, the APD model can value MBS backward provided that its refinancing and turnover constituents depend only on the current market. A likely implementation of this rule would rely on some experimental relationship between the SMMs and a relevant mortgage index. Although this is the way most mortgage practitioners envision *prepayment modeling*, it is not the only possible approach. In fact, the refinancing be-

havior of homeowners also depends on the type of mortgage in hand. Given coupon and market, the economic incentive to prepay vanishes when maturity, balloon, or ARM reset date approach. Hence, each type of mortgage and each seasoning stage call for its own refinancing model.

An attractive alternative would be linking the refinancing speed of a mortgage (still measured on the grid nodes, separately for the active and passive pieces) directly to its price appreciation, using path-independent specification $\text{RefiSMM}(\text{Price})$ instead of $\text{RefiSMM}(\text{Rate})$. This is the same hint as the one used for valuation of American option bonds except the refinancing model can still be an exogenous S-curve, not the “optimal” or “rational” exercise rule. This model would state the refinancing speed, RefiSMM , as a function of the pool’s price, for example, 15 CPR if collateral is priced at 102, 30 CPR for 105, and so on, asymptotically approaching its “ultimate” speed. Formulas (4), (5) still allow computing the active, passive, and total speeds. In particular, the passive component will still run off at a beta-reduced speed for the same price premium as the active component.

In essence, variable λ in the pricing PDE (2) becomes a function of the unknown P . Such an equation will still be path-independent, presenting no theoretical or computational issues for the backward solution. Moreover, if the refinancing behavior is indeed driven by price appreciation and such a universal relationship can be experimentally established, then the APD modeling approach and its backward implementation becomes a natural, if not the only, way to price an MBS. Any Monte-Carlo-based valuation method simply would not allow assessing future prices and, hence, prepayment speeds.

Arguably, the $\text{RefiSMM}(\text{Price})$ function can be viewed as one universal refinancing rule that can serve many collateral types. Furthermore, such a model can directly account for additional loan-specific transaction costs and cost saving

opportunities. For example, the knowledge of prepayment penalties, average loan sizes, or state-imposed taxes can easily be used to modify the S-curve.

Furthermore, the RefiSMM(Price) formulation can be used for modeling collateral behavior for CMOs as well. Although a typical CMO is path-dependent well beyond its collateral and necessitates Monte Carlo sampling, it is the prepayment modeling stage that can be done via the APD scheme. We will start with valuing collateral first on the grid or a tree, and then compute and store ActiveSMM and PassiveSMM for every node of the tree as a result of the backward inducting process described in this entry. We then run Monte Carlo simulations for the CMO in question and apply precomputed SMMs. As we pointed out, the key compositional variable $\psi(t)$ is known going forward (but not backward), thereby enabling construction of the full prepayment rate, hence, the cash flow, for every node and every path.

This approach's details and an illustration of how the same S-curve can "serve" both fixed-rate and adjustable-rate ARMs are given in Levin (2006). Pricing PDE (2) with $\lambda = \lambda(P)$ has been given mathematical consideration by Goncharov (2003, 2006), who studied the existence and uniqueness of its solution.

KEY POINTS

- The prices of mortgage-backed securities follow a partial differential equation that includes interest rates, coupon rates, and prepayment rates. Even for a simple mortgage pass-through security, this valuation PDE is path-dependent as it depends on the attained stage of burnout (hence, on past refinancing incentives).
- The active-passive decomposition model splits a pool into two path-independent, mutually exclusive borrower groups. APD naturally simulates the burnout effect.
- For mortgage pass-through securities (and their strip derivatives), APD splits valuation

into two quick backward induction steps and produces the entire pricing grid for risk measurement at no additional cost (unlike Monte Carlo simulation).

- Whereas CMOs will still rely on Monte Carlo simulation as being heavily path-dependent beyond the burnout, they will benefit from better prepay modeling.
- The backward induction pricing technique makes future values accessible and new valuation and modeling tasks feasible. For example, one can assume that the refinancing curve is a function of a loan's objective price rather than interest rates. Such an approach can be viewed as a universal model that applies to both fixed and adjustable rate pools.

NOTES

1. We can relax this condition just assuming that the historical turnover rate is known, not necessarily constant.
2. When using finite difference grids for solving the pricing PDE, the ghost nodes are part of the grid.
3. Historical calibration of the Hull-White model to the swaption volatility surface often reveals a small-to-zero level of the mean reversion constant.
4. One reason a borrower is "passive" can be due to credit-related issues.
5. See Longstaff (2003) and Stanton (1995).

REFERENCES

- Davidson, A. (1987). Understanding premium mortgage-backed securities: Observations and analysis, in F. Fabozzi (ed.), *Mortgage-Backed Securities: New Strategies, Applications and Research*. Chicago: Probus Publishing, pp. 191–204.
- Davidson, A., Herskovitz, M., and Van Drunen, L. (1988). The refinancing threshold model: An economic approach to valuing MBS. *Journal of Real Estate Finance and Economics* 1, June: 117–130.

- Dorigan, M., Fabozzi, F. J., and Kalotay, A. (2001). Valuation of floating-rate bonds, in F. Fabozzi (ed.), *Professional Perspectives on Fixed Income Portfolio Management*, Volume 2. Hoboken, NJ: John Wiley & Sons.
- Fabozzi, F., and Fong, G. (1994). *Advanced Fixed Income Portfolio Management*. Homewood, IL: Irwin Professional Publishing.
- Goncharov, Ye. (2003). *Mathematical Theory of Mortgage Modeling*. Ph.D. dissertation, University of Illinois at Chicago.
- Goncharov, Ye. (2006). An intensity-based approach to the valuation of mortgage contracts and computation of the endogenous mortgage rate. *International Journal of Theoretical and Applied Finance* 9, 6: 889–914.
- Hayre, L. (1994). A simple statistical framework for modeling burnout and refinancing behavior. *Journal of Fixed Income* 4, 3: 69–74.
- Hayre, L. (2000). Anatomy of prepayment. *Journal of Fixed Income* 10, 1: 19–49.
- Hull, J. (2005). *Options, Futures, and Other Derivatives*, 6th ed. Upper Saddle River, NJ: Prentice Hall.
- Kalotay, A., and Young, D. (2002). An implied prepayment model for mortgage-backed securities. Presentation at the Bachelier Congress, Crete, Greece.
- Levin, A. (1998). Deriving closed-form solutions for Gaussian pricing models: A systematic time-domain approach. *International Journal of Theoretical and Applied Finance* 1, 3: 348–376.
- Levin, A. (2001). Active-passive decomposition in burnout modeling. *Journal of Fixed Income* 10, 4: 27–40.
- Levin, A. (2002). Mortgage pricing on low-dimensional grids. In F. J. Fabozzi (ed.), *Interest Rate, Term Structure, and Valuation Modeling*. Hoboken, NJ: John Wiley & Sons.
- Levin, A. (2003). Divide and conquer: Exploring new OAS horizons, Part I: Active-passive decomposition. *AD&Co's Quantitative Perspectives*.
- Levin, A. (2006). Universal value-space refinancing model. *AD&Co Annual Conference*.
- Longstaff, F. (2003). Optimal recursive refinancing and the valuation of mortgage-backed securities. *Derivatives 2003: Reports from the Frontiers*, NYU/Stern Conference proceedings (book 2).
- Stanton, R. (1995). Rational prepayments and the valuation of mortgage-backed securities. *Reviews of Financial Studies* 8: 677–708.

Analysis of Nonagency Mortgage-Backed Securities

WILLIAM S. BERLINER

Executive Vice President, Manhattan Advisory Services Inc.

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

ANAND K. BHATTACHARYA, PhD

Professor of Finance Practice, Department of Finance, W. P. Carey School of Business, Arizona State University

Abstract: The transformation of groups of mortgage loans with common attributes into tradable and liquid MBS occurs using one of two mechanisms. Loans that meet the guidelines of the agencies (i.e., Fannie Mae, Freddie Mac, and Ginnie Mae) in terms of credit quality, underwriting standards, and balance are assigned an insurance premium (called a guaranty fee) by the agency in question and securitized as an agency pool. Loans that either do not qualify for agency treatment, or for which agency pooling execution is not efficient, can be securitized in nonagency or “private-label” transactions when such transactions are economically feasible. These types of securities do not have an agency guaranty, and must therefore be issued under the registration entity or “shelf” of the issuer. Although the analysis of private-label mortgage-backed securities utilizes many of the techniques employed to assess agency securities, the analysis must be extended in order to incorporate credit risk and adjust returns for expected principal losses, requiring additional analysis and metrics.

While the evaluation of private-label mortgage-backed securities (MBS) utilizes many of the techniques used in the evaluation of agency MBS (i.e., Ginnie Mae, Fannie Mae, and Freddie Mac MBS), the need to incorporate credit risk and adjust returns for expected principal losses requires additional analysis and metrics. The fact that the credit risk in these securities is not

assumed by the government, either explicitly or implicitly, forces investors to evaluate and judge both the timing of the return of principal as well as the amount of principal, if any, that investors can expect to receive. Moreover, credit analysis has moved up what is called the credit stack. A major change stemming from the subprime mortgage crisis is that investors

The authors acknowledge the contributions in one section of this entry by Paul Jacob.

can no longer assume that senior private-label mortgage-backed securities have virtually the same credit risk as agency MBS. Any bond that does not have agency credit support must be treated as a “credit piece” requiring the analysis of a variety of internal and external factors.

In this entry, we outline the various elements that drive the performance of nonagency MBS, and also examine the interactions of these factors.¹ We then examine a useful framework for understanding the evolution of a population’s credit profile and discuss a variety of techniques used to evaluate the credit risk and expected returns of private-label securities.

FACTORS IMPACTING RETURNS FROM NONAGENCY MBS

The analysis of agency MBS is focused on estimating the timing of principal cash flows since the government backing of these securities eliminates investors’ exposure to principal writedowns. Private-label securities require layers of additional analysis. This is because of the introduction of a series of additional factors that determine the bond’s cash flows and thus their projected returns. These factors can be broadly characterized as:

- The amount of principal expected to be returned.
- The timing of principal returns.
- The allocation of principal within the transaction.

Before proceeding, it will be helpful to review a few concepts. Prepayments on nonagency securities must be classified based on their causation. Unlike agency securities, the return of principal to the securitization (or, more specifically, the investment trust) must be treated differently depending on whether it resulted from a voluntary action by the borrower or is forced by credit-related difficulties. Modeling the impact of voluntary prepayments is relatively straightforward; investors can assume

that 100% of principal being prepaid will be returned on the next payment date. By contrast, projecting the impact of involuntary prepayments requires an estimate of both how much of every principal dollar prepaid will actually be paid to the investor, as well as when principal payments will be received by the trust.

The Amount and Timing of Principal Return

Before proceeding, a brief discussion of terminology will be helpful. For private-label securities *voluntary prepayments* encompass traditional prepayment activity. *Involuntary prepayments* are credit-related prepayments that result from defaults or other events specifically related to credit events (such as short sales of homes), while also accounting for the likelihood that less than the full amount of principal will be returned to the transaction (or, more accurately, the trust holding the deal’s collateral). Voluntary prepayments are typically quoted as VPRs, which stands for *voluntary prepayment rate*. They are calculated similar to a conditional prepayment rate (CPR), in which a monthly percentage of prepaid principal (sometimes denoted by VMM) is annualized. Involuntary prepayment speeds are quoted as *conditional default rates* (CDRs)², which are the annualized rate of default. CDRs are calculated by annualizing the monthly rate of default as a percentage of the current balance, or the MDR. The sum of the monthly VMMs and MDRs equals the total deal single monthly mortgage (SMM) rate for any particular month.

The issue of how much principal is projected to be received as a result of involuntary prepayments is a straightforward function of the assumed default rate and *loss severity*. Loss severities are simply the percentage of the defaulted principal that ultimately will not be returned to the investment trust. The inverse of loss severity is the *recovery percentage*.

The issues associated with the timing of principal return are more complex. Since the CDR is

by definition the involuntary prepayment rate, a higher default rate assumes the faster return of at least some principal to investors. As a result, the faster return of principal to the trust due to higher default rates can offset the effects of principal loss. This effect is a function of the price of the security, the loss severity, and the tranche's position in the transaction's structure (i.e., under what circumstances the security will absorb losses).

In addition, the amount of time between when a default occurs and recovered principal is received by the trust (the *lag*) can have a major influence on investor returns, especially for bonds that are more junior in priority. A longer lag between the time of default and the receipt of recovered principal delays the write-down of the junior bond's principal value. This means that the investor may receive interest payments for a longer period of time, improving the value of securities for which the interest payments comprise the bulk of expected cash flows. In fact, lower-priority subordinates are sometimes referred to as credit IOs, since investors assume that no principal will be returned, and the only cash flows that they expect to receive are coupon payments. Since the outstanding principal is written off more slowly, investors holding the tranche receive a larger and longer stream of interest payments as the lag extends.

There are a variety of factors that influence the lag. Both the amount of seriously delinquent loans at a point in time and the actions of servicers play major roles in the timing of defaults and principal recoveries. The period after 2007, for example, saw a huge increase in the number of seriously delinquent loans outstanding. At the same time, servicers (i.e., the entities that process borrower payments and manage the foreclosure process) were unable to effectively manage the huge surge in problem loans. This resulted in an enormous backup in the foreclosure pipeline, and led to long lags between the time when loans stopped performing and the properties were liquidated.

Legal and political factors also impact lag times. Since real estate transactions are governed by state and local laws, there are differences in the timing of principal returns based on the state in which a loan resides. Some states, which are referenced as judicial states, require that a foreclosure be approved by a judge, which typically slows the foreclosure process. Foreclosures in nonjudicial states can be processed faster, resulting in shorter lags. Also, the foreclosure process itself can become a matter of controversy. In 2010, for example, problems with the legal documentation of foreclosure filings led to the suspension of foreclosure proceedings in some states, as well as calls for a national foreclosure moratorium.

Generally speaking, the amount and timing of cash flows to the trust are impacted by a variety of actions and decisions taken by both borrowers and servicers, and are also influenced by exogenous factors. We discuss how these behaviors can be understood and modeled later in this entry.

Deal-Specific Factors

There are also a series of other subtle and obscure factors that can impact the cash flows and returns of nonagency securities. Some of these factors result from decisions by the servicer, while others vary depending on how an individual transaction's governing documents were written. These factors include (but are not limited to) the following:

- Servicers are required to advance principal and interest on delinquent loans. However, the governing documents of most deals state that the servicer is not required to advance any amount it deems "nonrecoverable" through the foreclosure process. The interpretation of "recoverability" depends on servicers' policies with respect to how long they will advance against seriously delinquent loans, along with the loan-to-value ratios (LTVs) of properties backing these loans. (Since expected recoveries are a function of

the current LTV, servicers often will stop advancing on loans where the current LTV exceeds a certain threshold.)

- The treatment of “modified” loans (i.e., loans for which the terms were altered in order to help borrowers meet their obligations) within individual transactions was rarely outlined in deals issued prior to the mortgage crisis. For example, there have been controversies regarding whether “forborne” (i.e., deferred) principal resulting from loan modifications should be written off immediately (which typically benefits the senior bondholders in a transaction) or deferred until the point where principal losses are realized by the trust, which would result in more interest flowing to the subordinates.
- The allocation of losses due to principal and interest “shortfalls” can become highly complex and deal-specific, particularly once the subordinate bonds in an overcollateralization structure are paid off. For example, some deals (typically those issued before mid-2005) only allow for the balances of senior bonds to be reduced by payments actually made by borrowers. These structures can experience a phenomenon called “negative overcollateralization,” which means that losses for the seniors are “implied.” As a result, losses on the senior tranches are only realized when the collateral pool is entirely paid off and the trust is terminated with some bond balances still outstanding.

One conclusion that can be drawn is that investors in private-label MBS must have the will-

ingness and ability to read and understand the documents governing their holdings. Events and factors that were either not contemplated or were viewed as highly improbable can, under adverse conditions, become important in determining investor returns.

UNDERSTANDING THE EVOLUTION OF CREDIT PERFORMANCE WITHIN A TRANSACTION

As discussed previously, the actions and decisions taken by both borrowers and servicers, along with outside environmental factors, determine both the amount and timing of cash flows received by the trust. This behavior can be conceptualized through the use of transition matrices. Such matrices show the probability of loans moving from one credit status (or “state”) to another in any month. This technique is often used as a foundation for formally modeling voluntary and involuntary speeds. We address it here, however, to help conceptualize the “life cycle” of a transaction’s credit profile. The methodology offers useful techniques for demonstrating how the credit problems of obligors evolve into delinquencies and defaults and flow through a transaction over time. It is also useful in describing and quantifying how changes in the overall credit environment might impact the performance of a loan population.

Table 1 contains a hypothetical example of a roll matrix for a loan population, which can be defined either narrowly (e.g., for a single

Table 1 Hypothetical Transition Matrix

		T_1 (“to”) State									
		Payoff	Current	D30	D60	D90+	Bk	Fcl	REO	Liq	Total
T_0 (“from”) State	Current	0.6%	94.6%	4.6%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	100.0%
	D30	0.2%	20.0%	42.4%	36.9%	0.0%	0.4%	0.0%	0.0%	0.1%	100.0%
	D60	0.1%	2.8%	8.9%	34.1%	52.8%	0.5%	0.7%	0.0%	0.2%	100.0%
	D90+	0.1%	1.9%	0.7%	1.0%	85.7%	0.7%	8.3%	0.2%	1.5%	100.0%
	Bk	0.1%	0.1%	0.3%	0.2%	3.7%	86.8%	8.3%	0.4%	0.1%	100.0%
	Fcl	0.1%	0.7%	0.1%	0.0%	4.2%	1.3%	88.7%	3.4%	1.5%	100.0%
	REO	0.7%	0.0%	0.0%	0.0%	0.2%	0.1%	0.4%	82.3%	16.3%	100.0%

Table 2 Applying the Current Population Profile to the Transition Matrix

A. Current Deal Profile										
		Percent of UPB								
Current		61.1%								
D30		4.6%								
D60		2.1%								
D90+		16.6%								
Bk		2.2%								
Fcl		11.4%								
REO		2.0%								
B. Multiply Current Performance by Transition Matrix										
		T_1 ("to") State								
		Payoff	Current	D30	D60	D90+	Bk	Fcl	REO	Liq
T_0 ("from") State	Current	0.3%	57.8%	2.8%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%
	D30	0.0%	0.9%	2.0%	1.7%	0.0%	0.0%	0.0%	0.0%	0.0%
	D60	0.0%	0.1%	0.2%	0.7%	1.1%	0.0%	0.0%	0.0%	0.0%
	D90+	0.0%	0.3%	0.1%	0.2%	14.2%	0.1%	1.4%	0.0%	0.2%
	Bk	0.0%	0.0%	0.0%	0.0%	0.1%	1.9%	0.2%	0.0%	0.0%
	Fcl	0.0%	0.1%	0.0%	0.0%	0.5%	0.1%	10.1%	0.4%	0.2%
	REO	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.6%	0.3%
	Subtotal	0.4%	59.2%	5.1%	2.6%	15.9%	2.3%	11.7%	2.1%	0.8%
Normalized Total ^a			59.9%	5.1%	2.7%	16.1%	2.3%	11.8%	2.1%	100.0%

^aExcluding payoffs and liquidations.

transaction) or more broadly (to represent a particular product and vintage). The vertical axis of the matrix shows the current (or "from") states of the population, while the horizontal axis shows the future (i.e., "to") states of the loans, typically one month hence. The horizontal axis also allows for two additional states, which would represent termination of the loans either through "payoff" (i.e., prepaid voluntarily) or "liquidation" (involuntarily prepaid). Each row must sum to 100%, as every loan in the population at time zero must transition to some state in the following month.

The matrix itself can be created through a variety of techniques. In some cases, the matrix simply represents historical experience (over either a short- or long-term horizon), while other analysts use loan-level simulations to generate the matrix. Note that not all cells have values greater than zero, as some transitions are impossible; for example, a loan cannot go from current to 60-days delinquent without first residing in the 30-days delinquent bucket.

Once a transition matrix is created, it can be applied to the population's current profile (i.e., at time T_0) as a means of projecting the population's credit performance in a future month. Table 2 illustrates the matrix math involved in generating the population's profile in month T_1 , treating the T_0 profile as a 1×7 matrix shown in Table 2(A) to be multiplied times the 7×9 transition matrix in Table 1.³ Table 2(B) shows the resulting profile one month hence (i.e., at time T_1) after summing each column, along with the percentage of loans that drop out of the population through voluntary or involuntary prepayment. The remaining population profile is then normalized by dividing the percentages of remaining loans in each credit state (i.e., excluding loans that are paid off or liquidated) by the remaining percentage in the pool. (In the exhibit, 98.8% represents the portion of the population that remains active; the 59.2% of loans expected to be current in month T_1 is divided by this percentage to get the 59.9% normalized total.)

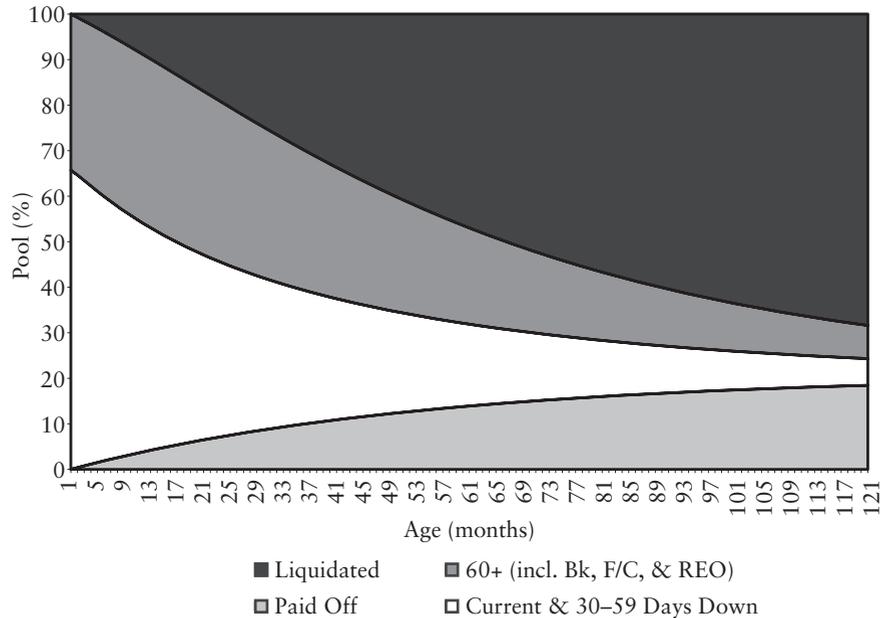


Figure 1 Projected Long-Term Performance Trends for Subject Deal

The process can be performed iteratively in order to show how the population's profile can, given unchanged transition behavior, be expected to evolve over time. (This means that the profile at T_1 can be multiplied by the transition matrix to generate a profile for time T_2 , etc.) Figure 1 shows the projected profile of the population over the next 10 years by iteratively applying the transition matrix in Table 1 to the evolving population. The chart indicates that liquidated loans (i.e., loans that go into default and are removed from the pool through the foreclosure process) will comprise the largest single cohort in around three years if current transition probabilities hold. Moreover, two-thirds of the current population can be expected to be liquidated in 10 years.

The iterative calculation can also be used to generate projections for voluntary and involuntary prepayment speeds. VPR and CDR vectors can then be utilized in yield and cash flow calculators.⁴ Figure 2 shows the vectors generated by the analysis over 120 months.

Interestingly, the vectors are neither constant nor linear; note that the CDR vector increases

fairly steadily for the first few years before leveling off around month 60. This pattern highlights the intrinsic nature of population transitions. Loans flow through the different credit states at varying rates that are a function of transition probabilities captured by the matrix. Therefore, the levels of VPRs and CDRs over time will vary even if transition activity is assumed to remain stable.

However, transition patterns normally do vary over time, reflecting changes in the economic and lending landscape as well as in the actions of servicers. The impact of changing behaviors can be captured by altering the transition matrix at a point in time. For example, a move on the part of servicers to more aggressively clean up the foreclosure pipeline would be captured in a transition framework by increasing the percentages in "late-stage" transitions (i.e., D90+ to FC, FC to REO, and REO to Liq) at a point in the future. Conversely, a full foreclosure moratorium (which was discussed in 2010) would be taken into account by changing all probabilities "from" the D90+, FC and REO buckets to zero for the expected length

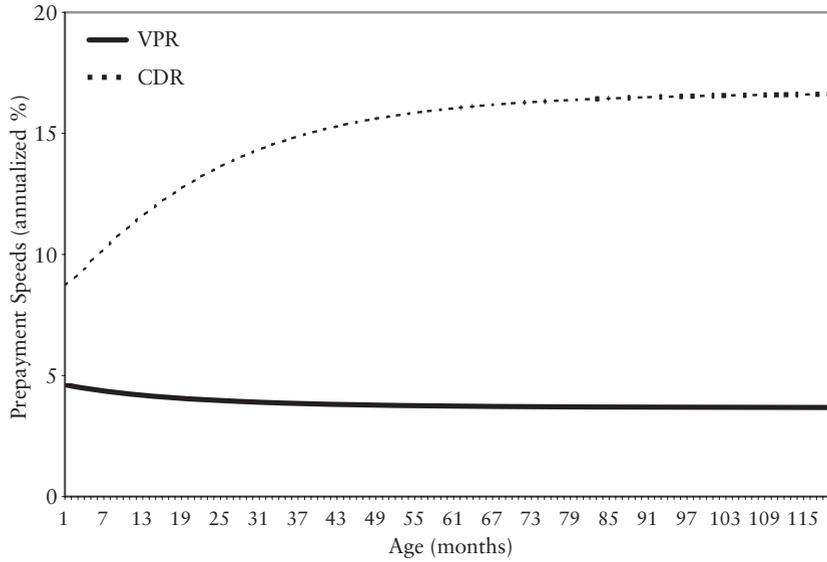


Figure 2 VPRs and CDRs for Subject Deal Using Unchanged Transition Matrix

of the moratorium. Finally, improved borrower performance would be captured by increasing “cures,” (i.e., D30 and D60 to Current) while decreasing the Current to D30 percentage. The updated matrix would be utilized at the point when the changes in behavior were expected to go into effect.

Incorporating such changes in servicer and/or borrower behavior would result in discontinuities in the VPR and CDR vectors. Along with the base vector, Figure 3 shows the projected CDRs for the subject population if the vectors are calculated using transition matrices after month 12 that reflect the “Foreclosure

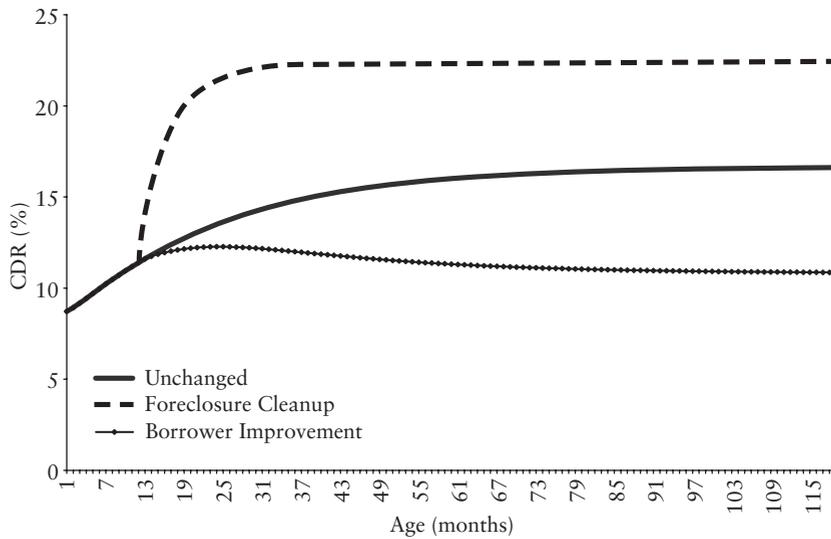


Figure 3 CDR Projections for Different Scenarios after 12 Months

Cleanup” and “Borrower Improvement” scenarios described above.

THE PROCESS OF ESTIMATING PRIVATE-LABEL MBS RETURNS

The analysis and valuation of private-label MBS is complicated by the need to project and account for a number of variables over and above those required to evaluate agency securities. As noted previously, the analysis requires additional metrics necessary to project the principal and interest cash flows paid to the trust, as well as how they will be allocated to the different tranches under a variety of scenarios.

The additional complexity associated with private-label MBS means that the dominant metric used to assess expected returns is *loss-adjusted yield*. This represents the internal rate of return (IRR) for a security’s projected cash flows using the additional factors and variables discussed previously after adjusting for the normal MBS-specific issues such as payment frequency and delay. The increased complexity associated with the product means that some methodologies, such as total return analysis, are infrequently utilized in evaluating credit pieces. For example, total return requires the estimation of a terminal value at the horizon for each scenario being analyzed. The complexity involved in projecting future prices makes them, and thus the analysis, quite subjective.

In the following sections, we illustrate the technique described in this entry using a series of tranches, as well as the collateral, from a representative 2007-vintage hybrid ARM transaction.⁵ The three tranches examined include a super-senior (SS) tranche with 24.2% original credit support; a senior mezzanine (SM) tranche (i.e., a bond originally rated triple-A but junior in priority to the SS) with original credit support of 5.25%; and a subordinate (“sub”) bond or tranche that originally had 3.85% credit enhancement.

Differentiating between Collateral and Tranche Losses

The various factors outlined above have interesting effects and interactions within individual transactions with respect to losses. For one thing, it is important to differentiate between losses on a deal’s collateral pool (i.e., at the trust level) and those impacting individual bonds within a transaction. Private-label MBS have a variety of internal mechanisms that allocate cash flows and principal losses within the structure to tranches having different degrees of seniority. Therefore, losses absorbed by individual bonds are a function of both the losses absorbed by the trust and the amount of credit support available to them.

Figure 4 shows projected losses, as a percentage of original face, for both the overall collateral pool of the deal as well as the three tranches described above. Losses were calculated using different loss severity assumptions while assuming a constant 4% VPR and CDR. (These levels are hypothetical and used for illustrative purposes only.) While the line showing projected losses on the collateral has a linear upward slope, the profile of projected losses for the tranches are quite different. For example, the SS tranche suffers no losses until severities are greater than 50%, while the SM begins to experience losses at severities greater than 40%. The sub tranche, however, has a unique loss profile. It experiences no losses until severities exceed 30%, but at that point losses spike higher; virtually the entire principal value of the bond is written off once the assumed loss severity reaches 45%. The chart highlights a critical conclusion; in addition to being different from the collateral, each bond’s exposure to losses is a function of its place in the transaction’s capital structure.

The Interaction of Credit Inputs

There are also a series of interesting observations that can be made by comparing the yields of the three bonds under a variety of scenarios.

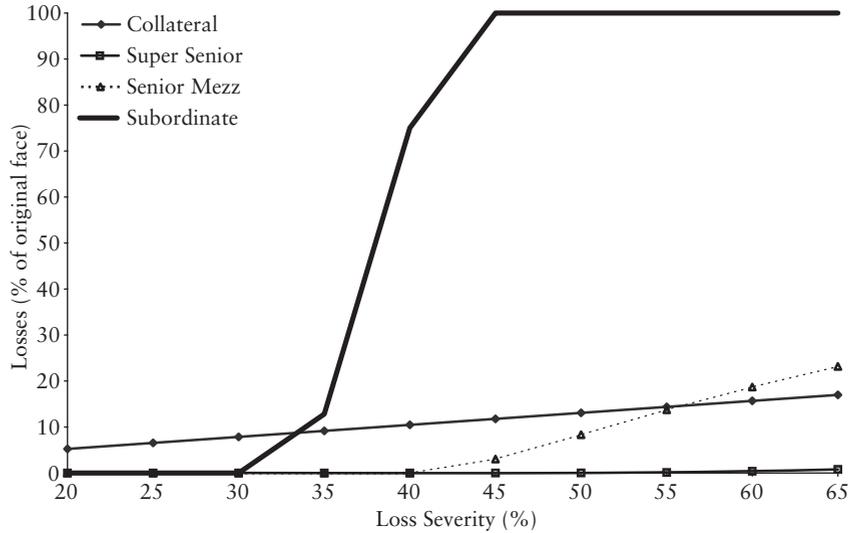


Figure 4 Collateral and Bond Projected Losses at Different Loss Severities (assumption: 4% VPR, 4% CDR, 12-month lag)

For the purposes of the analysis, the bonds were all run at the hypothetical level of a 10% yield to assumptions of 4% VPR, a 6% CDR, a 60% loss severity, and a 12-month lag. (This resulted in prices of 64-12, 48-00, and 6-22 for the three securities.) Using those base-case prices, we ran a few representative scenarios in which different variables were altered, with the goal of ex-

ploring some of the subtleties of the different tranches' returns.

Figure 5 shows yields on the three securities calculated using different CDR projections, assuming a constant 4% VPR along with a 60% loss severity and a 12-month lag. The yield on the SS tranche remains fairly stable (and actually increases slightly until the CDR reaches

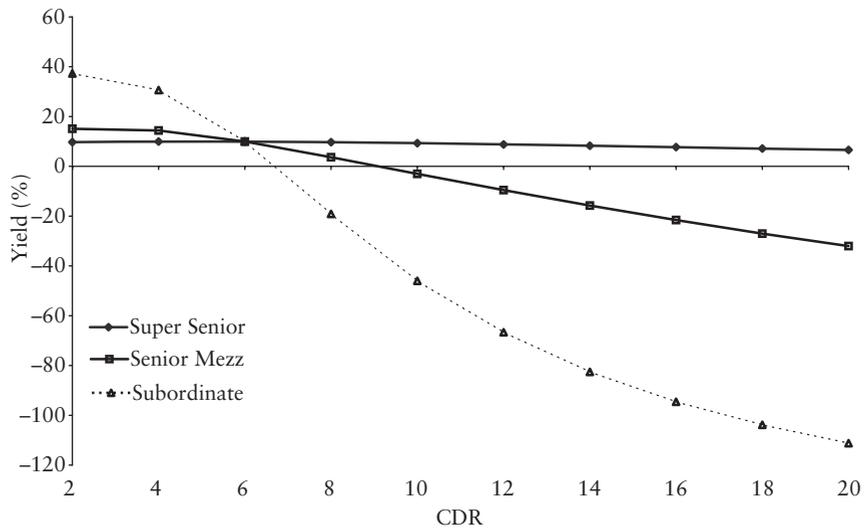


Figure 5 Projected Yields on Different Tranches Using Different CDR (assumption: 4% VPR, 60% severity, 12-month lag)

6%). The reason for this behavior is that faster CDRs effectively increase the overall rate of prepayments to the SS tranche; however, the bond does not absorb losses until the 6% CDR level is breached due to its credit support. Given the tranche's highly discounted dollar price, the faster rate of prepayments increases its yield. By contrast, yields on the more junior tranches decline as CDRs are increased since both bonds realized losses once their limited credit support is exhausted.

The rate of voluntary prepayments also influences returns for some bonds in the transaction. Figure 6 shows projected yields on the three tranches at different assumed VPRs, using a constant 6% CDR (and, as before, 60% severity and a 12-month lag assumption). While their profiles partially reflect the impact of faster prepayments on bonds with deeply discounted prices, voluntary prepayments also have a subtle impact on nonagency MBS. When voluntary prepayments increase, principal is paid back to investors at 100% of face value. This means that there is less principal outstanding that can later go into default, even if the CDR and loss sever-

ity are held constant. As a result, yields for the senior tranches are influenced (and in the SM's case, strongly so) by the expected voluntary prepayment speed.

By contrast, the yield on the sub tranche class is insensitive to changes in the VPR assumption, in part as a result of its place in the deal's structure. (Subordinates generally don't receive voluntary prepayments in an overcollateralization structure unless the deal "steps down," which does not happen under these assumptions.) Its returns, however, are highly sensitive to the combination of assumptions used for CDRs, loss severities, and lags. In particular, the severity assumption plays a key role despite the fact that the bond does not receive principal under most scenarios. As a credit IO, the tranche's outstanding principal value serves as its notional value by dictating how much interest is paid to investors in any single month. Since the severity strongly influences how fast the tranche's face value is written off, it (along with the lag assumption) dictates how long the bond will remain outstanding and thus how much interest investors can expect to receive.

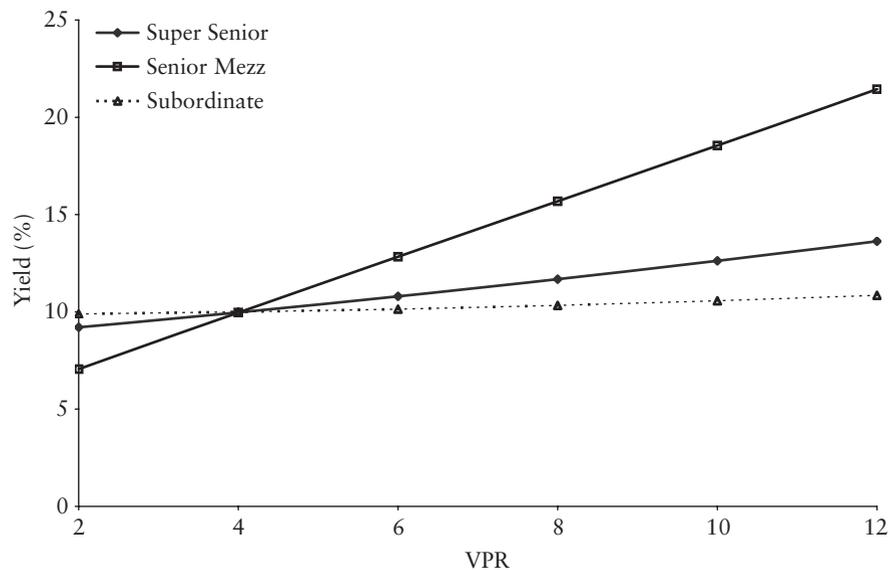


Figure 6 Projected Yields on Different Tranches Using Different VPR (assumption: 6% CDR, 60% severity, 12-month lag)

Evaluating Available Credit Support

Before evaluating projected yields and cash flows for a tranche, a prudent step is to assess the security's remaining credit support relative to the expected level of losses. The objective is to evaluate whether a bond's remaining credit support (i.e., the amount and proportion of bonds junior in priority) is adequate given the losses that the transaction is expected to absorb. The following discussion outlines a simple yet useful methodology for gauging a security's credit support relative to expected losses by using its current performance profile.

The analysis begins by evaluating a transaction's capital structure. Table 3(A) shows the original and current credit structure of a deal.

(While hypothetical, the deal's structure and profile is representative of transactions issued in 2006 and 2007.) The next step, shown in Table 3(B), uses a simple technique to estimate future cumulative losses for the transaction. Utilizing the current performance profile of the transaction, each performance cohort is assigned a probability of ultimate default, along with an assumed loss severity. The example uses a 10% estimate of ultimate default on current loans, a 50% estimate for loans that are D30, while 100% of loans that are seriously delinquent (D90, FC, and REO) are expected to ultimately default. (Note that loans in bankruptcy are not included in this calculation since they are generally captured in other

Table 3 Calculating "Coverage Ratios" for Tranches in a Transaction

A. Original and Current Deal Credit Structure				
Tranche	Orig. Rating	Orig. C/E	Curr. C/E	Curr. Factor
A1 (super senior)	AAA	25.0%	23.2%	0.6950
A2 (senior mezz)	AAA	7.5%	4.5%	0.6950
M1	AA	4.0%	2.4%	1.0000
M2	A	3.5%	1.6%	1.0000
M3	BBB	3.0%	0.8%	1.0000
M4	BB	2.5%	0.0%	1.0000
M5	B	1.5%	0.0%	0.0180
M6	NR	0.0%	n/a	0.0000
B. Current Credit Profile of Transaction				
Performance	UPB	Eventual Default	Assumed Severity	Expected Loss
Current	63.1%	10%	75%	4.73%
D30	4.4%	50%	75%	1.65%
D60	3.1%	90%	75%	2.11%
D90	14.6%	100%	75%	10.96%
FC	12.8%	100%	75%	9.61%
REO	2.0%	100%	75%	1.48%
Total				30.54%
C. Calculating Coverage Ratio				
Tranche		Curr. C/E		Coverage Ratio (curr. CE/expected loss)
A1 (super senior)		23.2%		0.760
A2 (senior mezz)		4.5%		0.147
M1		2.4%		0.079
M2		1.6%		0.052
M3		0.8%		0.026
M4		0.0%		0.000
M5		0.0%		0.000
M6		n/a		n/a

delinquency buckets.) Each delinquency cohort is multiplied by its assigned percentages and the loss severity assumption; the sum of these figures represents the percentage of losses that the deal will ultimately be expected to absorb.

The final step is to divide each tranche's current credit support percentage by the transaction's total expected losses, as shown in Table 3(C). This coverage ratio measures how much credit support is available to each tranche if the expected losses are eventually realized. In the example, the 76% coverage ratio of the A1 tranche suggests that the bond is likely to experience significant future losses, despite its current sizeable cushion. Expected losses will probably also be large enough to eventually cause the other outstanding bonds in the capital structure (i.e., the A2 down to the M5) to be entirely written down.

While this analysis serves as a useful first step in evaluating individual tranches, it is limited by its simplistic approach. The default percentages assigned to each credit bucket are arbitrary, and also cannot account for changes in the credit environment. It also doesn't take into account the issue of time, that is, when losses will accrue and bonds will be written down. This limits its usefulness in evaluating credit IOs and more junior securities. Finally, the analysis doesn't take some forms of credit support, such as excess spread and insurance wraps, into account.

Despite its limitations, however, the methodology serves as a useful first step in evaluating the credit enhancement currently supporting a tranche. In addition, investors evaluating potential purchases of newer securities will find this and related techniques particularly helpful in evaluating both the adequacy of a bond's credit support and whether it is vulnerable to a downgrade by the rating agencies.

Yield and Loss Matrix Analysis

As noted previously, the complexities associated with the product have made loss-adjusted

yield the primary metric for evaluating and comparing credit-related MBS. However, standard yield matrices must be altered in order to account for the numerous additional inputs and outputs necessary to properly evaluate private-label MBS. The additional inputs include separate entries for voluntary and involuntary prepayments, along with the inclusion of expected loss severities, lags, and servicer advances. In some cases, the analysis must also account for the presence of insurance wraps and how long they might remain in place; expectations for how long servicers will continue to advance principal and interest; and whether the deal will pass its triggers (i.e., the tests that dictate cash flow distributions within individual transactions).

In addition, a number of additional outputs are necessary in order to assess a bond's value. In addition to average life, spreads, and durations, investors need to assess expected losses on both the tranche and the deal's collateral at different levels of the inputs. Also useful are the points in time, if applicable, that the bond will experience its first principal loss, along with the amount of liquidations and losses previously realized.

Table 4 contains examples of yield matrices that might be used to evaluate the super-senior and senior mezzanine tranches introduced in an earlier section. Table 4(A) and (B) shows tables for the SS (super senior) tranche and the senior mezzanine (SM) tranche, respectively, priced (as before) at a 10% loss-adjusted yield to a 4% VPR/6% CDR base assumption. The tables in the exhibit show loss-adjusted yields and credit performance data for a range of CDRs, while holding the other variables (i.e., VPR, loss severity, and lag) constant. In addition to yields and average lives, the matrices show the durations, the dates of the first writedown, and the percentages of bond and collateral losses at the different CDR assumptions (which are the same for both tranches in this case).

However, the necessity of holding multiple inputs constant makes this format somewhat

Table 4 Example of Yield Tables for Private-Label MBS Tranches Pricing at 10% Yield at 4% VPR/6% CDR, 60% Severity, and a 12-Month Lag

A. Super-Senior Tranche (Px 64-12)						
VPR	4	4	4	4	4	4
CDR	2	4	6	8	10	12
Yield	9.725	9.978	9.977	9.711	9.310	8.835
WAL	10.39	9.96	8.93	7.95	7.11	6.4
Duration	6.49	6.09	5.71	5.36	5.05	4.77
First-Loss Dt	N/A	09/25/2032	08/25/2023	11/25/2019	12/25/2017	09/25/2016
% Tranche Loss (orig. face)	0.0	0.4	3.7	7.6	11.1	13.9
% Collat. Loss (orig. face)	9.0	15.5	20.4	24.0	26.8	28.9
% Collat. Loss (curr. face)	13.0	22.5	29.6	34.8	38.8	41.9
B. Senior Mezzanine Tranche (Px 48-00)						
VPR	4	4	4	4	4	4
CDR	2	4	6	8	10	12
Yield	15.066	14.406	9.995	3.784	-2.785	-9.137
WAL	10.38	6.69	4.1	3.01	2.46	2.13
Duration	5.18	4.37	3.59	3.08	2.74	2.52
First Loss Dt	N/A	06/25/2019	10/25/2015	05/25/2014	08/25/2013	04/25/2013
% Tranche Loss (orig. face)	0.0	18.8	34.0	41.2	45.1	47.6
% Collat. Loss (orig. face)	9.0	15.5	20.4	24.0	26.8	28.9
% Collat. Loss (curr. face)	13.0	22.5	29.6	34.8	38.8	41.9

awkward and time-consuming. For example, the tables would need to be recalculated multiple times in order to account for other assumptions for VPRs, loss severities, and lags. An alternative and somewhat more flexible scheme displays two variables as the axes, with yields and/or bond losses as the output (creating three-dimensional “surfaces” of yields and losses). Table 5 contains a matrix for the SM tranche showing VPRs on the vertical axis and CDRs on the horizontal, while holding the loss severity and lag assumptions constant. As with other forms of matrices, however, this format

is also limited to showing two variables at any one time. Additional matrices would need to be constructed in order to display different factors, depending on how relevant they were to the analysis.

Model-Generated Analysis

The variables used in the above analysis can be generated in a variety of ways, depending on both investors’ practices and the prevailing circumstances. During periods of relatively stable credit and housing performance, for example,

Table 5 Yield and Bond Loss Matrix for Senior Mezzanine Tranche at Base-Case Price

VPR	2	13.420/0.0%	12.166/25.1%	7.081/39.9%	0.309/46.0%	-6.665/49.2%	-13.249/51.1%
	4	15.067/0.0%	14.402/18.8%	9.992/34.1%	3.781/41.2%	-2.786/45.1%	-9.137/47.6%
	6	16.893/0.0%	16.681/13.9%	12.847/29.1%	7.128/36.9%	0.922/41.4%	-5.209/44.4%
	8	18.890/0.0%	19.010/9.9%	15.681/24.8%	10.396/33.1%	4.515/38.1%	-1.409/41.4%
	10	21.050/0.0%	21.394/6.6%	18.515/21.1%	13.621/29.7%	8.033/35.0%	2.306/38.6%
	12	23.368/0.0%	23.843/4.0%	21.368/17.9%	16.830/26.6%	11.509/32.2%	5.968/36.0%
		2	4	6	8	10	12
CDR							

some investors may choose to simply utilize recent history for inputs such as VPRs, CDRs, and loss severities, while making subjective adjustments based on an examination of the transaction's current collateral profile.

Alternatively, some investors may choose to utilize more sophisticated analysis, which can incorporate both the attributes of a deal's collateral along with exogenous economic and market variables. The models can be further incorporated into integrated systems that generate yield and loss figures while simultaneously analyzing and stratifying the collateral. Partial output from such an integrated system is shown in Table 6. The exhibit shows a yield matrix from Vichara Technology's system for the SS tranche.⁶ The matrix shows a variety of outputs at different multiples of the prepayment and default models, assuming unchanged home prices and interest rates. In addition, separate tables generated by the analysis (not shown) display the current credit structure of the deal, the tranche's cash flows, and analyses of the collateral. (The model also allows for the generation of a "credit OAS," although this metric is not widely utilized by investors at this writing due to its sensitivity to modeling error.)

Additional analysis can be generated for different home price appreciation (HPA) and interest rate assumptions. For example, a conservative set of assumptions might call for a 100 basis point parallel increase in rates accompanied by a 10% immediate decline in home prices. In addition, models for HPA that project different appreciation rates based on geographic and economic factors can also be utilized.

In the case of private-label securities, the normal challenge of assessing a model's "reasonableness" is complicated by the interactive nature of the variables. Unlike agency securities, where "model-equivalent CPRs" can be easily estimated (i.e., the bond's average life is iteratively calculated at various CPRs until it equals the model's calculated WAL), the division of prepayments into voluntary and involuntary categories means that a model-

equivalent CDR cannot be calculated unless the VPR is held constant, and vice versa. This necessitates the need for additional output in order to view and judge the model's VPR and CDR projections.

Interpreting the Outputs

The analysis and valuation of most securities (and virtually all fixed income investments) can be broadly summarized as assessing the "correct" level of expected returns given both market conditions and the bond's risks. This means that a number of factors need to be evaluated, including:

- The security's base-case yields and returns.
- Its returns in best- and worst-case scenarios.
- The likelihood of different scenarios being realized.

The relative complexity of analyzing private-label MBS, particularly compared to evaluating agency-backed securities, results from both the multiplicity of factors influencing returns as well as the many exogenous elements that drive these factors.

For example, a cursory evaluation of the yield matrix for the SM tranche (contained in Table 4(B)) indicates that the bond's projected yields decline rapidly as CDRs are increased. However, the matrix in Table 5 also shows that the tranche's yields remain relatively high if VPRs increase commensurately with CDRs (i.e., in the lower-right quadrant of the matrix). Alternatively, its projected yields are negative when higher CDRs are paired with lower VPRs (in the upper-right quadrant), while yields greater than 20% can be achieved with a combination of fast VPRs and slowing CDRs (the lower-left quadrant). If an investor decides that the combination of VPRs and CDRs in the upper-right quadrant represents a likely scenario, the negative yields projected for such scenarios indicates that the base-case yield assumption is too low to compensate investors for the risks being accepted.

Utilizing just these two variables, the analysis requires investors to assess the returns of

Table 6 Partial Output of Integrated Model for SS Bond

HPI FLAT/+0 IR Shock Scenario				
Percent of Prepay Model	Analytics	Percent of Default Model		
		75%	100%	125%
75%	Yield	10.561	8.416	7.295
	Price	64.38	64.38	64.38
	WAL	5.454	5.295	4.979
	MDuration	3.661	3.878	3.855
	Convexity	0.263	0.303	0.308
	Present Value	157,012,886	157,012,886	157,012,886
	Present Value + Accrued	157,021,422	157,021,422	157,021,422
	Collateral Loss %	37.84%	44.74%	47.97%
	Bond Collateral Loss	37.84%	44.74%	47.97%
	Bond Principal Window	1-333	1-356	1-379
	Bond Principal Writedown	31,919,013	51,373,711	61,802,013
	First Period Writedown	37	31	27
	Bond Principal Writedown	13.09%	21.06%	25.34%
Total Interest Shortfall	—	—	—	
100%	Yield	14.344	10.668	8.359
	Price	64.38	64.38	64.38
	WAL	4.252	4.336	4.268
	MDuration	2.695	3.062	3.250
	Convexity	0.144	0.186	0.216
	Present Value	157,012,886	157,012,886	157,012,886
	Present Value + Accrued	157,021,422	157,021,422	157,021,422
	Collateral Loss %	29.62%	39.49%	45.29%
	Bond Collateral Loss	29.62%	39.49%	45.29%
	Bond Principal Window	1-328	1-355	1-388
	Bond Principal Writedown	17,307,456	40,600,481	56,914,682
	First Period Writedown	40	30	27
	Bond Principal Writedown	7.10%	16.65%	23.33%
Total Interest Shortfall	—	—	—	
125%	Yield	17.718	14.530	10.856
	Price	64.38	64.38	64.38
	WAL	3.495	3.576	3.631
	MDuration	2.183	2.379	2.650
	Convexity	0.094	0.112	0.140
	Present Value	157,012,886	157,012,886	157,012,886
	Present Value + Accrued	157,021,422	157,021,422	157,021,422
	Collateral Loss	24.23%	32.30%	40.38%
	Bond Collateral Loss	24.23%	32.30%	40.38%
	Bond Principal Window	1-328	1-356	1-394
	Bond Principal Writedown	9,427,796	26,037,741	45,698,364
	First Period Writedown	44	32	27
	Bond Principal Writedown	3.87%	10.68%	18.74%
Total Interest Shortfall	—	—	—	

Source: Vichara Technologies. Analysis utilizes deal libraries of Intex Solutions, and models and data provided by CoreLogic.

potential investments in a range of different prepayment and default scenarios with varying degrees of plausibility. Further inquiries should be made regarding expected principal losses on the investment under the assumed

scenarios, taking the availability and adequacy of credit support into account. The sensitivity of the bond's returns to changes in other relevant factors must then be examined. As an example, expectations for real estate prices will directly

impact expected loss severities, which will in turn affect an investor's willingness to buy securities that are more junior in priority. Another example relates to the state of the foreclosure pipeline and its influence on lags. During much of 2009 and 2010, the backup in the foreclosure pipeline meant that buying credit IOs, which benefited from the extended lag, was a profitable strategy as long as servicers continued to advance P&I.

The most difficult aspect of the analysis is generating expectations for factors that are difficult or impossible to quantify. The previous example of the value of credit IOs serves as an example. In addition to the dearth of significant information from servicers (who treat much of the information as having proprietary value), certain factors simply defy quantification. In addition, investors must continuously check their analysis to be certain that they understand what factors are driving their results. This means that the sort of analyses performed earlier in this entry (particularly in the section describing the interaction of factors) is highly useful in developing intuitions for how bonds can be expected to perform under varying conditions.

Note that this entry's discussions were focused on the evaluation of legacy bonds, that is, private-label MBS issued in the period prior to mid-2007. The techniques described in this entry, however, can also be used to evaluate newly issued securities, although some adjustments to the methodologies might need to be made. Investors analyzing the adequacy of credit support using the "coverage ratio" methodology demonstrated in Table 3, for example, would need to replace the use of a transaction's current credit profile with alternative ways of predicting future losses.

Finally, noticeably absent from these discussions were any mention of the rating agencies. Bond ratings cannot and should never substitute for rigorous analysis, as investors that experienced the post-2007 credit meltdown can attest. Ratings are relevant mainly due to constraints and restrictions on the holdings of regulated investors; when bond holdings are

downgraded to below investment grade, many investors are forced to liquidate them, causing their prices to crater. Techniques similar to the coverage ratios outlined previously can be used to monitor the adequacy of bonds' credit support and identify bonds that are vulnerable to being downgraded.

KEY POINTS

- In the analysis of agency MBS, since the government backing of these securities eliminates investors' exposure to principal writedown, the focus is on estimating the timing of principal cash flows.
- Private-label securities require layers of additional analysis because of the introduction of a series of additional factors that determine the bond's cash flows and thus their projected returns. These factors can be broadly characterized as (1) the amount of principal expected to be returned, (2) the timing of principal returns, and (3) the allocation of principal within the transaction.
- The issue of how much principal is projected to be received as a result of prepayments is a straightforward function of the assumed default rate and loss severity. Loss severity is measured as the percentage of the defaulted principal that will ultimately not be returned to the investment trust and of this measure is the recovery percentage.
- The amount and timing of cash flows to the trust are impacted by a variety of actions and decisions taken by both borrower and servicers, and are also influenced by exogenous factors.
- The analysis and valuation of private-label MBS is complicated by the need to project and account for a number of variables over and above those required to evaluate agency securities, requiring additional metrics necessary to project the principal and interest cash flows paid to the trust, as well as how they will be allocated to the different tranches, under a variety of scenarios.

- The additional complexity associated with private-label MBS means that the dominant metric used to assess expected returns is loss-adjusted yield.
 - Before evaluating projected yields and cash flows for a tranche, a prudent step is to assess the security's remaining credit support relative to the expected level of losses. The objective is to evaluate whether a bond's remaining credit support (i.e., the amount and proportion of bonds junior in priority) is adequate given the losses that the transaction is expected to absorb.
 - The relative complexity of analyzing private-label MBS, particularly compared to evaluating agency-backed securities, results from both the multiplicity of factors influencing returns as well as the many exogenous elements that drive these factors. The most difficult aspect of the analysis is generating expectations for factors that are difficult or impossible to quantify.
2. See Chapter 4 in Fabozzi, Bhattacharya, and Berliner (2011).
 3. The example uses the common notation where loans that are 30 to 59 days delinquent are shown as D30, loans that are 90 or more days delinquent are D90+, and so on. "Pay-off" accounts for loans that are voluntarily prepaid; "Liq" are seriously delinquent loans that are liquidated, with T_1 representing the month when recoveries are received by the trust.
 4. The vectors technically are not the equivalent of VPRs and CPRs since they don't account for the effects of amortization on the cash flows.
 5. The analysis utilized CWALT 07-HY8C A1, A2, and M1.
 6. The system utilizes the deal libraries of Intex Solutions; the analysis shown used models and data provided by CoreLogic.

NOTES

1. For an explanation of nonagency MBS, see Fabozzi (2005) and Fabozzi, Bhattacharya, and Berliner (2011).

REFERENCES

- Fabozzi, F. J. (2005). *The Handbook of Mortgage-Backed Securities*. New York: McGraw-Hill.
- Fabozzi, F. J., Bhattacharya, A. K., and Berliner, W. S. (2011). *Mortgage-Backed Securities, 2nd ed.* Hoboken, NJ: John Wiley & Sons.

Measurement of Prepayments for Residential Mortgage-Backed Securities

WILLIAM S. BERLINER

Executive Vice President, Manhattan Advisory Services Inc.

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

ANAND K. BHATTACHARYA, PhD

Professor of Finance Practice, Department of Finance, W. P. Carey School of Business, Arizona State University

Abstract: The valuation of residential mortgage-backed securities begins with a projection of a subject security's cash flow. The monthly cash flow from the underlying pool of mortgage loans includes three components: (1) scheduled principal payments (also referred to as amortization), (2) interest payments, and (3) any prepayments. Prepayments are any payments made by borrowers that are in excess of the scheduled principal payment. Consequently, the cash flow depends on the prepayment behavior of the borrowers in the mortgage pool. In addition to prepayments, the expected credit performance of the underlying loans must be projected to estimate a residential mortgage-backed securities cash flow. The sharp deterioration in mortgage performance that emerged in late 2006 led to the realization that prepayments and defaults often had related effects on the performance of these securities, even though they represent very different phenomena. As a result, new terminology has emerged to clarify the different circumstances that result in the early return of principal to investors. Understanding the terms used in the market to define prepayments and default experience, as well as the methodologies used to generate these metrics, is important for the following reasons: efficient risk-based pricing at the origination level; evaluation of relative value within the residential mortgage-backed securities sector (as well as across the fixed income universe); effective hedging and management of prepayment and credit risk exposure; and ex post performance attribution.

Securities backed by a pool of residential mortgage loans, referred to as *mortgage-backed securities* (MBS) or mortgage-related securities, have complex cash flow characteristics compared to the traditional government, corporate or municipal security. *Residential MBS* are classified as *agency MBS* and *nonagency MBS*. The former include MBS issued by Ginnie Mae (a federally-related government entity) and two government-sponsored enterprises (Fannie Mae and Freddie Mac). Residential MBS not issued by agency MBS are called nonagency or *private label MBS*. In turn nonagency MBS are categorized based on the credit quality of the underlying borrower or lien. There are nonagency MBS backed by prime loans, along with those backed by borrowers with blemished credit histories or an inferior lien on the mortgaged property (e.g., a second mortgage lien). The latter nonagency MBS are generically referred to as subprime MBS.¹

Complicating the cash flows projection of a residential MBS is that borrowers can prepay their loans and will in fact do so for a variety of reasons. Such *prepayments* can occur for a variety of reasons. Virtually all mortgage loans have a “due on sale” clause, which means that the remaining balance of the loan must be paid when the house is sold. Existing mortgages can also be refinanced by the obligor if the prevailing level of mortgage rates declines, or if a more attractive financing vehicle is proposed to them. In addition, homeowners can make partial prepayments on their loan, which serve to reduce the remaining balance and shorten the loan’s remaining term. Prepayments strongly impact the returns and performance of MBS, and investors devote significant resources to studying and modeling them.

For the holder of a mortgage-related security asset, the borrower’s prepayment option creates a unique form of risk. In cases where the obligor refinances the loan in order to capitalize on a drop in market rates, the investor has a high-yielding asset pay off, and it can be replaced only with an asset carrying a lower yield.

Prepayment risk is analogous to “call risk” for corporate and municipal bonds in terms of its impact on returns, and also creates uncertainty with respect to the timing of investors’ cash flows. In addition, changing prepayment “speeds” due to interest rate moves causes variations in the cash flows of mortgages and securities collateralized by mortgage products, strongly influencing their relative performance and making them difficult and expensive to hedge.

Prepayments are phenomena resulting from decisions made by the borrower and/or the lender and occur for the following reasons: (1) sale of the property (due to normal mobility, as well as death and divorce); (2) destruction of the property by fire or other disaster, (3) default on the part of the borrower, and (4) refinancing. Prepayments attributable to the first two reasons are referred to under the broad rubric of “turnover.” Turnover rates tend to be fairly stable over time, but are strongly influenced by the health of the housing market, specifically the levels of real estate appreciation and the volume of existing home sales. Refinancing activity is categorized as either “rate and term” or “cash-out” refinancings. Rate-and-term (or “no cash”) transactions generally depend on a borrower’s ability to obtain a new loan with either a lower rate or a smaller payment. This activity is therefore dependent on the level of interest rates, the shape of the yield curve, and the availability of alternative loan products. These factors also impact cash-out activity, although the primary driver of cash-out refinancings remains home price appreciation; the ability to borrow additional funds against a property is contingent on the property having appreciated in price.

The paradigm in mortgages is thus fairly straightforward. Mortgages with low note rates (that are “out-of-the-money,” to borrow a term from the option market) normally prepay fairly slowly and steadily, while loans carrying higher rates (and are “in-the-money”) are prone to experience spikes in prepayments due to

refinancings when rates decline. In turn, the relationship between a loan's note rate and the prevailing level of mortgage rates dictates whether the borrower has an incentive to refinance.

It is important to understand how changes in prepayment rates impact the performance of mortgages and MBS. Since prepayments increase as bond prices rise and market yields are declining, mortgages shorten in average life and duration when the bond market rallies, constraining their price appreciation. Conversely, rising yields cause prepayments to slow and bond durations to extend, resulting in a greater drop in price than experienced by more traditional (i.e., option-free) fixed income products. As a result, the price performance of mortgages and MBS tends to lag that of comparable fixed maturity instruments (such as Treasury notes) when the prevailing level of yields changes. This phenomenon is generically described as *negative convexity*. The effect of changing prepayment speeds on mortgage durations, based on movements in interest rates, is precisely the opposite of what a bondholder would desire. (Fixed income portfolio managers, for example, extend durations as rates decline, and shorten them when rates rise.) The price performance of mortgages and MBS is, therefore, decidedly nonlinear in nature, and the product will underperform assets that do not exhibit negatively convex behavior as rates fluctuate.

Consequently, it is essential for participants in the residential MBS market to understand the general prepayment and credit performance nomenclature. The market is characterized by the usage of a variety of terms; some terms describe general phenomena, while others are specific to certain types of loan products and assets. In this entry, the basic terms used to characterize residential mortgage-related prepayments and losses are discussed. Our focus is on describing the terminology and outlining the methodologies used in calculating relevant metrics, not on the determinants of prepayment and default behavior.

PREPAYMENT TERMINOLOGY

For fixed-rate fully amortizing assets, such as, home equity loans (HELs), and manufactured housing loans (MHs), the monthly scheduled payment (consisting of scheduled principal and interest) is constant throughout the amortization term. If the borrower pays more than the monthly scheduled payment, the extra payment will be used to pay down the outstanding balance faster than the original amortization schedule, resulting in a prepayment (or, as it is sometimes referenced, an unscheduled principal payment). If the outstanding balance is paid off in full, the prepayment is a complete prepayment; if only a portion of the outstanding balance is prepaid, the prepayment is called either a partial prepayment or curtailment. Prepayments can be the result of natural turnover, refinancings, defaults, partial paydowns, and credit-related events.

The evaluation of prepayments is further complicated by the fact that there is an interplay between *defaults*, which are effectively credit-related prepayments, and prepayments attributable specifically to declining interest rates. In agency MBS (i.e., pools issued by Ginnie Mae, Fannie Mae, and Freddie Mac) there have at times been large numbers of seriously delinquent loans in pools for which Freddie Mac and Fannie Mae continued to pay interest and scheduled principal. In 2010, however, the two government-sponsored enterprises (Fannie Mae and Freddie Mac) changed their policies and began buying loans that were 120 days or more delinquent out of pools. These buyouts initially resulted in a surge in prepayment speeds. Moreover, the new policy meant that pools containing large numbers of lower-quality loans would tend to experience consistently faster prepayment speeds than those pools backed by better-credit loans.

However, for private-label MBS, prepayments resulting from credit events must be treated differently than those attributable to

refinancings. This is because a default means that the investor will probably not receive the entire amount of the defaulted principal, but only the amount recovered after the foreclosure process is completed. Moreover, the timing of payments is also at issue. There is typically a sizeable delay between the time a borrower becomes delinquent on a loan and its ultimate liquidation. This has resulted in the convention where prepayments in private-label securities are separated into *voluntary* and *involuntary prepayments*. Voluntary prepayments occur as a result of a refinancing, the sale of the property, or other events (e.g., the death of the property owner) where the full principal amount is paid immediately to the bondholder. Involuntary prepayments occur as a result of a credit event, for which both the timing and net principal received are uncertain.

Prepayments and defaults can be analyzed on both the loan and pool level. Loan-level prepayment analysis, which requires detailed loan-level information, is more accurate than pool-level prepayment analysis, but is also more computationally intensive. Additionally, this type of analysis allows the inclusion of specific obligor and property characteristics as determinants of prepayments and defaults. Loan-level analysis involves tracking defaults and prepayments on an individual loan basis, projecting each loan's cash flows, and combining these amounts to calculate aggregated metrics. Due to the diversity of the characteristics of the underlying loans in most deals, loan level analysis is generally more accurate and has greater predictive capabilities.

CALCULATING PREPAYMENT SPEEDS

The first critical step in calculating prepayment speed is to define a prepayment. For the purposes of this discussion, a prepayment is defined as the early return of principal to the

investor. By definition, this means that amortization (or scheduled principal payments) must be excluded from the calculation, leaving only unscheduled principal payments to be analyzed.

Conditional Prepayment Rate

The approach most commonly used to generate prepayment speeds is to calculate monthly prepaid principal as a percentage of the security's outstanding balance and then annualize that percentage. Most current approaches to prepayment calculations either quote this annualized periodic speed, known as the *conditional prepayment rate* (CPR) directly or use it as an input to generate other quotation benchmarks.² This methodology is useful in that it allows analysts to both calculate the historical prepayment experience of a security, as well as project prepayment speeds (and thus a security's cash flows) into the future. When used as part of a model to generate projected cash flows, the CPR calculation assumes that some fraction of the unpaid principal balance (or UPB) of the pool is prepaid each month for the remaining term of the mortgage. The advantages of this approach are its simplicity and its flexibility. For example, changes in economic conditions that impact prepayment rates or changes in the historical prepayment pattern of a pool can be analyzed quickly. In addition, the CPR can be used as an input to other models and quotation mechanisms, as noted already.

The CPR is an annual rate. However, because mortgage cash flows are a monthly phenomenon, calculating the CPR requires the generation of a monthly prepayment rate, called the *single monthly mortality rate* (SMM). The SMM is the most fundamental measure of prepayment speeds. SMM measures the monthly prepayment amount as a percentage of the previous month's outstanding balance minus the scheduled principal payment. Mathematically, the SMM is calculated as follows:

$$SMM = \frac{\text{Total payment, including prepayments} - \text{Scheduled interest payment} - \text{Scheduled principal payment}}{[\text{Unpaid principal balance} - \text{Scheduled principal payment}]}$$

For example, if the pool balance at month zero is \$10,000,000, assuming an interest rate of 12%, the scheduled principal and interest payments are \$2,861.26 and \$100,000 in month one, respectively. If the actual payment received by investors in month one is \$202,891.25, the SMM rate is 1%, calculated as

$$SMM = \frac{(\$202,891.25 - \$100,000 - \$2,861.26)}{(\$10,000,000 - \$2,861.26)} = 1\%$$

Therefore, if a mortgage loan prepaid at 1% SMM in a particular month, this means that 1% of that month’s scheduled balance (last month’s outstanding balance minus the scheduled principal payment) has been prepaid.

Given the SMM, a CPR can be computed using the following formula:

$$CPR = 1 - (1 - SMM)^{12}$$

For example, if the SMM is 1%, then the CPR is

$$CPR = 1 - (0.99)^{12} = 11.36\%$$

Conversely, CPRs can be converted into SMMs (and thus be used to generate monthly cash flows) through the following formula:

$$SMM = 1 - (1 - CPR)^{1/12}$$

For example, suppose that the CPR used to estimate prepayments is 6%. The corresponding SMM is

$$SMM = 1 - (1 - 0.06)^{1/12} = 1 - 0.94^{0.08333} = 0.5143\%$$

PSA Prepayment Benchmark

The *Public Securities Association (PSA) prepayment benchmark* is expressed as a monthly series of annual prepayment rates.³ The basic PSA model assumes that prepayment rates are low for newly originated mortgages and then increase linearly as the mortgages age or season.

The PSA standard benchmark assumes the following prepayment rates for 30-year mortgages:

1. A CPR of 0.2% for the first month, increased by 0.2% per year per month for the next 29 months when it reaches 6% per year.
2. A 6% CPR for the remaining years.

This benchmark, referred to as “100% PSA” or simply “100 PSA,” is graphically depicted in the middle graph in Figure 1. Mathematically, 100 PSA can be expressed as follows:

$$\begin{aligned} \text{If } t \leq 30 \text{ then } CPR &= 6\% \times (t/30) \\ \text{If } t > 30 \text{ then } CPR &= 6\% \end{aligned}$$

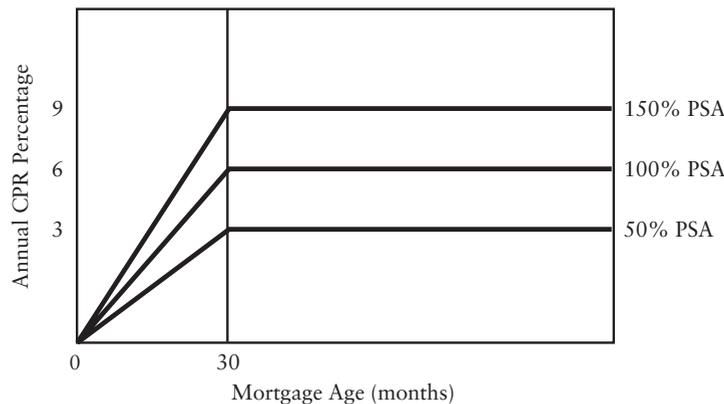


Figure 1 Graphical Depiction of 50 PSA, 100 PSA, and 300 PSA

where t is the number of months since the mortgage was originated. Since the CPR prior to month 30 rises at a constant rate, this period is sometimes referred to as the “ramp,” and loans are considered to be “on the ramp” when they are less than 30 months old.

Slower or faster speeds are then referred to as some percentage of PSA. For example, 50 PSA means one-half the CPR of the PSA benchmark prepayment rate; 150 PSA means 1.5 times the CPR of the PSA benchmark prepayment rate; 300 PSA means three times the CPR of the benchmark prepayment rate. This is illustrated graphically in Figure 1 for 50 PSA, 100 PSA, and 150 PSA. A prepayment rate of 0 PSA means that no prepayments are assumed.

It is important to note that mortgage pools will typically be comprised of loans having different origination months and, therefore, different ages. In practice, the weighted average loan age (WALA) of a pool or security is used as a proxy for its age. However, a large dispersion of loan ages within a pool will distort the PSA calculation.

It is helpful to outline the CPRs and SMMs assumed at different PSA assumptions for different loan ages. The SMMs for month 5, month 20, and months 31 through 360 assuming 100 PSA are calculated as follows:

For month 5:

$$\begin{aligned} \text{CPR} &= 6\% (5/30) = 1\% = 0.01 \\ \text{SMM} &= 1 - (1 - 0.01)^{1/12} = 1 - (0.99)^{0.083333} \\ &= 0.000837 \end{aligned}$$

For month 20:

$$\begin{aligned} \text{CPR} &= 6\% (20/30) = 4\% = 0.04 \\ \text{SMM} &= 1 - (1 - 0.04)^{1/12} = 1 - (0.96)^{0.083333} \\ &= 0.003396 \end{aligned}$$

For months 31–360:

$$\begin{aligned} \text{CPR} &= 6\% \\ \text{SMM} &= 1 - (1 - 0.06)^{1/12} = 1 - (0.94)^{0.083333} \\ &= 0.005143 \end{aligned}$$

The SMMs for month 5, month 20, and months 31 through 360 assuming 165 PSA are computed as follows:

For month 5:

$$\begin{aligned} \text{CPR} &= 6\% (5/30) = 1\% = 0.01 \\ 165 \text{ PSA} &= 1.65 (0.01) = 0.0165 \\ \text{SMM} &= 1 - (1 - 0.0165)^{1/12} \\ &= 1 - (0.9835)^{0.083333} = 0.001386 \end{aligned}$$

For month 20:

$$\begin{aligned} \text{CPR} &= 6\% (20/30) = 4\% = 0.04 \\ 165 \text{ PSA} &= 1.65 (0.04) = 0.066 \\ \text{SMM} &= 1 - (1 - 0.066)^{1/12} = 1 - (0.934)^{0.083333} \\ &= 0.005674 \end{aligned}$$

For months 31 through 360:

$$\begin{aligned} \text{CPR} &= 6\% \\ 165 \text{ PSA} &= 1.65 (0.06) = 0.099 \\ \text{SMM} &= 1 - (1 - 0.099)^{1/12} = 1 - (0.901)^{0.083333} \\ &= 0.007828 \end{aligned}$$

Notice that the SMM assuming 165 PSA is not 1.65 times the SMM at 100 PSA. Rather, the CPR for the pool’s age at 100 PSA is multiplied by 1.65 to generate the CPR representing 165 PSA at that age.

Illustration of Monthly Cash Flow Construction

We now show how to construct a monthly cash flow for a hypothetical agency pass-through given a PSA assumption. For the purpose of this illustration, the underlying mortgages for this hypothetical pass-through are assumed to be fixed rate fully amortizing mortgages with a weighted average coupon (WAC) rate of 6.0%. It will be assumed that the mortgage pass-through rate is 5.5% with a weighted average maturity (WAM) of 358 months.

Table 1 shows the cash flow for selected months assuming 100 PSA. The cash flow is broken down into three components: (1) interest (based on the pass-through rate), (2) the

Table 1 Monthly Cash Flow for a \$400 Million Mortgage Pass-Through with a 5.5% Pass-Through Rate, a WAC of 6.0%, and a WAM of 358 Months, Assuming 100% PSA

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Month	Outstanding Balance	SMM	Mortgage Payment	Net Interest	Scheduled Principal	Prepayments	Total Principal	Cash Flow
1	400,000,000	0.00050	2,402,998	1,833,333	402,998	200,350	603,349	2,436,682
2	399,396,651	0.00067	2,401,794	1,830,568	404,810	266,975	671,785	2,502,353
3	398,724,866	0.00084	2,400,187	1,827,489	406,562	333,463	740,025	2,567,514
4	397,984,841	0.00101	2,398,177	1,824,097	408,253	399,780	808,033	2,632,130
5	397,176,808	0.00117	2,395,766	1,820,394	409,882	465,892	875,773	2,696,167
6	396,301,034	0.00134	2,392,953	1,816,380	411,447	531,764	943,211	2,759,591
7	395,357,823	0.00151	2,389,738	1,812,057	412,949	597,362	1,010,311	2,822,368
8	394,347,512	0.00168	2,386,124	1,807,426	414,386	662,652	1,077,038	2,884,464
9	393,270,474	0.00185	2,382,110	1,802,490	415,758	727,600	1,143,357	2,945,847
10	392,127,117	0.00202	2,377,698	1,797,249	417,063	792,172	1,209,235	3,006,484
11	390,917,882	0.00219	2,372,890	1,791,707	418,300	856,336	1,274,636	3,066,343
12	389,643,247	0.00236	2,367,686	1,785,865	419,470	920,057	1,339,527	3,125,391
13	388,303,720	0.00253	2,362,089	1,779,725	420,571	983,303	1,403,873	3,183,599
14	386,899,847	0.00271	2,356,101	1,773,291	421,602	1,046,041	1,467,643	3,240,934
15	385,432,204	0.00288	2,349,724	1,766,564	422,563	1,108,239	1,530,802	3,297,366
16	383,901,402	0.00305	2,342,961	1,759,548	423,454	1,169,864	1,593,318	3,352,866
17	382,308,084	0.00322	2,335,813	1,752,245	424,273	1,230,887	1,655,159	3,407,405
18	380,652,925	0.00340	2,328,284	1,744,659	425,020	1,291,274	1,716,294	3,460,953
19	378,936,632	0.00357	2,320,377	1,736,793	425,694	1,350,996	1,776,690	3,513,483
20	377,159,941	0.00374	2,312,095	1,728,650	426,296	1,410,023	1,836,319	3,564,968
21	375,323,622	0.00392	2,303,442	1,720,233	426,824	1,468,325	1,895,148	3,615,382
22	373,428,474	0.00409	2,294,420	1,711,547	427,278	1,525,872	1,953,150	3,664,697
23	371,475,324	0.00427	2,285,034	1,702,595	427,657	1,582,637	2,010,294	3,712,889
24	369,465,030	0.00444	2,275,288	1,693,381	427,962	1,638,590	2,066,553	3,759,934
25	367,398,478	0.00462	2,265,185	1,683,910	428,192	1,693,706	2,121,898	3,805,808
26	365,276,580	0.00479	2,254,730	1,674,184	428,347	1,747,956	2,176,303	3,850,488
27	363,100,276	0.00497	2,243,928	1,664,210	428,427	1,801,315	2,229,742	3,893,952
28	360,870,534	0.00514	2,232,783	1,653,990	428,430	1,853,758	2,282,189	3,936,178
29	358,588,346	0.00514	2,221,300	1,643,530	428,358	1,842,021	2,270,379	3,913,909
30	356,317,967	0.00514	2,209,875	1,633,124	428,286	1,830,345	2,258,631	3,891,755
100	223,414,587	0.00514	1,540,329	1,023,984	423,256	1,146,847	1,570,104	2,594,087
101	221,844,483	0.00514	1,532,407	1,016,787	423,185	1,138,773	1,561,958	2,578,745
102	220,282,525	0.00514	1,524,526	1,009,628	423,114	1,130,740	1,553,853	2,563,482
103	218,728,672	0.00514	1,516,686	1,002,506	423,042	1,122,749	1,545,791	2,548,297
104	217,182,881	0.00514	1,508,885	995,422	422,971	1,114,799	1,537,770	2,533,191
105	215,645,111	0.00514	1,501,125	988,373	422,900	1,106,891	1,529,790	2,518,164
200	100,719,066	0.00514	919,770	461,629	416,174	515,859	932,033	1,393,662
201	99,787,032	0.00514	915,039	457,357	416,104	511,066	927,170	1,384,527
202	98,859,862	0.00514	910,333	453,108	416,034	506,298	922,332	1,375,439
203	97,937,531	0.00514	905,651	448,880	415,964	501,555	917,518	1,366,399
204	97,020,012	0.00514	900,994	444,675	415,893	496,836	912,730	1,357,405
205	96,107,283	0.00514	896,360	440,492	415,823	492,142	907,966	1,348,457
300	28,001,417	0.00514	549,218	128,340	409,211	141,907	551,118	679,457
301	27,450,299	0.00514	546,393	125,814	409,142	139,073	548,215	674,028
302	26,902,085	0.00514	543,583	123,301	409,073	136,254	545,326	668,628
303	26,356,758	0.00514	540,787	120,802	409,003	133,450	542,453	663,255
304	25,814,305	0.00514	538,006	118,316	408,934	130,660	539,595	657,910
305	25,274,710	0.00514	535,239	115,842	408,865	127,885	536,751	652,593
350	3,725,850	0.00514	424,402	17,077	405,773	17,075	422,848	439,925
351	3,303,002	0.00514	422,219	15,139	405,704	14,901	420,605	435,744
352	2,882,397	0.00514	420,048	13,211	405,636	12,738	418,374	431,585
353	2,464,023	0.00514	417,887	11,293	405,567	10,587	416,154	427,447
354	2,047,869	0.00514	415,738	9,386	405,499	8,447	413,946	423,332
355	1,633,924	0.00514	413,600	7,489	405,430	6,318	411,749	419,237
356	1,222,175	0.00514	411,473	5,602	405,362	4,201	409,563	415,164
357	812,613	0.00514	409,357	3,724	405,294	2,095	407,388	411,113
358	405,224	0.00514	407,251	1,857	405,225	0	405,225	407,082

^a Since the WAM is 358 months, the underlying mortgage pool is seasoned an average of two months. Therefore, the CPR for month 28 is 6%.

regularly scheduled principal payment, and (3) prepayments based on 100 PSA. Let's walk through Table 1 column by column:

Column 1. This is the month.

Column 2. This column gives the outstanding mortgage balance at the beginning of the month. It is equal to the outstanding balance at the beginning of the previous month reduced by the total principal payment in the previous month.

Column 3. This column shows the SMM for 100 PSA. Two things should be noted in this column. First, for month 1, the SMM is for a pass-through that has been seasoned three months because the WAM is 357 months. This results in a CPR of 0.8%. Second, from month 27 on, the SMM is 0.00514, which corresponds to a CPR of 6%.

Column 4. The aggregate monthly mortgage payments using a 6% note rate are shown in this column. Notice that the total monthly mortgage payment declines over time, as prepayments reduce the mortgage balance outstanding. (In the absence of prepayments, this figure would remain constant.) In essence, the payment is calculated each month as a function of the WAC, the remaining balance at the end of the prior month, and the remaining term (i.e., the original WAM minus the number of months since issuance). For example, the payment in month 10 of \$2,376,474 can be generated on a calculator by inputting \$391,508,422 as the balance or present value, 0.5% (6.0% divided by 12) as the rate, and 348 months as the remaining term.⁴

Column 5. The monthly interest paid to the pass-through investor is found in this column. This value is determined by multiplying the outstanding mortgage balance at the beginning of the month by the pass-through rate of 5.5% and dividing by 12.

Column 6. This column shows the scheduled principal repayment, or amortization. This is the difference between the total monthly

mortgage payment [the amount shown in column (4)] and the gross coupon interest for the month. The gross coupon interest is 6.0% multiplied by the outstanding mortgage balance at the beginning of the month, then divided by 12.

Column 7. The dollar value of prepayments for the month is reported in this column. This amount is calculated by using the following equation:

$$\begin{aligned} \text{Prepayments}_t &= \text{SMM}(\text{Beginning principal balance}_t \\ &\quad - \text{Scheduled principal balance}_t) \end{aligned}$$

So, for example, in month 100, the beginning mortgage balance is \$223,414,587, the scheduled principal payment is \$423,356, and the SMM at 100 PSA is 0.00514301 (only 0.00514 is shown in the table to save space), so the prepayment is

$$\begin{aligned} &0.00514301 \times (\$223,414,587 - \$423,356) \\ &= \$1,146,847 \end{aligned}$$

Column 8. The total principal payment, which is the sum of columns (6) and (7), is shown in this column.

Column 9. The projected monthly cash flow for this pass-through is shown in this last column. The monthly cash flow is the sum of the interest paid to the pass-through investor [column (5)] and the total principal payments for the month [column (8)].

Prospectus Prepayment Curve

A more recent addition to MBS prepayment terminology is the prospectus prepayment curve (PPC). While the logic underlying the PSA convention (i.e., that loans prepay faster as they age, all other factors constant) remains in force, a PPC curve allowed its creator (typically the underwriter of a private-label deal) to specify the prepayment ramp that was used to structure the deal. Evidence suggested that loans have seasoned faster than the 30 month period

implied by the PSA curve, especially for some products (such as alt-A loans) that were believed to season faster than normal. Rather than use a percentage of a publicly utilized ramp, PPC curves (which are quoted in a transaction's prospectus supplement) were used for many nonagency transactions between 2004 and 2007.

Typically, 100% PPC is the base-case prepayment assumption used to create a particular deal. PPC curves (or ramps) are generally specified as a beginning and terminal CPR, along with the associated time period. A typical ramp might be specified as "8–20% CPR over 12 months." This translates to an assumption of 8% CPR in the first month, increasing 1.09% per month for the next 11 months, and terminating at 20% CPR in month 12. However, there is no industry standardization for the usage of this terminology, as the specification is issue-dependent. As a result, investors must confirm how "100% PPC" is defined for each particular issue before performing further analysis.

The language utilized in a deal's prospectus supplement is illuminating. For example, the document for the CWALT 2005-J9 deal has language as follows:

Prepayments of mortgage loans commonly are measured relative to a prepayment standard or model. The model used in this prospectus supplement assumes a constant prepayment rate (i.e., CPR) or an assumed rate of prepayment each month of the then-outstanding principal balance of a pool of new mortgage loans. A 100% prepayment assumption for loan group 1 (the "prepayment assumption") assumes a CPR of 8.0% per annum of the then outstanding principal balance of the applicable mortgage loans in the first month of the life of the mortgage loans and an additional approximately 1.0909090909% (precisely 12%/11) per annum in the second through 11th months. Beginning in the 12th month and in each month thereafter during the life of the mortgage loans, a 100% prepayment assumption assumes a CPR of 20.0% per annum each month.

Note that the prospectus supplement does not directly refer to a "PPC," but rather defines

the prepayment ramp as "a 100% prepayment assumption."

Prepayment Conventions for Securities Backed by Home Equity and Manufactured Housing Loans

While the expression of prepayments in the MBS market is fairly standardized and comprises a combination of PSA curves and CPR calculations as previously described, a variety of descriptions are used to express the pay-down behavior of securities backed by home equity and manufactured housing loans. While issuance of securities backed by these loans fell out of favor in the mid-2000s, a brief discussion of these conventions will nonetheless be helpful in understanding how prepayment conventions have been adjusted in order to represent an asset's unique behavior. Despite the diversity in terminology, most of the concepts used to indicate prepayments for these two sectors of the mortgage market use the CPR concept as the numeraire while incorporating the PSA ramping methodology.

Home Equity Prepayment Speeds

In the early stages of the development of the securitized market for home equity loans, the majority of the loans were fixed rate, closed-end loans. Over the years, the balance has slowly shifted in favor of adjustable rate loans, particularly subprime ARMs. The earliest definition of prepayment speeds in the home equity market was the *home equity prepayment (HEP) curve*.⁵ The primary motivation for using a different prepayment methodology for home equity loans was to capture the faster seasoning ramp observed for the asset class. Typically, home equity loans season faster than traditional single-family loans, making the PSA ramp an inappropriate description of the behavior of prepayments.

The HEP curve reflects the observed behavior in historic HEL data—it has a ramp of

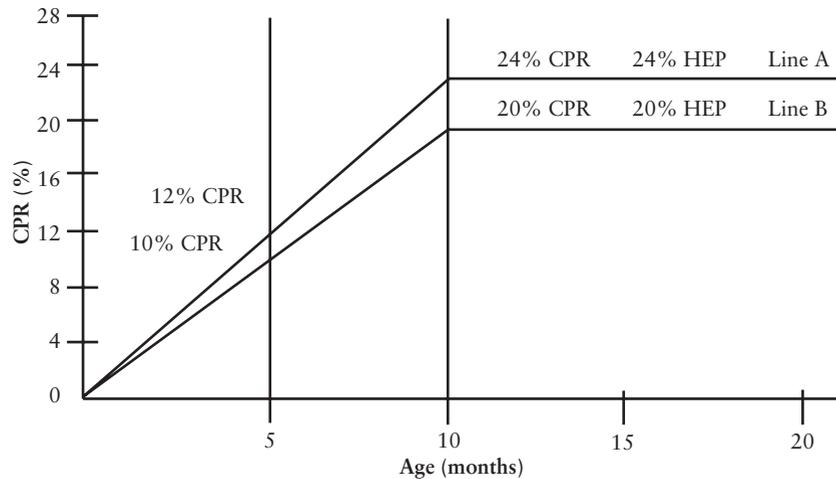


Figure 2 HEP Curves

10 months and a variable long-term CPR to reflect individual issuer speeds. A faster long-term speed means faster CPRs on the ramp because the ramp is fixed at 10 months regardless of the long-term speed. For example, a 20% HEP projection would mean a 10-month ramp increasing to 20% in the 10th month from 2% in the first month and a constant 20% thereafter. Figure 2 shows several HEP curves at 20% HEP and 24% HEP, where month 1 speeds of 2.4% CPR increase over 10 months to 24% CPR.

In addition to utilizing the HEP curve, a PPC ramp is also commonly used to define the base-case prepayment assumption for the product. As with other mortgage products, the specification of the ramp will be dependent on the attributes of the underlying loan collateral, with respect to both the beginning and terminal speeds as well as the duration of the ramp. Occasionally, deals are also priced to a constant CPR assumption, ignoring the impact of seasoning in generating the deal's cash flows.

Manufactured Housing Prepayment Curve

The *manufactured housing prepayment* (MHP) curve is a measure of prepayment behavior for manufactured housing, based on the Green Tree

Financial manufactured housing prepayment experience. MHP is similar to the PSA curve, except that the seasoning ramp is slightly different to account for the specific behavior of manufactured loans: 100% MHP is equivalent to 3.6% CPR at month zero and increases 0.1% CPR every month until month 24, when it plateaus at 6% CPR. Figure 3 shows the prepayment speeds at 50% MHP, 100% MHP, and 200% MHP.

DELINQUENCY, DEFAULT, AND LOSS TERMINOLOGY

The measurement of potential and actual cash flow impairment resulting from borrower credit problems is critically important to the analysis of private label or nonagency MBS. Historically, the importance of these measures stemmed from their role in allowing investors in subordinate MBS tranches to assess relative value and risk. However, the mortgage crisis that began in 2007 demonstrated to investors that all nonagency securities have exposure to defaults and losses; put differently, it is impossible to invest in nonagency MBS without taking on a material degree of credit risk. This means that any divergence in realized default and loss experience from investors' initial expectations can result in writedowns and losses on the investment.

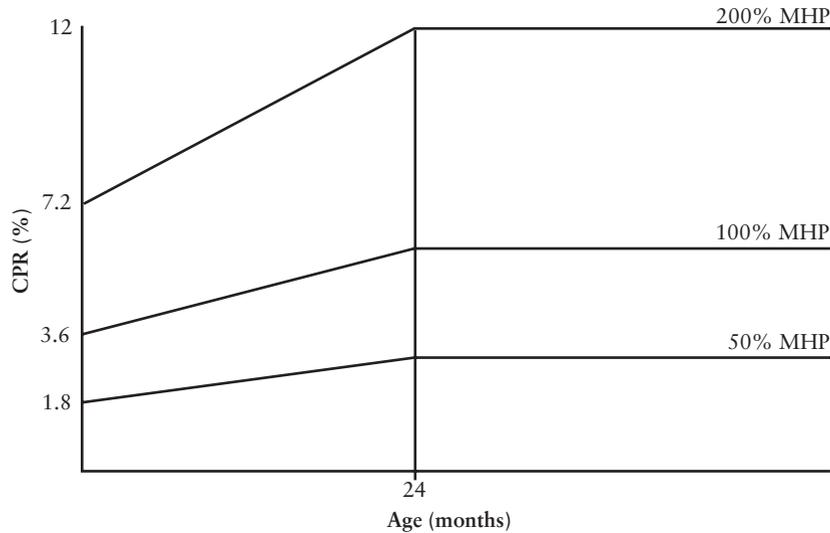


Figure 3 MHP Curves

Despite the importance of delinquencies, losses, and defaults in the mortgage-related markets, the terminology is not standardized. For instance, static pool losses may be reported on a monthly or annualized basis as a percentage of either current or original balance, with the metric based upon current balance being the preferred method to ensure consistency with prepayment reporting.

Before we discuss the measurement of defaults and losses, it is instructive to briefly review the various outcomes of a loan when the obligor ceases making scheduled payments. A loan becomes delinquent when the obligor fails to make the contractual payment on the stated date. If the underlying property has appreciated from the initial purchase price, the homeowner can often sell the home and use the proceeds to settle the mortgage debt. (This generally is categorized as a voluntary prepayment and is considered part of housing turnover.) If the homeowner cannot sell the property at a high enough price and remains delinquent, the loan is declared to be in default once all collection (and modification) efforts have failed. At that point, the issuer (or the servicer) has several options. There may either be a short sale, where the borrower sells the property in a negotiated

transaction subject to approval by the servicer; alternatively, the property may go into the foreclosure or repossession process and be eventually sold by the servicer. Therefore, the process chain is delinquency to default to foreclosure (or repossession) to liquidation, at which time the severity of loss can be assessed.

Delinquency Measures

As mentioned, when a borrower fails to make one or more timely payments, the loan is said to be delinquent. Delinquency measures are designed to gauge whether borrowers are current on their loan payment as well as to stratify unpaid loans according to the seriousness of the delinquency. The calculation method used is determined by the servicer. When the underlying pool of assets is comprised of mortgage loans, the two commonly used methods for classifying delinquencies are those recommended by the now-defunct Office of Thrift Supervision (OTS) and the Mortgage Bankers Association (MBA).

The OTS method uses the following loan delinquency classifications:

- Payment due date to 30 days late: Current
- 30–60 days late: 30 days delinquent

- 60–90 days late: 60 days delinquent
- More than 90 days late: 90+ days delinquent

The MBA method is a somewhat more stringent classification method, classifying a loan as 30 days delinquent once payments are not received after the due date. Thus, a loan classified as “current” under the OTS method would be listed as “30 days delinquent” under the MBA method. The two methods can report significantly different delinquencies.⁶

Default Measures

The conditions that result in classification of some loans as delinquent (such as the loss of a job or illness) may change, resulting in the resumption of timely principal and interest payments. However, some portion of the loans classified as delinquent typically end up in default. By definition, default is the point where the borrower loses title to the property in question.

Two broadly used measures for quantifying default are the cumulative default rate and the conditional default rate. The cumulative default rate (denoted as the CDX) is the proportion of the total face value of loans in a pool that have gone into default as a percentage of the total face value of the security.

The *conditional default rate* (CDR) is the annualized value of the unpaid principal balance of newly defaulted loans over the course of a month as a percentage of the unpaid balance of the pool (before scheduled principal payment) at the beginning of the month. It is computed by first calculating the *monthly default rate* (MDR) as shown below:

$$\text{MDR for month } t = \frac{\text{Default loan balance in month } t}{\text{Beginning balance for month } t - \text{Scheduled principal payment in month } t}$$

This is then annualized as follows to get the CDR:

$$\text{CDR}_t = 1 - (1 - \text{Default rate for month } t)^{12}$$

Note that the conversion of MDR to CDR is identical to the formula for converting SMMs

to CPRs. As described earlier, the default rate represents involuntary prepayments, and the CDR represents the involuntary prepayment speed calculated for nonagency MBS. Voluntary prepayment speeds (i.e., those resulting from refinancing activity and housing turnover) must be calculated separately.

Let’s use the following as an example. Assume that a nonagency pool⁷ with an 8% note rate and 300 months left to maturity has a balance at time t of \$10,000,000. The pool’s scheduled monthly payment is \$77,181.62, comprised of \$66,666.67 in interest and \$10,514.96 in scheduled principal. Assume that the pool receives \$20,000 of voluntary prepayments and \$15,000 in involuntary prepayments.⁸

The monthly voluntary prepayment speed is calculated as follows:

$$\text{Voluntary SMM} = \frac{\$20,000}{\$10,000,000 - \$10,514.96} = 0.002$$

This can then be converted to 2.37% CPR.

The MDR is calculated similarly:

$$\text{MDR} = \frac{\$15,000}{\$10,000,000 - \$10,514.96} = 0.0015$$

which can be converted to 1.78% CDR.

In some cases, the involuntary and voluntary prepayment speeds are combined to calculate a single prepayment speed. In this case, the calculation of a “total CPR” is as follows:

$$\text{Total SMM} = \frac{\$35,000}{\$10,000,000 - \$10,514.96} = 0.0035$$

which can be converted to a total CPR of 4.12%.

There are a number of issues implied by these calculations. First, note that the voluntary SMM and MDR equals the pool’s total SMM. (It is not true, however, that CPRs and CDRs sum to equal the total pool CPR; it is only the monthly rates that are additive.) In using the output of a model, it is also important to ascertain what the vendor means when they quote a “CPR.” Since many systems will show CPRs as the annualized rate of all prepayments (i.e., total CPRs)

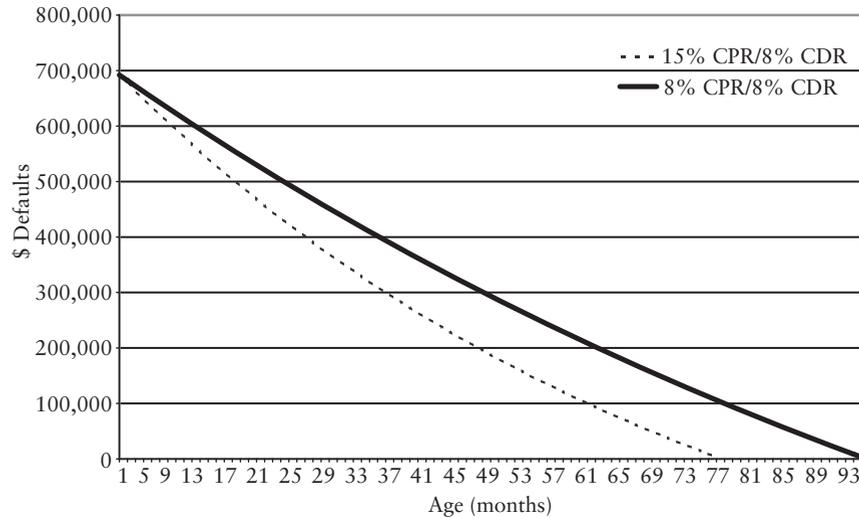


Figure 4 Monthly Dollar Amounts of Defaults on a \$100 Million Pool Using 8% CDR at Different Voluntary Prepayment Speeds

and show CDRs separately, the voluntary prepayment speed must be calculated independently. This can be accomplished by deannualizing the CPRs and CDRs (i.e., converting them to SMMs and MDRs), subtracting the MDR from the SMM, and annualizing the difference. In the above example, the voluntary SMM is 0.0035 less 0.0015 or 0.002, which annualizes to a 2.37% voluntary CPR.

Also note that the CDR metric measures only the amount of defaults and not the amount of losses because actual losses depends upon the amounts that can be recovered on loans in default, adjusted for the costs of collection and servicer advances, if applicable. In the extreme case, if there is full recovery of the unpaid principal balance of the defaulted loans, the losses will be zero except for the costs of recovery. However, depending upon the timing of the recovery of the defaulted loan balances, the cash flows to certain bondholders may be interrupted.

There is also an interesting and important relationship between the voluntary prepayment speed and the dollar amount of defaults in a pool. Every dollar of principal that is prepaid voluntarily is returned at 100 cents on the dollar

and cannot subsequently go into default. Therefore, the dollar amount of a pool's principal that goes into default declines as voluntary prepayment speeds increase, even if the assumed CDR remains constant. This is illustrated in Figure 4. The figure shows the projected dollar amounts of defaults on a \$100 million pool with an 8.5% note rate at 8% CDR for two different voluntary CPRs. At a combination of 15% CPR and 8% CDR, the pool is expected to lose a total of \$21.9 million in face value; the projected amount of defaulted principal using 8% CPR and 8% CDR increases to \$29.0 million.

As with prepayment analysis, there are disadvantages to using constant CDRs that tend to distort credit analysis. A constant CDR assumption is not necessarily consistent with the actual behavior of defaults, and also does not allow the analysis to take variations in the timing of defaults into account. As with prepayments, credit problems have historically tended to be very low immediately after the loans are closed, but generally increase with time as the pool in question ages.

One time-honored methodology is to utilize the Standard Default Assumption (SDA) convention, which assumes that defaults (as

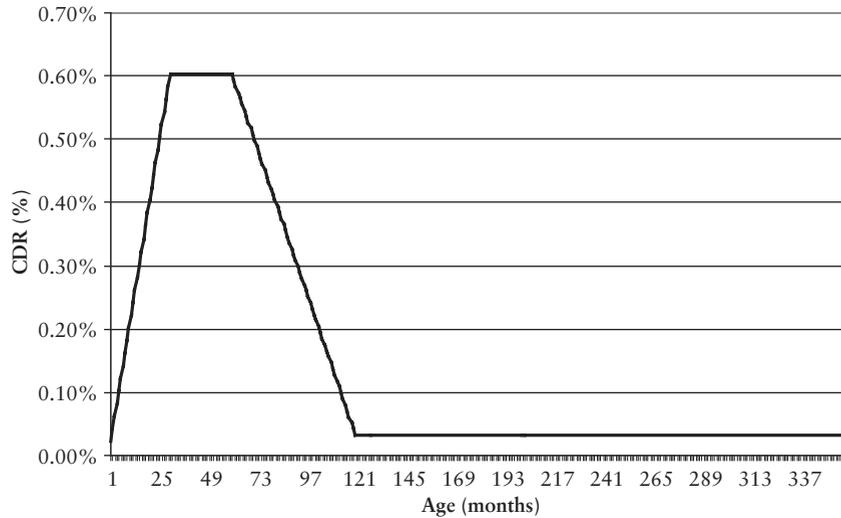


Figure 5 100% SDA Without Effects of Prepayments

measured in annual terms using CDRs) have a fairly consistent pattern over the life of the pool. The SDA model is similar in concept to the PSA convention used in prepayment analysis, and is specified as follows:

- 0.02% initial CDR, rising 0.02% CDR until reaching 0.6% CDR in month 30.
- A constant 0.6% CDR from months 30 to 60.
- A linear decline of 0.0095% between months 61 and 120, reaching 0.03% in month 120.
- A constant 0.03% CDR for the remaining term.

The base SDA curve is shown in Figure 5.

In addition to the prescribed CDR curve described above, the base SDA model explicitly accounts for the effects of voluntary prepayments by assuming a prepayment speed of 150% PSA. One hundred percent SDA at 150% PSA results in cumulative defaults of around 2.73%. The dollar amount of monthly defaults is calculated as the product of monthly default rates or MDRs (i.e., the deannualized CDR) and the monthly balance factor at the projected prepayment speed. Cumulative defaults are the sum of this vector. Table 2 shows how 100% SDA would be calculated, assuming a 6.0% coupon pass-through (as in the prior examples).

A depiction of monthly defaults using the base assumptions of the SDA model at 150% PSA is shown in Figure 6.

Loss Severity Measures

Where the lender has a lien on the property, a portion of the value of the loan can be recovered through the legal recovery process (i.e., through foreclosure and repossession) and subsequent sale of the asset. The difference between the proceeds received from the recovery process (after all transaction costs) and principal balance of the loan is the loss in dollars. The historical loss severity rate in any month is defined as follows:

$$\text{Loss severity rate} = 1 - \frac{\text{Liquidation Proceeds}}{\text{Liquidation Balance}_t}$$

The loss severity rate ranges from 0 to 1 (or 0% to 100%). If the loss severity rate is zero, then liquidation proceeds are equal to the liquidated loan balance. A loss severity rate of 1 (or 100%) means that there are no liquidation proceeds. The loss rate is equal to the annual default rate multiplied by the loss assumption severity. In projecting future cash flows and losses, investors will often use a constant loss

Table 2 Calculation of Monthly Defaults Using 100% SDA at 150% PSA for a Pass-Through with a 5.5% Pass-Through Rate, a WAC of 6.0%, and a WAM of 357 Months

(1)	(2)	(3)	(4)	(5)
Month	100% SDA (in CDRs)	100% SDA (in MDRs) ^a	Bond Factor (@ 150% PSA)	Factor-Adjusted MDR ^b
1	0.080%	0.007%	0.99798	0.0067%
2	0.100%	0.008%	0.99571	0.0083%
3	0.120%	0.010%	0.99318	0.0099%
4	0.140%	0.012%	0.99041	0.0116%
5	0.160%	0.013%	0.98738	0.0132%
6	0.180%	0.015%	0.98410	0.0148%
7	0.200%	0.017%	0.98057	0.0164%
8	0.220%	0.018%	0.97680	0.0179%
9	0.240%	0.020%	0.97278	0.0195%
10	0.260%	0.022%	0.96853	0.0210%
11	0.280%	0.023%	0.96403	0.0225%
12	0.300%	0.025%	0.95930	0.0240%
13	0.320%	0.027%	0.95433	0.0255%
14	0.340%	0.028%	0.94914	0.0269%
15	0.360%	0.030%	0.94372	0.0284%
16	0.380%	0.032%	0.93807	0.0298%
17	0.400%	0.033%	0.93220	0.0311%
18	0.420%	0.035%	0.92612	0.0325%
19	0.440%	0.037%	0.91982	0.0338%
20	0.460%	0.038%	0.91332	0.0351%
21	0.480%	0.040%	0.90661	0.0363%
22	0.500%	0.042%	0.89970	0.0376%
23	0.520%	0.043%	0.89260	0.0388%
24	0.540%	0.045%	0.88531	0.0399%
25	0.560%	0.047%	0.87783	0.0411%
26	0.580%	0.048%	0.87017	0.0422%
27	0.600%	0.050%	0.86233	0.0432%
28	0.600%	0.050%	0.85456	0.0428%
29	0.600%	0.050%	0.84685	0.0425%
30	0.600%	0.050%	0.83920	0.0421%
100	0.192%	0.016%	0.43487	0.0069%
101	0.182%	0.015%	0.43064	0.0065%
102	0.173%	0.014%	0.42644	0.0061%
103	0.163%	0.014%	0.42228	0.0057%
104	0.154%	0.013%	0.41815	0.0054%
105	0.144%	0.012%	0.41406	0.0050%
200	0.030%	0.003%	0.14894	0.0004%
201	0.030%	0.003%	0.14715	0.0004%
202	0.030%	0.003%	0.14538	0.0004%
203	0.030%	0.003%	0.14363	0.0004%
204	0.030%	0.003%	0.14188	0.0004%
205	0.030%	0.003%	0.14016	0.0004%
300	0.030%	0.003%	0.03093	0.0001%
301	0.030%	0.003%	0.03022	0.0001%
302	0.030%	0.003%	0.02952	0.0001%
303	0.030%	0.003%	0.02882	0.0001%
304	0.030%	0.003%	0.02814	0.0001%
305	0.030%	0.003%	0.02745	0.0001%
350	0.030%	0.003%	0.00289	0.0000%
351	0.030%	0.003%	0.00247	0.0000%
352	0.030%	0.003%	0.00204	0.0000%
353	0.030%	0.003%	0.00163	0.0000%
354	0.030%	0.003%	0.00121	0.0000%
355	0.030%	0.003%	0.00080	0.0000%
356	0.030%	0.003%	0.00040	0.0000%
357	0.030%	0.003%	0.00000	0.0000%
Cumulative Defaults				2.75%

^a CDRs are converted to MDRs by using the following formula:

$$\text{MDR} = 1 - (1 - \text{CDR})^{1/2}$$

^b Column (3) \times (4)

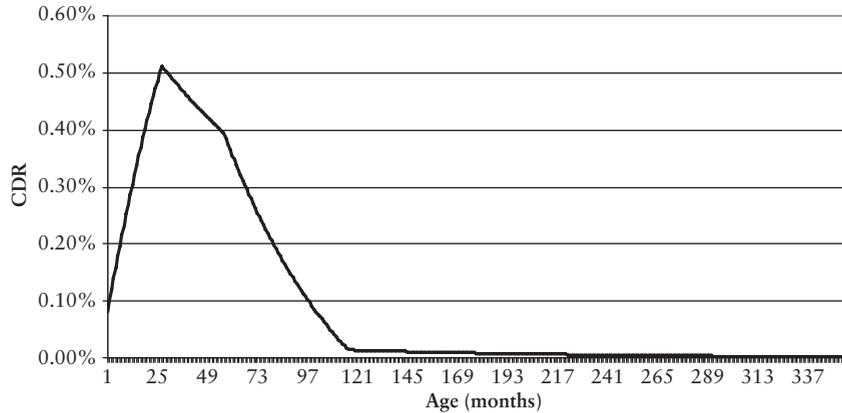


Figure 6 Monthly CDRs for 100% SDA Using 150% PSA

severity assumption based on a combination of loan attributes, projected changes in home prices, and the length of time until liquidation. The percentage of loss severity is then applied to the monthly default amount (generated by using the applicable MDR) in order to calculate monthly losses.

Default and loss severity assumptions (which translate into expected losses) are critical metrics for holders of mortgages and MBS that have exposure to mortgage credit performance. From the viewpoint of issuers, the assumptions used to value and capitalize investments in retained tranches are critical for assessing a firm's value, as any deterioration in the performance of retained tranches can negatively impact overall corporate valuations. Investors in whole-loan mortgages and subordinate MBS routinely use the credit metrics discussed above to analyze the relative value of different alternatives by generating default- and loss-adjusted returns and valuations.

KEY POINTS

- The monthly cash flow from the underlying pool of mortgage loans for a residential mortgage-backed security includes scheduled principal payments, interest payments, and any principal payments made by borrowers that is in excess of the scheduled principal payment. The last component is referred to as prepayments.
- The valuation of residential mortgage-backed securities requires the generation of a residential MBS's cash flow. Prepayment speeds and default rates must be projected in order to do so.
- The performance of a residential MBS depends on the prepayments and performance of the loan pool.
- The measurement of potential and actual cash flow impairment resulting from borrower credit problems is critically important to the analysis of nonagency or private label MBS.
- Complicating the evaluation of prepayments is the interplay between defaults, which are effectively credit-related prepayments, and prepayments attributable specifically to declining interest rates.
- The approach most commonly used to measure prepayment speeds is the conditional prepayment rate, which calculates monthly prepaid principal (i.e., that excludes scheduled principal amortizations) as a percentage of the security's outstanding balance and then annualizes that percentage. The CPR is an annual rate; the corresponding monthly rate is the single monthly mortality rate.
- The Public Securities Association (PSA) prepayment benchmark is expressed as a monthly series of annual prepayment rates

that assumes prepayment rates are low for newly originated mortgages and then will speed up as the mortgages age.

- A loan is classified as delinquent when a borrower fails to make one or more timely payments. Measures of delinquency are designed to gauge whether borrowers are current on their loan payment as well as stratifying unpaid loans according to the seriousness of the delinquency. The calculation method used is determined by the servicer. The two commonly used methods for classifying delinquencies are those recommended by the now-defunct Office of Thrift Supervision (OTS) and the Mortgage Bankers Association (MBA).
- Cumulative default rate and conditional default rate are the two broadly used metrics for quantifying defaults for a mortgage pool. The cumulative default rate is the proportion of the total face value of loans in a pool that have gone into default as a percentage of the total face value of the collateral pool. The conditional default rate is the annualized value of the unpaid principal balance of newly defaulted loans over the course of a month as a percentage of the unpaid balance of the pool (before scheduled principal payment) at the beginning of the month. To compute this measure, the monthly default rate must first be calculated.

NOTES

1. For a detailed discussion of the types of mortgage loans and residential MBS, see Fabozzi, Bhattacharya, and Berliner (2011).
2. Also called the *constant prepayment rate*.
3. This benchmark is commonly referred to as a “prepayment model,” suggesting that it can be used to estimate prepayments. Characterization of this benchmark as a prepayment model is inaccurate. It is simply a market convention. While the PSA has changed its name to the Securities Industry and Financial Markets Association, or SIFMA, the benchmark is still referred to as the “PSA prepayment benchmark.”
4. The calculation can also be presented as a series of formulas, which are available in Chapter 21 Fabozzi (2006).
5. The HEP curve was developed by Prudential Securities based on the prepayment experience of \$10 billion of home equity loan deals.
6. For example, a June 9, 2000, report by Moody’s titled, “Contradictions in Terms: Variations in Terminology in the Mortgage Market,” shows that the reported delinquencies can differ dramatically when the different conventions are used.
7. For clarity’s sake, we assume a simple pool with no credit enhancement.
8. These payments are reported in the monthly remittance reports compiled by a transaction’s trustee.

REFERENCES

- Fabozzi, F. J. (2006). *Fixed Income Mathematics: Analytical and Statistical Techniques*. New York: McGraw-Hill.
- Fabozzi, F. J., Bhattacharya, A. K., and Berliner, S. (2011). *Mortgage-Backed Securities*, 2nd ed. Hoboken, NJ: John Wiley & Sons.

Prepayments and Factors Influencing the Return of Principal for Residential Mortgage-Backed Securities

WILLIAM S. BERLINER

Executive Vice President, Manhattan Advisory Services Inc.

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

ANAND K. BHATTACHARYA, PhD

Professor of Finance Practice, Department of Finance, W. P. Carey School of Business, Arizona State University

Abstract: Prepayments and their impact on principal cash flows are critical components of the valuation, trading, and risk management of residential mortgage-backed securities. Because of this, substantial resources are expended by investors and dealers in understanding and modeling prepayment “speeds.” However, prepayment behavior is not static and has evolved repeatedly since the first prepayment waves in the early 1990s. Moreover, the very definition of “prepayments” has evolved from one focused primarily on borrowers’ refinancing options to one encompassing a plethora of actions and decisions.

In general, a mortgage is a loan that is secured by underlying assets that can be repossessed in the event of default. In the residential housing market, a mortgage is defined as a loan made to the owner of a one- to four-family residential dwelling and secured by the underlying property (i.e., the land, the structure and any improvements). The fundamental unit in the *residential mortgage-backed securities* (MBS) market is the pool. At its lowest common denominator, mortgage-backed pools are aggregations of large numbers of mortgage loans with similar

(but not identical) characteristics. Loans with a commonality of attributes such as note rate (i.e., the interest rate paid by the borrower on the loan), term to maturity, credit quality, loan balance, and product type are combined using a variety of legal mechanisms to create relatively fungible investment vehicles.

To value a residential MBS, a financial modeler must project the cash flow. For an individual mortgage, the monthly cash flow includes the scheduled principal payments (also referred to as amortization), interest

payments, and any prepayments. *Prepayments* are any payments made by borrowers that are in excess of the scheduled principal payment. Consequently, the cash flow depends on the prepayment behavior of the borrowers in the mortgage pool. This risk faced by investors is referred to as prepayment risk and is similar to the risk faced by investors in callable corporate bonds.

Both the valuation and the subsequent performance of a residential MBS depend on prepayments—projected in the former case and realized in the latter case. In this entry, we discuss the underlying factors impacting principal repayment rates. We also draw distinctions between the traditional view of prepayments and a broader one that puts credit-related factors into context.

PREPAYMENT FUNDAMENTALS

Traditional prepayment analysis has focused on borrowers' option to retire their loans prior to maturity. Virtually all mortgage loans allow for the early repayment of principal. Prepayment behavior can be divided into several categories. The first of these is referred to as *turnover*, which occurs when the underlying property is sold and the associated loan is retired. Turnover can occur for a number of reasons:

- The homeowner moves or trades up to a larger house.
- The obligor relocates as part of changes in their job or employment.
- The property is sold subsequent to the death of the homeowner or as part of a divorce settlement.
- The property is destroyed by a fire or other natural disaster.

In all these cases, the resulting proceeds (from either the property's sale or an insurance settlement) are passed on as prepaid principal to the holder of the mortgage. In the event of the sale

of the property, the loan is paid off from the proceeds of the sale; in fact, most loans contain a "due-on-sale" clause ensuring that the loan is retired once the property is sold. Properties are also sold in the event that the obligors encounter financial difficulties. While we discuss credit-related factors at several points in this entry, it is important to note that prepayments resulting from credit events are sometimes taken into account under the broad umbrella of "turnover."

A second form of prepayment can be broadly ascribed to refinancing. This behavior can take a number of forms. A *rate-and-term refinancing* is undertaken solely to reduce the borrower's monthly payment, most commonly due to a decline in the level of consumer mortgage rates. Such a change puts the market rate for new mortgages below the rate of existing loans, creating incentives to refinance. A related activity takes place when borrowers refinance in order to liquefy their home's equity by increasing the balance on their new loan. Such transactions, referred to as *cash-out refinancings*, often are taken as an alternative to second lien loans. Cash-out activity is strongly correlated with rates of home price appreciation which, logically enough, creates the borrower equity extracted through the transaction. Such activity can also be relatively insensitive to traditional refinancing incentives, and has at times boosted prepayment speeds for lower-coupon MBS.

At various points in time, borrowers have also been inclined to refinance from one product into a different one that offers a payment savings. A simple form of *product transition* is to refinance from a fixed-rate loan into an adjustable-rate mortgage (ARM) that offers a lower rate. Borrowers have also transitioned into products with alternative amortization schemes, such as interest-only and negative amortization loans, in order to reduce their monthly payment burdens. Such transitions are contingent on the availability and popularity of alternative products, as well as borrowers' ability (either through lower rates or other nontraditional means) to achieve payment reductions.

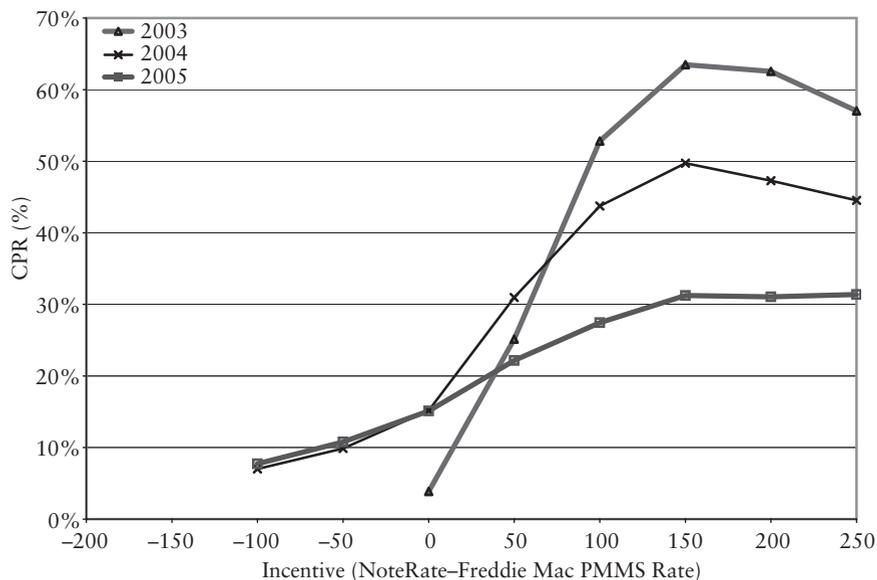


Figure 1 Prepayment S-Curves for Different Years for 30-year Fixed Rate Conventional Loans
Data Source: eMBS.

Another critical factor in prepayments is based on the borrower's financial situation. However, the impact of borrower credit on prepayments is quite complex. Prepayments often result directly from changes to homeowners' financial situation. At its simplest, principal is returned to investors when borrowers default on their loans, although the amount and timing of principal cash flows is subject to many variables. However, credit-related factors also exert more subtle effects on prepayment behavior. For example, borrowers with weak credit, or who don't have significant equity in their home, may not be able to take advantage of declining interest rates by obtaining new loans.

Taken together, these factors and activities result in prepayment speeds that vary across the MBS market. The most common way to assess prepayment speeds within a product group is by a simple view of prepayment speeds as measured by conditional prepayment rates (CPRs) at various levels of refinancing incentive. Prepayment *S-curves* show prepayment speeds for different levels of mortgage rates and/or refinancing incentives. S-curves can be created us-

ing a number of different methodologies and data sources. Either projected or historical prepayment speeds can be shown; additionally, the level of prepayments can be compared by showing either the absolute level of rates or the relative degree of refinancing incentive.

An example of S-curves for different periods of time is shown in Figure 1. The figure shows historical prepayment speeds for 30-year conventional fixed-rate pools exhibited by refinancing incentive (defined as the cohort's WAC less the Freddie Mac 30-year fixed survey rate for that period). The different shapes of the S-curves are indicative of different consumer behaviors. For example, the curve for 2003 was quite steep, indicating that borrowers were extremely sensitive to refinancing opportunities; borrowers that had an incentive to refinance (or, to borrow a term from the option market, were "in-the-money") did so in large numbers. At the same time, prepayments on "out-of-the-money" pools (i.e., those with lower weighted-average coupons (WACs) and no apparent refinancing incentive) were relatively slow, reflecting slow housing turnover

and limited cash-out activity. By contrast, the S-curves for 2004 and 2005 were increasingly flat. This reflected faster housing turnover, brisk levels of cash-out activity, and growing product transition activity for loans with minimal or negative incentives, while in-the-money borrowers were less responsive to apparent refinancing opportunities.

The following subsections discuss the primary drivers of prepayment speeds in more detail.

Turnover

As previously described, *turnover* refers to activity in which the underlying property is sold or liquidated, with the proceeds of the sale subsequently passed through to the holder of the mortgage as a prepayment. There are a number of ways to observe the level of turnover. A simple way to assess turnover is to look at the prepayment speeds of out-of-the-money MBS pools, such as, for example, prepayment speeds on Fannie 4.0s when mortgage rates are 5% or higher.

However, prepayment speeds for lower-coupon MBS can also be influenced by factors other than turnover. For example, high levels of cash-out refinancings (when borrowers refinance primarily to monetize the equity in their homes) will also increase prepayment speeds on out-of-the-money coupons. Product transition activity, which was widespread from 2004 through early 2007, can also distort the normal calculation of “in-the-moneyness.” As discussed later in this entry, transitions typically are associated with the widespread availability and popularity of products that allow borrowers to reduce their monthly payment obligations through either lower loan rates or alternative amortization schemes.

A truer estimate of housing turnover can be obtained by calculating existing home sales for single-family homes as a percentage of the number of such homes owned. Existing home sales data are published monthly by the National Association of Realtors, while the number of

single-family homes outstanding is reported by the Census Bureau on a quarterly basis, subject to periodic adjustments. Research indicates that turnover has varied over time, primarily reflecting changes in the level of home sales.

It is tempting to associate elevated housing turnover with robust growth in home prices. Purely speaking, however, housing turnover is not directly associated with real estate price appreciation, but rather with the level of home sales activity and the number of completed transactions. While home prices and sales are highly correlated, it is conceivable that home prices could stagnate while sales activity remains firm, and vice versa.

Refinancing

Refinancing (“refi”) activity can be broadly defined as transactions where borrowers replace their existing mortgage with a new loan, using the proceeds from the new loans to pay off their preexisting mortgage obligations. While it encompasses a number of different activities, it most commonly occurs when the prevailing level of interest rates declines to the point where borrowers can take out new loans and reduce their monthly payments (after accounting for transaction costs and potential penalties).¹

As noted already, refinancing activity can be broadly categorized as rate-and-term refinancings, where borrowers act solely to reduce their mortgage payments, and cash-out refinancings for which the new loan is larger than the one being retired. Rate-and-term refis are easily conceptualized as a form of option exercise. In a fashion similar to a corporation calling a debt issue, homeowners can reduce their required debt service obligations by calling their current loans carrying above-market rates and issuing new debt.

However, the nature of mortgage lending complicates borrowers’ refinancing decisions. Homeowners refinancing their loans are subject to a variety of costs and fees, many of which are fixed. The expected monthly savings, by

contrast, is a function of the size of the loan in question. This implies that refinancing incentives are strongly impacted by loan size, as smaller loans typically require a greater refinancing incentive in order to trigger refinancing activity. Take, for example, two loans with 5% note rates and balances of \$200,000 and \$400,000, respectively. A 50 basis point rate savings reduces the payment on the \$200,000 loan by \$60 per month, while the same rate savings reduces the larger loan's monthly payment by roughly \$120. If both loans are subject to \$1,000 in refinancing costs, the borrower with the \$400,000 loan will recoup the initial outlay in month 8; the borrower with the smaller loan needs more than double the time to break even. This makes loan size a critical variable in modeling and projecting future prepayment speeds.

Cash-out refinancings are commonly viewed as a subset of overall refinancing activity. For example, Freddie Mac defines cash-out refis as transactions where the new loan is at least 5% larger than the original one, and reports cash-outs as a percentage of overall prepayment activity. The level of cash-out activity has varied significantly over time. For example, the relative level of cash-out activity was extremely high in the late 1980s and 1990s, as well as in the period between 2003 and 2007.

The primary driver of cash-out activity at any point in time is the amount of equity borrowers have in their homes. In turn, equity is a function of both the original equity in the home (i.e., the inverse of a loan's loan-to-value (LTV) ratio) and the rate of home price appreciation since the home was purchased.

Aggregate refinancing incentives can be observed by examining the distribution of note rates within the MBS universe at various points in time. Keep in mind that the outstanding mortgage population is always changing, as new loans are issued and older loans are retired. The distribution of note rates for the population of outstanding loans is strongly impacted by refinancing activity, which can be thought

of as recycling older high-rate loans into new mortgages with lower rates.

A useful technique is to compare the outstanding balances and the cumulative percentages of note rates for MBS products at different points in time. The cumulative balance percentages are calculated as follows:

- Divide the outstanding market balances into discrete segments or "buckets" by WAC. (The following analysis uses 12.5 basis point WAC buckets.)
- For each WAC bucket, calculate the percentage of the remaining balances with note rates equal to and below that bucket.

For example, if the lowest WAC bucket is 5.0% to 5.124% and it represents 2% of the remaining balance, its cumulative percentage is 2%. If the next WAC bucket (5.125% to 5.249%) comprises 6% of the unpaid balance of the market, its cumulative balance is therefore 8%. This process is completed for all WAC buckets. This technique is particularly useful in assessing the "refinanceability" of the market at particular points in time.

FACTORS INFLUENCING PREPAYMENT SPEEDS

In understanding and evaluating prepayment behavior, the level of consumer mortgage rates is the single factor upon which most attention is paid. However, there is no single "market" rate that analysts can observe. There are always differences in the rate offerings of different lenders; since loans are the "product" they offer, it's not surprising that there are pricing discrepancies. Individual lenders also have a variety of offerings, with different combinations of interest rates and up-front fees (or "points," which vary inversely with the rate offered). While these options give borrowers choices between up-front costs and monthly payments, the relationship between rates and points is highly lender-specific and a function of their pricing algorithms. Finally, lenders seek

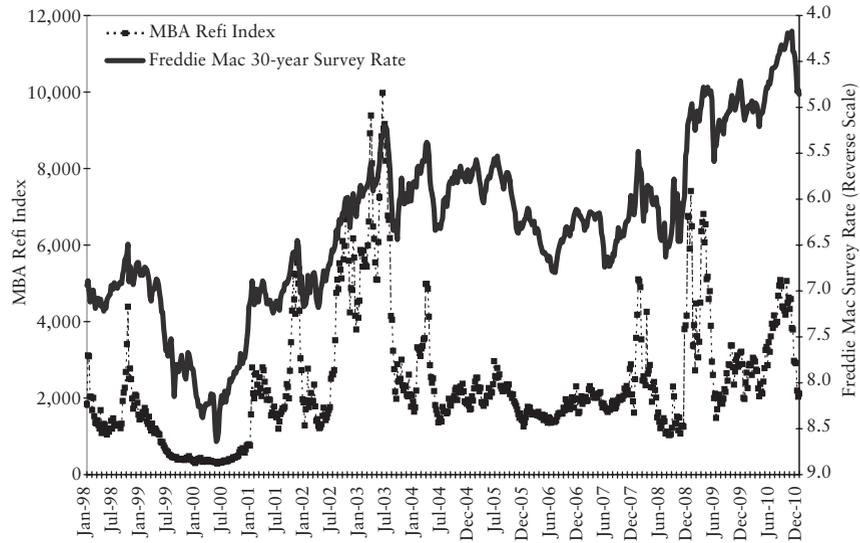


Figure 2 MBA Refi Index versus Freddie Mac 30-year Survey Rate
Data Sources: Mortgage Bankers Association and Freddie Mac.

to price in the risk of loans to various borrowers in a series of activities broadly classified under “risk-based pricing.”²

However, a variety of outside factors that influence prepayment speeds and refinancing behavior can be outlined. These include exogenous factors, mortgage industry economics, and consumer behaviors and preferences.

Borrower Inefficiencies

Rational borrowers will always seek to lower their borrowing costs by refinancing their debts. Refinancing opportunities present themselves to both institutional and individual borrowers. Unlike corporations and municipalities, however, residential borrowers are relatively inefficient in capitalizing on refinancing opportunities. (If mortgagors were efficient, for example, few if any premium pools would be outstanding; however, there were approximately \$110 billion of 30-year Fannie Maes with coupons of 6.5% and higher at the end of 2010.)

Borrower inefficiencies exist for a number of reasons. Homeowners have varying degrees of awareness of financial market rates and condi-

tions, and as a result are not always cognizant of refinancing opportunities. Borrowers often hear about declines in rates from their friends and co-workers; they also may read about it in the financial press or see it discussed on news programs. These are collectively referred to as *media effects*. While the growth of the financial press (with information available from print, television, and the Internet) has improved refinancing efficiency over time, it often takes a significant and noteworthy drop in rates to generate conversation and media “buzz.” This explains the tendency for refinancings to occur in waves, as illustrated in Figure 2. The figure shows mortgage rates (using Freddie Mac’s 30-year survey rate as a proxy, shown on a reverse scale) versus refinancing activity, using the Mortgage Bankers Association’s refinancing applications index. The figure indicates that refinancing activity often remains tepid for long periods of time, but spikes when mortgage rates decline beyond some indeterminate threshold.

In addition, the costs associated with refinancing alter the refinancing economics for borrowers. The need to overcome cost hurdles serves to inhibit refinancing activity and complicates

refinancing decisions. As noted previously, this is particularly relevant for borrowers with smaller loan balances, who typically require a greater refinancing incentive before engaging in rate-and-term refinancings.

Refinancing efficiency has also been impacted by the structure of the mortgage industry. Beginning in the mid-1990s, lenders became increasingly adept at marketing their products and generating refinancing activity. Some of these activities involve directly contacting existing customers, while others involve mass marketing through television commercials, print advertisements, and direct mail and phone solicitations. Also contributing to the marketing effort was a cadre of mortgage brokers and other “third-party originators” who acted as agents linking lenders and borrowers. These developments contributed to improved refinancing efficiency.

The events that culminated in the financial crisis in 2008, however, led to sharp contraction in “wholesale” lending activities. Brokers were blamed for poor loan quality and sloppy paperwork; since they did not make loans directly, they arguably had no incentive to insure the quality of their loans. As a result, many smaller originators that were dependent on the wholesale channel failed, while a number of large originators curtailed or severely limited their interaction with third-party lenders. This development in turn served to impair borrowers’ ability and/or willingness to capitalize on refinancing opportunities.

Finally, additional factors impact refinancing activities. After 2007, for example, a combination of significantly tighter lending standards, fewer product offerings, and declining borrower equity due to falling home prices acted to further depress refinancing activity. Referring to Figure 2, the inability of the MBA’s refi index to reach and maintain high levels reflected the fact that the pool of borrowers with the ability to refinance was quickly exhausted when mortgage rates plummeted beginning in early 2009.

Product Choices and Transitions

Both rate-and-term and cash-out refinancing activity is at times influenced by *product transitions*. This means that borrowers can lower their monthly payment by refinancing from one product into another. This type of activity has varied over time, depending on the availability, popularity, and pricing of alternative products. When the yield curve has been relatively steep, for example, large numbers of borrowers have sometimes refinanced out of fixed rate loans into adjustable rate products.

Transition activity has varied substantially over time, however, driven by both lender offerings and consumer preferences. Prior to mid-2003, for example, ARMs were a niche product targeted primarily to first-time home buyers. In the summer of 2003, however, ARM volumes rose fairly dramatically, as consumers refinanced out of fixed rate products into newly popular hybrid ARMs. This reflected both consumers’ increased comfort with adjustable rate loans as well as marketing efforts by mortgage lenders designed to maintain issuance volumes. By mid-2007, borrowers once again eschewed ARMs, in part due to bad publicity emphasizing their riskiness.

These abrupt changes in behavior are illustrated in Figure 3. The figure contains a scatterchart showing the Freddie Mac 30-year fixed survey rate on the horizontal axis, and the percentage of loans taken as ARMs on the vertical axis. The figure demonstrates the existence of three distinct regimes. ARMs were relatively unpopular in the years prior to mid-2003, and only reflected a large share of activity when mortgage rates were relatively high. From mid-2003 through early 2008, by contrast, the percentage of ARMs was relatively high irrespective of the level of mortgage rates and, by implication, refi activity. After the beginning of 2008, ARMs again fell out of favor; by 2010 they comprised less than 10% of new loan applications.

The varying popularity of fixed-to-ARM refinancings has several implications. Because

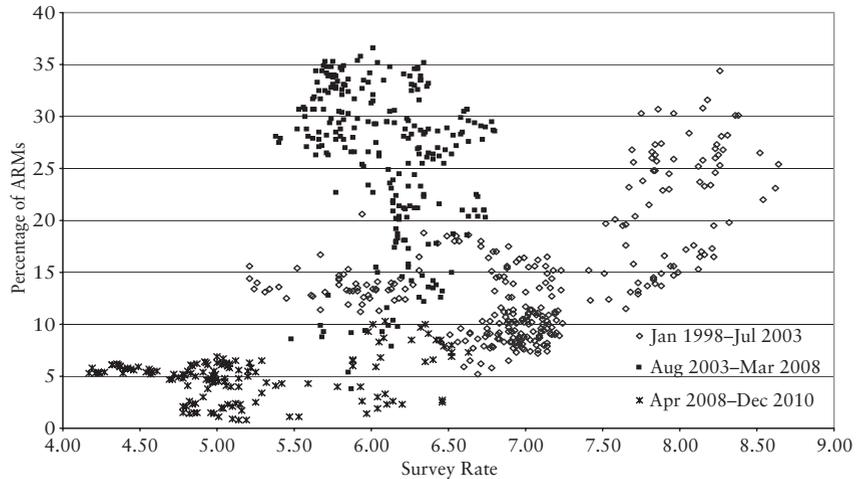


Figure 3 Freddie Mac 30-Year Survey Rate versus MBA's ARM Percentage
Data Sources: Mortgage Bankers Association and Freddie Mac.

of the generally upward slope of the yield curves, ARM rates are typically lower than fixed rates. This means that borrowers willing to utilize adjustable rate products will be presented with an apparent refinancing incentive more often than those borrowers that eschew ARMs and will only consider fixed rate products. (Of course, this savings is only guaranteed for an ARM's fixed rate or "teaser" period.) Taking available ARM rates into account means that more borrowers can reduce their mortgage rates by refinancing.³ As a result, regimes where ARMs are a popular product choice (due to consumer preferences and/or a steep yield curve) are characterized by steady levels of refinancing activity and relatively flat S-curves.

Alternatively, when short rates rise and push ARM rates higher, fixed-to-ARM refinancing incentives are reduced. In fact, regimes associated with flat yield curves are often characterized by ARM-to-fixed transitions, as borrowers seek to lock in lower long-term rates. Taken together, these phenomena indicate that refinancing behavior is not simply dictated by the level of intermediate and long interest rates. The levels of *all* interest rates, as well as the shape of the

yield curve, are important drivers of refinancing incentives and prepayment activity.

Large-scale transitions also have been observed as borrowers utilized loan products with alternative amortization schedules and payment schemes. As a simple example, a borrower with a \$200,000 loan balance and a 30-year loan with a fixed 5% note rate would have monthly P&I payments of \$1,074. If they refinanced into an interest-only loan with the same term and note rate, their new monthly payment would be \$833, an initial savings of \$240. However, the savings would only be available for the period that the borrower was allowed to make interest-only payments; after that point, the loan is "recast" (i.e., the payments are recalculated) over the remaining term. If the borrower chooses a new loan with an interest-only period of 10 years, the post-recast monthly payment would be \$1,320, significantly higher than the payment on the original loan.

The borrower's decision thus trades off early savings for a sharply increase monthly payment (or *payment shock*) at the recast. While such decisions were popular during the period of widespread product transitions, the mortgage crisis of 2007 led to the realization that these

types of transitions exposed both borrowers and lenders to serious embedded risks. As a result, transitions into alternative payment products became fairly rare by 2008.

Changes in Homeowner Equity and Credit

As noted earlier, the experience of the post-2007 period has highlighted the interrelationship between prepayments and home prices and, by extension, borrower credit. We already highlighted the importance of cash-out refinancings and the critical role that home price appreciation plays in this activity. In addition, deteriorating borrower credit (of which homeowner equity is a crucial element) often directly results in prepayments, as we discuss next.

However, changing home prices and borrower credit have other subtle affects on prepayments. For example, borrowers often are presented with an enhanced refinancing incentive when their credit improves. If they took loans with relatively high rates because of risk-based pricing, they can capitalize on their improved situation by refinancing. Such “credit curing” can be related to economic factors such as improving labor markets and consumer credit conditions, particularly when observing local or regional activity. A similar phenomenon is associated with rapid increases of home prices. Borrowers with high LTVs who were saddled with higher risk-based mortgage rates and/or mortgage insurance premiums can lower their payments once their homes appreciate in value, even if the overall level of mortgage rates remains unchanged.

Alternatively, borrower credit can also act to slow prepayment speeds. Borrowers with deteriorating credit may not be able to capitalize on declining interest rates if they cannot obtain new loans because of tighter credit standards. Declining real estate values can also prevent homeowners from refinancing existing loans by reducing or eliminating their equity. If homeowners’ equity disappears or becomes nega-

tive (a situation often referenced as “being underwater”), they may lose the ability to obtain new loans. Moreover, significant declines in home values ultimately serve to constrain homeowners from selling their properties, as they would be forced to realize large losses on their homes. These developments are collectively called *prepayment lock-in*, and serve to slow both refinancing- and turnover-related prepayments.

Time

Prepayment rates vary with the passage of time. In addition to purely random variations, fairly predictable changes occur to prepayment speeds due to factors that are independent of interest rates. The behavior of borrowers undergoes a variety of secular and cyclical changes as time elapses; in addition, the composition of closed loan populations (i.e., loans collateralizing a pool) changes as the pool ages and loans drop out for any number of reasons.

Time-related factors mean that evaluating any MBS at a single constant speed is unrealistic. This realization was first incorporated into the PSA prepayment benchmark, which recognized the fact that loans are more likely to prepay as they age (or season). Borrowers are disinclined to prepay their loans immediately after issuance, but become increasingly open to the possibility as time elapses. This is due to a variety of factors:

- Borrowers typically are reluctant to undertake the effort and expense of refinancing until their loans are at least a few months old.
- Borrowers are unlikely to sell their properties and move immediately after purchasing a home. This is true even for homeowners that relocate frequently; evidence suggests that they tend to stay in their homes for at least a year.
- It takes some time for borrowers to build equity in their property (assuming, of course, a regime of rising home prices).

The key insight introduced by the PSA model is the concept that prepayment speeds are not constant over time, especially early in loans' lives. It is, however, simplistic in its assumption of a constant prepayment speed after 30 months, and does not account for other time-related behaviors. One such factor is seasonality, which suggests that prepayments typically increase during spring and summer months. Another behavior, burnout, accounts for the observation that loans remaining in a population are less likely to refinance after a certain point in time. The underlying logic is that borrowers that have not availed themselves of refinancing opportunities lack the ability and/or the inclination to do so.

The combination of these behaviors means that a time series of CPRs generated by a prepayment model (as well as the realized prepayment speeds for any security)—the *CPR vector*—will look very different from the equivalent speeds quoted as percentages of the PSA model.

Time-related changes to prepayment speeds are even more profound for mortgage products that do not require fixed monthly payments over their life. For example, ARMs typically experience a spike in prepayment speeds as the loans approach their first reset date. (For example, the monthly payments on 5/1 hybrid ARMs change when the loans reset at month 60.) Interest-only loans exhibit comparable behavior, as their required monthly payments increase once the IO period expires. All such products exhibit prepayment patterns reflecting variations in the loans' monthly payments and, by implication, refinancing incentives.

The spike in ARM speeds at their reset results from a variety of factors. Unlike homeowners in Europe, U.S. borrowers have traditionally been somewhat averse to adjustable-rate loans. This means that borrowers often prepay hybrid ARMs simply to avoid being exposed to changing interest rates and variable payments. It also is a function of the level of the benchmark rate at the reset; in regimes where the yield curve is flat or inverted, the new loan rates are often

higher than the teaser rate. The resulting payment shock creates a refinancing incentive for borrowers during periods when the new rate is higher than that for either a new ARM or a fixed rate loan.

Empirical evidence shows a sharp increase in CPRs at the reset; in addition, models also project a cyclical increase in speed every 12 months thereafter, corresponding with the annual rate resets for the loans as well as normal seasonal patterns.⁴

DEFAULTS AND "INVOLUNTARY" PREPAYMENTS

The mortgage crisis that erupted in early 2007 underscored the critical role of credit performance in all sectors of the mortgage and MBS markets. In the past, investors assumed that senior nonagency MBS were "money-good" by virtue of their triple-A ratings. The collapse of mortgage performance both reinforced the importance of sound credit analysis of private-label securities, while also giving investors a painful and expensive lesson on the factors influencing residential mortgage credit performance.

Factors Influencing Default Frequency and Credit Performance

The general thinking has long been that borrower equity simply provides a cushion for the lender in cases when the home must be repossessed. However, a critical lesson learned from the post-2006 experience is that borrower credit performance and home prices are strongly interrelated at a number of levels, and that high-LTV loans have, all else being equal, an increased likelihood of default.

At its most basic, appreciating home prices give borrowers the ability to monetize their home's equity in order to meet their financial obligations and mitigate cash flow problems. In addition, steady or rising home prices also impact the resolution of troubled loans.

Delinquent borrowers that have equity in their homes can sell their properties and, using the net proceeds, pay off their loans instead of going into foreclosure. In theory, borrowers should never default if their homes' values are great enough to extinguish the loan and pay the associated costs. Borrowers whose homes have declined to the point where their LTVs are greater than 100% (i.e., where their loans are greater than the value of their homes) do not have this option. This accounts for why some loan vintages (such as the year 2000) have experienced relatively high levels of delinquency but limited defaults and losses; borrowers in financial difficulty were able to sell their homes and emerge "whole."⁵

The decline in home prices that began in 2007 resulted in unexpectedly large increases in defaults. The loss of home equity induced numerous borrowers to exercise the option embedded in any collateralized loan that allows the collateral to revert back to the lender. It is axiomatic in corporate credit theory that borrowers are expected to default on loans once the value of the loans' collateral declines below the value of the loans themselves. However, the mortgage sector has long operated under the assumption that obligors rarely walk away from the properties because of the importance of dwellings to families' well-being. This behavior was untested until 2007, in large part because home prices have never before experienced significant and widespread declines. However, the new phenomenon of the "strategic default" emerged during the mortgage crisis, where large numbers of homeowners with income and assets sufficient to service their loans nevertheless ceased making monthly mortgage payments.

The emergence of this activity has a number of implications. The most important realization is that home prices and mortgage credit performance are closely linked. In this light, the strong credit performance exhibited by the mortgage market since the 1950s was arguably skewed higher by decades of steady home price appreciation. This assertion implicitly argues that res-

idential mortgage loans are riskier assets than previously assumed. In addition, mortgage underwriters have placed undue faith in metrics such as credit scores which, while valuable, cannot serve as reliable proxies for borrowers' willingness to service their loans during times of financial distress.

Voluntary and Involuntary Prepayments

Once borrowers cease making regular payments, the loans eventually go into default, meaning that the borrowers lose title to the underlying properties. The properties are subsequently liquidated, typically by being placed in foreclosure; this means that the servicer eventually takes possession of the property and sells it. The proceeds of the sale, less associated costs, are categorized as recovered principal or *recoveries*. Since recoveries are typically less than the amount of the loan, some entity must absorb a principal loss.

Losses for agency MBS are absorbed by the entity or agency that guaranteed them. At some point, seriously delinquent loans in agency pools are classified as "nonperforming" and subsequently bought out of the pools, either by the GSEs or (in the case of FHA and VA loans) the servicer. Because of the principal guaranty, the full face value of principal is quickly returned to investors. This means that all unscheduled principal payments can be captured in a single "prepayment speed" reported for the security in question. This measure is calculated based on the total principal repaid on the pool and the breakdown (either reported or estimated) between amortizations and prepayments (i.e., between scheduled and unscheduled principal payments). As a result, many agency securities exhibited increased prepayment speeds during periods of poor credit performance and widespread delinquencies, particularly when the agencies change their buyout policies.⁶ (This also blurs the line between credit-related prepayments and normal housing "turnover.")

By contrast, traditional and credit-related prepayments must be calculated and reported separately for nonagency securities. This is because of the fact that credit support for these securities is internal; deals are structured such that senior bonds in a transaction have priority over other bonds in receiving principal and interest. Since the transaction itself will absorb incurred losses, traditional prepayments (which return all of principal to the security holder) and credit-related prepayments (which result in shortfalls that must be allocated within structures) must be segregated. As a result, private-label securities report both *voluntary prepayments*, which encompass traditional prepayment activity, and credit-related *involuntary prepayments*. The latter result from defaults or other events specifically related to credit events (such as short sales of homes), while also accounting for the likelihood that less than the full amount of principal will be returned to the transaction (or, more accurately, the trust holding the deal's collateral).

These factors complicate the projection and calculation of prepayment speeds for private-label securities. Voluntary prepayments are typically quoted as *VPRs*, which stands for voluntary prepayment rate. They are calculated similarly to a CPR, in which a monthly percentage of prepaid principal (sometimes called a VMM) is annualized. Involuntary prepayment speeds are quoted as conditional default rates (CDRs) which are calculated by annualizing the monthly default rates or MDRs. Note that the sum of the monthly VMMs and MDRs equals the total deal SMM for any particular month.

Involuntary prepayments require additional metrics to be reported. In addition to the rate of default, an estimate must be made of the loss severity (which indicates how much of the defaulted principal amount is returned to investors) as well as the lag between the time when loans go into default (i.e., when the borrowers lose title to the properties) and when the trusts receive the recovered principal.

Interactions Between Prepayments and Defaults

There are some interesting interactions between voluntary and involuntary prepayment speeds that impact the analysis of private-label securities. All things equal, fast prepayments enhance the performance of these securities; faster return of principal means that there is less principal outstanding to go into default. At the same assumed CDR, faster voluntary prepayment speeds (i.e., a higher VPR assumption) will typically result in higher projected yields and returns.

This assertion is somewhat simplistic, however, since it doesn't take the changing composition of the pool into account. For example, it is unlikely that the CDR would remain constant under the different VPR assumptions, as the profile of any closed population of mortgages changes over time. In addition to home prices and economic conditions, the composition of the collateral pool backing a transaction evolves as the result of attrition. Loans pay off over time as a result of both voluntary and involuntary factors. Voluntary prepayments negatively impact the composition of a pool because "better" borrowers (i.e., those with stronger credit and/or more equity in their homes) are able to take advantage of refinancing opportunities; since weaker borrowers are locked into their existing loans, the credit profile of the remaining population deteriorates. This is known as *adverse selection*, and suggests that the credit quality of a pool typically declines over time, all things equal.

The high level of defaults experienced during the mortgage crisis also created a new and unanticipated phenomenon. High levels of defaults means that weaker borrowers are dropping out of the collateral pools. In turn, the remaining borrowers generally have stronger credit, meaning that the population's credit profile improves over time. This is especially noteworthy during periods of declining home prices. Borrowers with poor credit (i.e., both those unable or unwilling to service their loans)

go into default in large numbers, while stronger borrowers who are nonetheless “locked in” by a lack of equity continue to service their loans and remain in the pool. This process is sometimes called *favorable selection*, and was most prominently observed in subprime and alt-A pools, which experienced very high levels of defaults.

Neither the processes of adverse nor favorable selection take place in a vacuum. For example, the performance of a cohort assumed to be adversely selected (i.e., having experienced relatively high levels of voluntary prepayments) will improve in the face of home price appreciation. Alternatively, a population of subprime loans may experience a renewed surge in defaults if money-market rates increase sharply. Since many subprime loans have adjustable-note rates with very high loan margins, rising rates create widespread payment shock that challenges the ability of borrowers to service their loans.

KEY POINTS

- Traditional prepayment analysis has focused on borrowers’ option to retire their loans prior to maturity.
- The two primary drivers of prepayment behavior are turnover and refinancing.
- Turnover occurs when the underlying properties are sold and the associated loan is retired.
- Refinancing behavior includes rate-and-term refinancing (undertaken to reduce the borrower’s monthly payment, most commonly due to a decline in the level of consumer mortgage rates) and cash-out refinancing (often are taken as an alternative to second lien loans and strongly correlated with rates of home price appreciation).
- The most common way to assess prepayment speeds within a product group at various levels of refinancing incentive is with the prepayment S-curves. These curves show pre-

payment speeds for different levels of mortgage rates and/or refinancing incentives.

- In understanding and evaluating prepayment behavior, the level of consumer mortgage rates is the single factor to which most attention is paid.
- Outside factors that influence prepayment speeds and refinancing behavior include exogenous factors, mortgage industry economics, and consumer behaviors and preferences.

NOTES

1. For a more detailed discussion, see Chapter 3 in Fabozzi, Bhattacharya, and Berliner (2011).
2. See Bhattacharya, Berliner, and Fabozzi (2008).
3. If ARM rates are low enough, virtually the entire fixed rate coupon stack can be considered in-the-money.
4. See Bhattacharya, Berliner, and Fabozzi (2008).
5. In these cases, the transaction is recorded as a home sale and captured under “turnover.”
6. In early 2010, Fannie Mae and Freddie Mac instituted policies in which loans that were 120 days or more delinquent were automatically bought out of pools. Prior to that, buy-outs had been left to their discretion. The process of buying out large numbers of seriously delinquent loans led to sharp short-term spikes in prepayment speeds, as well as huge writedowns for Fannie Mae and Freddie Mac.

REFERENCES

- Bhattacharya, A. K., Berliner, W. S., and Fabozzi, F. J. (2008). The interaction of MBS markets and primary mortgage rates. *Journal of Structured Finance* 14, 3: 16–36.
- Fabozzi, F. J., Bhattacharya, A. K., and Berliner, W. S. (2011). *Mortgage-Backed Securities*, 2nd ed. Hoboken, NJ: John Wiley & Sons.

Operational Risk

Operational Risk

ANNA CHERNOBAI, PhD

Assistant Professor of Finance, M.J. Whitman School of Management, Syracuse University

SVETLOZAR T. RACHEV, PhD, DrSci

Frey Family Foundation Chair-Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: At one time the belief was that financial institutions are exposed to two main risks. Operational risk was regarded as a mere part of “other” risks. That view has changed. This risk is now viewed as a major risk faced by financial institutions as the world financial system has been shaken by a number of banking failures since the mid 1980s, and the risks—that internationally active banks, in particular, have had to deal with—have become more complex and challenging. More than 100 operational losses exceeding \$100 million in value each and a number of losses exceeding \$1 billion have impacted financial firms globally since the end of the 1980s. There is no question that the cause is unrelated to market or credit risks. Such large-scale losses have resulted in bankruptcies, mergers, or substantial equity price declines of a large number of highly recognized financial institutions.

A long-held belief is that credit risk and market risk have been considered the two largest contributors to the risks faced by financial entities such as banks, insurance companies, and asset management firms. Credit risk is the risk of counterparty failure; market risk is the loss due to changes in market indicators, such as equity prices, interest rates, and exchange rates. It is now recognized that operational risk is a major risk faced by financial entities. In general terms, *operational risk* is the risk of loss resulting from inadequate or failed internal processes, people, or systems or from external events. This risk encompasses legal risks, which includes, but is

not limited to, exposure to fines, penalties, or punitive damages resulting from supervisory actions, as well as private settlements.

Operational losses have been reflected in banks’ balance sheets for many decades. Operational risk affects the soundness and operating efficiency of all banking activities and all business units. Most of the losses are relatively small in magnitude—the fact that these losses are frequent makes them predictable and often preventable. Examples of such operational losses include losses resulting from accidental accounting errors, minor credit card fraud, or equipment failures. Operational risk-related

events that are often more severe in the magnitude of incurred loss include tax noncompliance, unauthorized trading activities, major internal fraudulent activities, business disruptions due to natural disasters, and vandalism.

Until around the 1990s, the latter events have been infrequent, and even if they did occur, banks were capable of sustaining the losses without major consequences. This is quite understandable because the operations within the banking industry until about the middle of the 1980s have been subject to numerous restrictions, keeping trading volumes relatively modest and diversity of operations limited. Therefore, the significance of operational risk (whose impact is positively correlated with income size and dispersion of business units) has been perceived as minor, with limited effect on management's decision making and capital allocation when compared to credit risk and market risk. However, serious changes in the global financial markets have caused noticeable shifts in banks' risk profiles.

In this entry, we discuss some key aspects that distinguish operational risk from credit risk and market risk. They are related to the arrival process of loss events, the loss severity, and the dependence structure of operational losses across a bank's business units.

OPERATIONAL RISK DEFINED

Let's begin by distinguishing operational risk from other categories of financial risk. Operational risk is, in large part, a firm-specific and nonsystematic risk.¹ Early publications of the Bank for International Settlements (BIS) defined operational risk as:²

- Other risks.
- "Any risk not categorized as market and credit risk."
- "The risk of loss arising from various types of human or technical errors."

Other definitions proposed in the literature include:

- Risk "arising from human and technical errors and accidents."³
- "A measure of the link between a firm's business activities and the variation in its business results."⁴
- "The risk associated with operating a business."⁵

The formal definition that is currently widely accepted was initially proposed by the British Bankers Association (2001) and adopted by the BIS in January 2001. Operational risk was defined as "the risk of direct or indirect loss resulting from inadequate or failed internal processes, people or systems or from external events."

The industry responded to this definition with criticism regarding the lack of a clear definition of "direct" and "indirect" losses. A refined definition of operational risk dropped the two terms, hence finalizing the definition of operational risk as:

Operational risk is the risk of loss resulting from inadequate or failed internal processes, people or systems, or from external events. (BIS, 2001b, p. 2)

This definition includes legal risk, but excludes strategic and reputational risk (these will be defined soon). The definition is "causal-based," providing a breakdown of operational risk into four categories based on its sources: (1) people, (2) processes, (3) systems, and (4) external factors. According to Barclays Bank, the major sources of operational risk include operational process reliability, IT security, outsourcing of operations, dependence on key suppliers, implementation of strategic change, integration of acquisitions, fraud, error, customer service quality, regulatory compliance, recruitment, training and retention of staff, and social and environmental impacts.⁶

Large banks and financial institutions sometimes prefer to use their own definition of operational risk. For example, Deutsche Bank defines operational risk as

*potential for incurring losses in relation to employees, contractual specifications and documentation, technology, infrastructure failure and disasters, external influences and customer relationships.*⁷

The Bank of Tokyo-Mitsubishi defines operational risk as “the risk of incurring losses that might be caused by negligence of proper operational processing, or by incidents or misconduct by either officers or staffs.”⁸

In October 2003, the U.S. Securities and Exchange Commission (SEC) defined operational risk as:

*the risk of loss due to the breakdown of controls within the firm including, but not limited to, unidentified limit excesses, unauthorized trading, fraud in trading or in back office functions, inexperienced personnel, and unstable and easily accessed computer systems.*⁹

OPERATIONAL RISK EXPOSURE INDICATORS

The probability of an operational risk event occurring increases with a larger number of personnel (due to increased possibility of committing an error) and with a greater transaction volume. Examples of operational risk exposure indicators include:¹⁰

- Gross income.
- Volume of trades or new deals.
- Value of assets under management.
- Value of transactions.
- Number of transactions.
- Number of employees.
- Employees’ years of experience.
- Capital structure (debt to equity ratio).
- Historical operational losses.
- Historical insurance claims for operational losses.

For example, larger banks are more likely to have larger operational losses. Shih, Samad-Khan, and Medapa (2000) measured the dependence between a bank size and operational loss amounts. They found that, on average, for every unit increase in a bank size, operational losses are predicted to increase by roughly a

fourth root of that. This means that when they regressed log-losses on a bank’s log-size, the estimated coefficient was approximately 0.25. In a different study, Chapelle, Crama, Hübner, and Peters (2005) estimated the coefficient to be 0.15.

CLASSIFICATION OF OPERATIONAL RISK

Operational risk can be classified according to

- The nature of the loss: internally inflicted or externally inflicted.
- The impact of the loss: direct losses or indirect losses.
- The degree of expectancy: expected or unexpected.
- Risk type, event type, and loss type.
- The magnitude (or severity) of loss and frequency of loss.

We discuss each one below.

Internal versus External Operational Losses

Operational losses can be either internally inflicted or result from external sources. *Internally* inflicted sources include most of the losses caused by human, process, and technology failures, such as those due to human errors, internal fraud, unauthorized trading, injuries, business delays due to computer failures, or telecommunication problems. *External* sources include man-made incidents such as external fraud, theft, computer hacking, terrorist activities, and natural disasters such as damage to physical assets due to hurricanes, floods, and fires.

Many of the internal operational failures can be prevented with appropriate internal management practices; for example, tightened controls and management of the personnel can help prevent some employee errors and internal fraud, and improved telecommunication networks can help prevent some technological failures.

External losses are very difficult to prevent. However, it is possible to design insurance or other hedging strategies to reduce or possibly eliminate externally inflicted losses.

Direct versus Indirect Operational Losses

Direct losses are the losses that directly arise from the associated events. For example, an incompetent currency trading can result in a loss for the bank due to adverse exchange rate movements. As another example, mistakenly charging a client \$50,000 instead of \$150,000 results in the loss for the bank in the amount of \$100,000. The Basel II Capital Accord sets guidelines regarding the estimation of the regulatory capital charge by banks based only on direct losses. Table 1 identifies the Basel II Capital Accord's categories and definitions of direct operational losses.

Indirect losses are generally opportunity costs and the losses associated with the costs of fixing an operational risk problem, such as near-miss losses, latent losses, or contingent losses.

Near-Miss Operational Losses

Near-miss losses (or near-misses) are the estimated losses from those events that could potentially occur but were successfully prevented. The rationale behind including near-misses into internal databases is as follows: The definition

of "risk" should not be solely based on the past history of actual events but instead should be a forward-looking concept and include both actual and potential events that could result in material losses. The mere fact that a loss was prevented in the past (be it by luck or by conscious managerial action) does not guarantee that it will be prevented in the future. Therefore, near-misses signal flaws in a bank's internal system and should be accounted for in internal models. It is also possible to view near-misses from quite the opposite perspective: The ability to prevent these losses before they happen demonstrates the bank's effective operational risk management practices. Therefore, the losses that would result had these events taken place should not be included in the internal databases.

Muermann and Oktem (2002, p. 30) define near-miss as:

an event, a sequence of events, or an observation of unusual occurrences that possesses the potential of improving a system's operability by reducing the risk of upsets some of which could eventually cause serious damage.

They assert that internal operational risk measurement models must include adequate management of near-misses.

Muermann and Oktem propose developing a pyramid-type three-level structure for the near-miss management system: corporate level, branch level, and individual level. At the corporate level within every bank, they propose

Table 1 Direct Loss Types and Their Definitions According to the Basel II Capital Accord

Loss Type	Contents
Write-downs	Direct reduction in value of assets due to theft, fraud, unauthorized activity, or market and credit losses arising as a result of operational events
Loss of recourse Restitution	Payments or disbursements made to incorrect parties and not recovered Payments to clients of principal and/or interest by way of restitution, or the cost of any other form of compensation paid to clients
Legal liability	Judgements, settlements, and other legal costs
Regulatory and compliance	Taxation penalties, fines, or the direct cost of any other penalties, such as license revocations
Loss of or damage to assets	Direct reduction in value of physical assets, including certificates, due to an accident, such as neglect, fire, and earthquake

Source: BIS (2001a), p. 23, with modifications.

establishing a Near-Miss Management Strategic Committee whose primary functions would include:

- Establishing guidelines for corporate and site near-miss structures.
- Developing criteria for classification of near-misses.
- Establishing prioritizing procedures for each near-miss class.
- Auditing the near-miss system.
- Integrating quality and other management tools into near-miss management practice.
- Identifying gaps in the near-miss management structure based on analysis of incidents with higher damage (beyond near-misses) and taking corrective actions.
- Developing guidelines for training site management and employees on near-miss system.

At the branch level, they propose establishing a Near-Miss Management Council for every business unit. The key responsibilities of the council would include:

- Adapting criteria set by Near-Miss Management Strategic Committee to the branch practices.
- Monitoring site near-miss practices.
- Promoting the program.
- Ensuring availability of necessary resources for analysis and corrective action, especially for high priority near-misses.
- Periodically analyzing reported near-misses for further improvement of the system.
- Training employees on NM implementation.

Finally, a successful near-miss management system relies on the individual actions by managers, supervisors, and employees. Appropriate training is necessary to recognize operational issues before they become a major problem and develop into operational losses for the bank.

Expected versus Unexpected Operational Losses

Some operational losses are expected, some are not. The *expected* losses are generally those that occur on a regular (such as every day) basis, such as minor employee errors and minor credit card fraud. *Unexpected* losses are those losses that generally cannot be easily foreseen, such as terrorist attacks, natural disasters, and large-scale internal fraud.

Operational Risk Type, Event Type, and Loss Type

Confusion arises in the operational risk literature because of the distinction between risk type (or hazard type), event type, and loss type. When banks record their operational loss data, it is crucial to record it separately according to event type and loss type, and correctly identify the risk type.¹¹ The distinction between the three is comparable to cause and effect:¹²

- *Hazard* constitutes one or more factors that increase the probability of occurrence of an event.
- *Event* is a single incident that leads directly to one or more effects (e.g., losses).
- *Loss* constitutes the amount of financial damage resulting from an event.

Thus, hazard potentially leads to event, and event is the cause of loss. Therefore, an event is the effect of a hazard while loss is the effect of an event.

Figure 1 illustrates the mechanism of operational loss occurrence. The following example, adopted from Mori and Harada (2001), further illustrates how the correct identification of the “event type” is critical in determining whether a loss of a particular “loss type” is attributed to market, credit, or operational risk.

Consider the following example:

- A reduction in the value of a bond due to a change in the market price.

Operational Risk

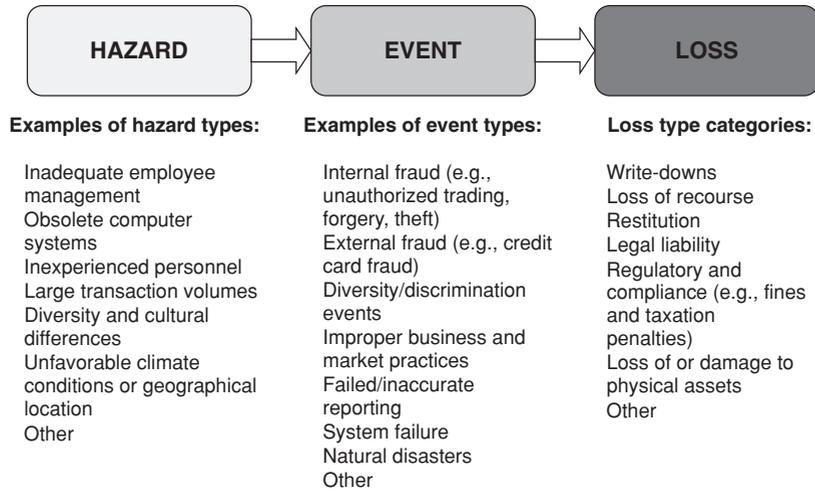


Figure 1 The Process of Operational Loss Occurrence
 Source: Mori and Harada (2001), p. 3, with modifications.

- A reduction in the value of a bond due to the bankruptcy of the issuer.
- A reduction in the value of a bond due to a delivery failure.

risk, credit risk, and operational risk, respectively. Accurate documentation of operational risk by the type of hazard, event, and loss is also essential for an understanding of operational risk.

In this example, the write-down of the bond (the loss type) belongs to the scope of market

The Basel II Capital Accord classifies operational risk into seven event-type groups (see

Table 2 Operational Risk Event Types and Their Descriptions According to the Basel II Capital Accord

Event Types and Descriptions According to Basel II	
Event Type	Definition and Categories
1. Internal Fraud	Acts intended to defraud, misappropriate property, or circumvent regulations, the law, or company policy, which involves at least one internal party. <i>Categories:</i> unauthorized activity and theft and fraud.
2. External Fraud	Acts of a type intended to defraud, misappropriate property, or circumvent the law, by a third party. <i>Categories:</i> (1) theft and fraud and (2) systems security.
3. Employment Practices and Workplace Safety	Acts inconsistent with employment, health, or safety laws or agreements, from payment of personal injury claims, or from diversity/discrimination events. <i>Categories:</i> (1) employee relations, (2) safe environment, and (3) diversity and discrimination.
4. Clients, Products, and Business Practices	Unintentional or negligent failure to meet a professional obligation to specific clients (including fiduciary and suitability requirements), or from the nature or design of a product. <i>Categories:</i> (1) suitability, disclosure, and fiduciary, (2) improper business or market practices, (3) product flaws, (4) selection, sponsorship, and exposure, and (5) advisory activities.
5. Damage to Physical Assets	Loss or damage to physical assets from natural disaster or other events. <i>Categories:</i> disasters and other events.
6. Business Disruption and System Failures	Disruption of business or system failures. <i>Categories:</i> systems.
7. Execution, Delivery, and Process Management	Failed transaction processing or process management, from relations with trade counterparties and vendors. <i>Categories:</i> (1) transaction capture, execution, and maintenance, (2) monitoring and reporting, (3) customer intake and documentation, (4) customer/client account management, (5) trade counterparties, and (6) vendors and suppliers.

Source: BIS (2001b), pp. 21–23.

Table 2) and six operational loss types (see Table 1).

Operational Loss Severity and Frequency

We have already stated that expected losses generally refer to the losses of low severity (or magnitude) and high frequency. Generalizing this idea, operational losses can be broadly classified into four main groups:

1. Low frequency/low severity.
2. High frequency/low severity.
3. High frequency/high severity.
4. Low frequency/high severity.

The idea is illustrated in the top half of Figure 2.

According to Samad-Khan (2005), the third group is implausible. More precisely, he suggests classifying each of the frequency and

severity of operational losses into three groups: low, medium, and high. This creates a 3 × 3 matrix of all possible “frequency/severity” combinations. He states that “medium frequency/high severity,” “high frequency/medium severity,” and “high frequency/high severity” losses are unrealistic.

Recently, the financial industry also agreed that the first group is not feasible. Therefore, the two remaining categories of operational losses that the financial industry needs to focus on are “high frequency/low severity” and “low severity/high frequency” losses. The idea is illustrated in the bottom half of Figure 2.

The losses of “high frequency/low severity” are relatively unimportant for an institution and can often be prevented. What poses the greatest damage is the “low frequency/high severity” losses. Banks must be particularly attentive to these losses as these cause the greatest financial consequences to the institution, including

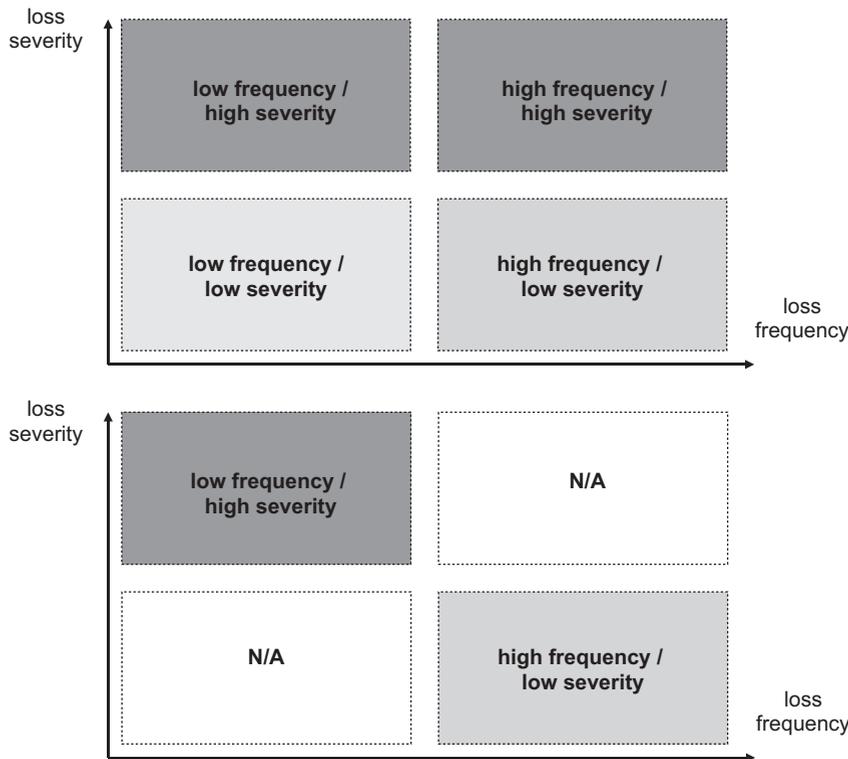


Figure 2 Classification of Operational Risk by Frequency and Severity: Unrealistic View (top) and Realistic View (bottom)

potential bankruptcy.¹³ Just a few of such events may result in bankruptcy or a significant decline in the market value of the bank. Therefore, it is critical for banks to be able to capture such losses in their internal risk models.

KEY POINTS

- Financial institutions bear various operational losses on a daily basis. Examples are losses resulting from employee errors, internal and external fraud, equipment failures, business disruptions due to natural disasters, and vandalism.
- Credit risk and market risk had been perceived as the two biggest sources of risk for financial institutions. Operational risk has been regarded as a mere part of “other” risks. Failures of major financial entities have made market participants aware of the importance of this risk.
- Operation risk is the risk of loss resulting from inadequate or failed internal processes, people, or systems or from external events. This definition identifies operational risk as coming from four major causes: processes, human, systems, and external factors.
- Operational risk can be classified according to several principles: nature of the loss (internally inflicted or externally inflicted), direct losses or indirect losses, degree of expectancy (expected or unexpected), risk type, event type or loss type, and by the magnitude (or severity) of loss and the frequency of loss.
- Operational risk can be the cause of reputational risk, a risk that can occur when the market reaction to an operational loss event results in reduction in the market value of a bank that is greater than the amount of the initial loss.

NOTES

1. However, operational risk is not entirely idiosyncratic. Two recent studies—Allen and Bali (2007) and Chernobai, Jorion, and Yu (2011)—found evidence of the effect of

macroeconomic factors on operational risk in banks.

2. See BIS (1998).
3. See Jorion (2000).
4. See King (2001).
5. See Crouhy, Galai, and Mark (2001).
6. See Barclays Bank Annual Report 2004, Form 20-F/A.
7. Deutsche Bank 2005 Annual Report, p. 45.
8. Bank of Tokyo-Mitsubishi Financial Performance, Form 20-F (2005), p. 124.
9. “Supervised Investment Bank Holding Companies,” SEC (2003), p. 62914.
10. Examples of operational risk exposure indicators are given in BIS (2001a, Annex 4), Haubenstock (2003), and Allen, Boudoukh, and Saunders (2004).
11. See the discussion of this issue in Mori and Harada (2001) and Alvarez (2002).
12. See Mori and Harada (2001).
13. The events that incur such losses are often called the “tail events.”

REFERENCES

- Allen, L., and Bali, T. (2007). Cyclicity in catastrophic and operational risk measurements, *Journal of Banking and Finance* 31, 4: 1191–1235.
- Allen, L., J. Boudoukh, and Saunders, A. (2004). *Understanding Market, Credit, and Operational Risk: The Value-at-Risk Approach*. Oxford: Blackwell Publishing.
- Alvarez, G. (2002). Operational risk event classification. <http://www.garp.com>.
- Bank for International Settlements (1998). Overview of the amendment to the capital accord to incorporate market risks. <http://www.bis.org>.
- Bank for International Settlements (2001a). Consultative document: Operational risk. <http://www.bis.org>.
- Bank for International Settlements (2001b). Working paper on the regulatory treatment of operational risk. <http://www.bis.org>.
- Chapelle, A., Crama, Y., Hübner, G., and Peters, J. (2005). Measuring and managing operational risk in the financial sector: An integrated framework. Technical report, National Bank of Belgium.

- Chernobai, A., Jorion, P., and Yu, F. (2011). The determinants of operational risk in U.S. financial institutions. *Journal of Financial and Quantitative Analysis* 46, 6: 1683–1725.
- Crouhy, M., Galai, D., and Mark, R. (2001). *Risk Management*. New York: McGraw-Hill.
- Haubenstock, M. (2003). The operational risk management framework. In C. Alexander (ed.), *Operational Risk: Regulation, Analysis, and Management*. Great Britain: Prentice Hall.
- Jorion, P. (2000). *Value-at-Risk: The New Benchmark for Managing Financial Risk*, 2nd ed. New York: McGraw-Hill.
- King, J. L. (2001). *Operational Risk: Measurement and Modelling*. Hoboken, NJ: John Wiley & Sons.
- Mori, T., and Harada, E. (2001). Internal measurement approach to operational risk capital charge. Technical report, Bank of Japan.
- Muermann, A., and Oktem, U. (2002). The near-miss management of operational risk. *The Journal of Risk Finance* 4, 1: 25–36.
- Samad-Khan, A. (2005). Why COSO is flawed. *Operational Risk*, January, 1–6.
- Shih, J., Samad-Khan, A. J., and Medapa, P. (2000). Is the size of an operational risk related to firm size? *Operational Risk*, January.

Operational Risk Models

ANNA CHERNOBAI, PhD

Assistant Professor of Finance, M. J. Whitman School of Management, Syracuse University

SVETLOZAR T. RACHEV, PhD, DrSci

Frey Family Foundation Chair Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: In general terms, operational risk is the risk of loss resulting from inadequate or failed internal processes, people, or systems or from external events. The models that have been proposed for assessing operational risk can be broadly classified into top-down models and bottom-up models. Top-down approaches quantify operational risk without attempting to identify the events or causes of losses. Bottom-up models quantify operational risk on a micro level being based on identified internal events. The obstacle hindering the implementation of these models is the scarcity of available historical operational loss data.

Identifying the core principles that underlie the *operational risk* process is the fundamental building block in deciding on the optimal model to be used. In this entry we provide an overview of models that have been put forward for the assessment of operational risk. These models are broadly classified into *top-down models* and *bottom-up models*.

Operational risk is distinct from credit risk and market risk, posing difficulties of implementation of the Basel II guidelines and strategic planning. We discuss some key aspects that distinguish operational risk from credit risk and market risk. They are related to the *arrival process* of loss events, the *loss severity*, and the dependence structure of operational losses across

a bank's business units. Finally in this entry we reconsider the normality assumption—an assumption often made in modeling financial data—and question its applicability for the purpose of operational risk modeling.

OPERATIONAL RISK MODELS

Broadly speaking, operational risk models stem from two fundamentally different approaches: (1) the top-down approach, and (2) the bottom-up approach. Figure 1 illustrates a possible categorization of quantitative models.

Top-down approaches quantify operational risk without attempting to identify the events

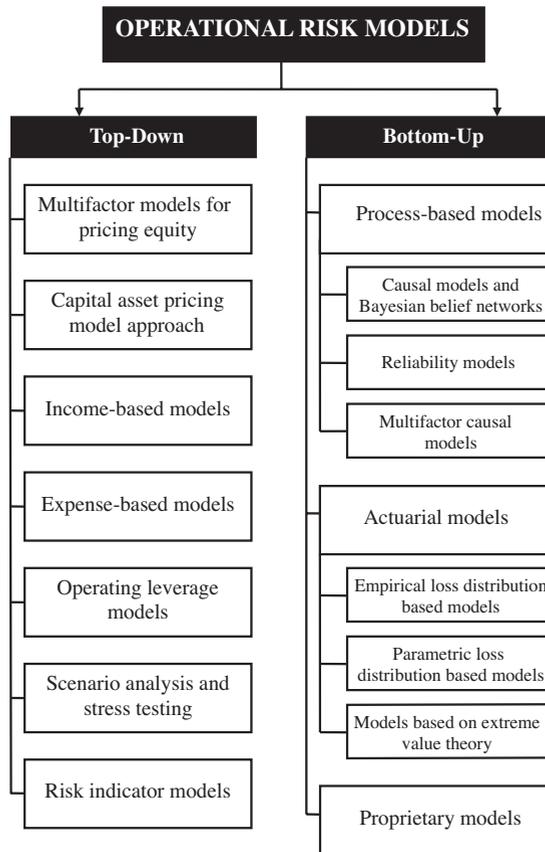


Figure 1 Topology of Operational Risk Models

or causes of losses.¹ That is, the losses are simply measured on a macro basis. The principal advantage of this approach is that little effort is required with collecting data and evaluating operational risk. Bottom-up approaches quantify operational risk on a micro level being based on identified internal events, and this information is then incorporated into the overall capital charge calculation. The advantage of bottom-up approaches over top-down approaches lies in their ability to explain the mechanism of how and why operational risk is formed within an institution. Banks can either start with top-down models and use them as a temporary tool to estimate the capital charge and then slowly shift to the more advanced bottom-up models, or they can adopt bottom-up models from the start, provided that they have robust databases.

Models Based on Top-Down Approaches

In this section we will provide a brief look at the seven top-down approaches shown in Figure 1.²

Multifactor Equity Pricing Models

Multifactor equity pricing models, also referred to as multifactor models, can be utilized to perform a global analysis of banking risks and may be used for the purpose of integrated risk management, in particular for publicly traded firms. The stock return process R_t can be estimated by regressing stock return on a large number of external risk factor indexes I_t related to market risk, credit risks, and other nonoperational risks (such as interest rate fluctuations, stock price movements, and macroeconomic effects). Operational risk is then measured as the volatility of the residual term. Such models rely on the assumption that operational risk is the residual banking risk, after credit and market risks are accounted for.³

$$R_t = a_t + b_1 I_{1t} + \dots + b_n I_{nt} + \varepsilon_t$$

in which ε_t is the residual term, a proxy for operational risk.

This approach relies on the widely known efficient market hypothesis that was introduced by Fama (1970), that states that in efficient capital markets all relevant past, publicly, and privately available information is reflected in current asset prices.

Capital Asset Pricing Model

Under the capital asset pricing model (CAPM) approach all risks are assumed to be measurable by the CAPM and represented by beta (β). CAPM, developed by Sharpe (1964), is an equilibrium model that describes the pricing of assets. It concludes that the expected security risk premium (i.e., expected return on security minus the risk-free rate of return) equals beta times the expected market risk premium

(i.e., expected return on the market minus the risk-free rate of return).

Under the CAPM approach, operational risk is obtained by measuring market, credit, and other risks' betas and deducting them from the total beta. With respect to applications to operational risk, the CAPM approach was discussed by Hiwatashi and Ashida (2002) and van den Brink (2002). According to van den Brink (2002), the CAPM approach has some limitations and so has not received a wide recognition for operational risk, but was in the past considered by Chase Manhattan Bank.

Income-Based Models

Income-based models resemble the multifactor equity price models: Operational risk is estimated as the residual variance by extracting market, credit, and other risks from the historical income (or earnings) volatility. Income-based models are described by Allen, Boudoukh, and Saunders (2004), who refer to these models as earnings at risk models and by Hiwatashi and Ashida (2002), who refer to them as the volatility approach. According to Cruz (2002), the profit and loss (P&L) volatility in a financial institution is attributed 50%, 15%, and 35% to credit risk, market risk, and operational and other risks, respectively.

Expense-Based Models

Expense-based models measure operational risk as fluctuations in historical expenses rather than income. The unexpected operational losses are captured by the volatility of direct expenses (as opposed to indirect expenses, such as opportunity costs, reputational risk, and strategic risk, that are outside the agreed scope of operational risk), adjusted for any structural changes within the bank.

Operating Leverage Models

Operating leverage models measure the relationship between operating expenses and total assets. Operating leverage is measured as a weighted combination of a fraction of fixed as-

sets and a portion of operating expenses. Examples of calculating operating leverage amount per business line include taking 10% of fixed assets plus 25% times three months' operating expenses for a particular business, or taking 2.5 times the monthly fixed expenses.⁴

Scenario Analysis and Stress Testing Models

Scenario analysis and stress testing models can be used for testing the robustness properties of loss models, in monetary terms, in the presence of potential events that are not part of banks' actual internal databases. These models, also called expert judgment models by van den Brink (2002), are estimated based on the "what if" scenarios generated with reference to expert opinion, external data, catastrophic events that occurred in other banks, or imaginary high-magnitude events. Experts estimate the expected risk amounts and their associated probabilities of occurrence. For any particular bank, examples of scenarios include:⁵

- Bank's inability to reconcile a new settlement system with the original system.
- A class action suit alleging incomplete disclosure.
- Massive technology failure.
- High-scale unauthorized trading (for example, adding the total loss borne by the Barings bank preceding its collapse into the database, and reevaluating the model).
- Doubling the bank's maximum historical loss amount.

Additionally, stress tests can be used to see the likely increase in risk exposure due to removing a control or reduction in risk exposure due to tightening of controls.

Risk Indicator Models

Risk indicator models rely on a number (one or more) of operational risk exposure indicators to track operational risk. In the operational risk literature, risk indicator models are also called indicator approach models,⁶ risk profiling

models,⁷ and peer-group comparison.⁸ A necessary aspect of such models is testing for possible correlations between risk factors. These models assume that there is a direct and significant relationship between the indicators and target variables. For example, Taylor and Hoffman (1999) illustrate how training expenditure has a reverse effect on the number of employee errors and customer complaints and Shih, Samad-Khan, and Medapa (2000) illustrate how a bank's size relates to the operational loss amount.

Risk indicator models may rely on a single indicator or multiple indicators. The former model is called the single-indicator approach;⁹ an example of such a model is the Basic Indicator Approach for quantification of the operational risk regulatory capital, proposed by the Basel II. The latter model is called the multi-indicator approach; an example of such a model is the Standardized Approach.

Models Based on Bottom-Up Approaches

An ideal internal operational risk assessment procedure would be to use a balanced approach, and include both top-down and bottom-up elements in the analysis.¹⁰ For example, scenario analysis can prove effective for backtesting purposes, and multifactor causal models are useful in performing operational Value-at-Risk (VaR) sensitivity analysis. Bottom-up approach models can be categorized into three groups:¹¹ process-based models, actuarial-type models (or statistical models), and proprietary models.

Process-Based Models

There are three types of process-based models: (1) causal models and Bayesian belief networks, (2) reliability models, and (3) multifactor causal models. We describe each below.

The first group of process-based models is the causal models and Bayesian belief networks. Also called causal network models, causal

models are subjective self-assessment models. Causal models form the basis of the scorecard models.¹² These models split banking activities into simple steps; for each step, bank management evaluates the number of days needed to complete the step, the number of failures and errors, and so on, and then records the results in a "process map" (or scorecards) in order to identify potential weak points in the operational cycle. Constructing associated event trees that detect a sequence of actions or events that may lead to an operational loss is part of the analysis.¹³ For each step, bank management estimates a probability of its occurrence, called the subjective (or prior) probability. The ultimate event's probability is measured by the posterior probability. Prior and posterior probabilities can be estimated using the Bayesian belief networks.¹⁴ A variation of the causal models, connectivity models, focuses on the ex ante cause of operational loss event, rather than the ex post effect.

The second group of process-based models encompasses reliability models. These models are based on the frequency distribution of the operational loss events and their interarrival times. Reliability models focus on measuring the likelihood that a particular event will occur at some point or interval of time. We discuss this model below.

If $f(t)$ is the density of a loss amount occurring at time t , then the reliability of the system is the probability of survival up to time t , denoted by $R(t)$ and calculated as

$$R(t) = 1 - \int_0^t f(s)ds$$

The *hazard rate* (or the *failure rate*), $h(t)$, is the rate at which losses occur per unit of time t , defined as

$$h(t) = \frac{f(t)}{R(t)}$$

In practical applications, it is often convenient to use the Poisson-type arrival model to describe the occurrence of operational loss events. Under the simple Poisson model with the

intensity rate λ (which represents the average number of events in any point of time), the inter-arrival times between the events (i.e., the time intervals between any two consecutive points of time in which an event takes place) follow an exponential distribution having density of form $f(t) = \lambda e^{-\lambda t}$ with mean interarrival time equal to $1/\lambda$. The parameter λ is then the hazard rate for the simple Poisson process.

Finally, the third group of process-based models is multifactor causal models. These models can be used for performing the factor analysis of operational risk. These are regression-type models that examine the sensitivity of aggregate operational losses (or, alternatively, VaR) to various internal risk factors (or risk drivers). Multifactor causal models have been discussed in the VaR and operational risk literature.¹⁵ Examples of control factors include system downtime in minutes per day, number of employees in the back office, data quality (such as the ratio of the number of transactions with no input errors to the total number of transactions), total number of transactions, skill levels, product complexity, level of automation, customer satisfaction, and so on. Cruz (2002) suggests using manageable explanatory factors. In multifactor causal models, operational losses OR_t , or VaR, in a particular business unit at a point t , are regressed on a number of control factors:

$$OR_t = a_t + b_1 X_{1t} + \dots + b_n X_{nt} + \varepsilon_t$$

where X_k , $k = 1, 2, \dots, n$, are the explanatory variables, and b 's are the estimated coefficients. The model is forward-looking (or ex ante) as operational risk drivers are predictive of future losses. Extensions to the simple regression model may include autoregressive models, regime-switching models, ARMA/GARCH models, and others.

Actuarial Models

Actuarial models (or statistical models) are generally parametric statistical models. They have two key components: (1) the loss frequency and

(2) the loss severity distributions of the historic operational loss data. Operational risk capital is measured by the VaR of the aggregated one-year losses.¹⁶

For the frequency of the loss data it is common to assume a Poisson process, with possible generalizations, such as a Cox process.

Actuarial models can differ by the type of the loss distribution. Empirical loss distribution models do not specify a particular class of loss distributions, but directly utilize the empirical distribution derived from the historic data. Parametric loss distribution models make use of a particular parametric distribution for the losses (or part of them), such as lognormal, Weibull, Pareto, and so on. Models based on extreme value theory (EVT) restrict attention to the tail events (i.e., the losses in the upper quantiles of the severity distribution), and VaR or other analyses are carried out upon fitting the generalized Pareto distribution to the data beyond a fixed high threshold. Van den Brink suggests using all three models simultaneously; Figure 2, inspired by his discussions, illustrates possible approaches. Yet another possibility is to fit an ARMA/GARCH model to the losses below a high threshold and the generalized Pareto distribution to the data exceeding it.

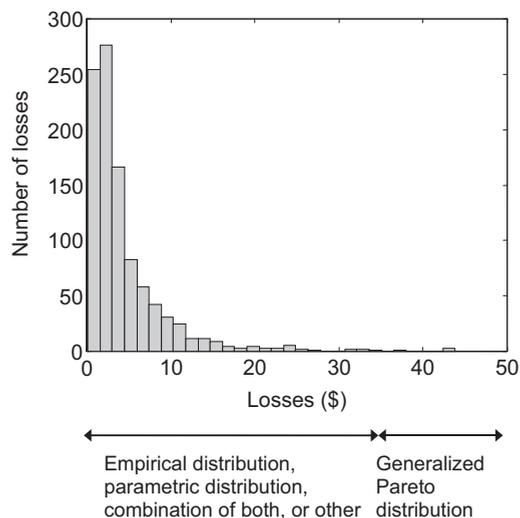


Figure 2 An Example of a Histogram of the Operational Loss Severity Distribution

Proprietary Models

Proprietary models for operational risk have been developed by major financial service companies and use a variety of bottom-up and top-down quantitative methodologies, as well as qualitative analysis, to evaluate operational risk. Banks can input their loss data into ready and systematized spreadsheets, which would be further categorized. The system then performs a qualitative and quantitative analysis of the data, and can carry out multiple tasks such as calculating regulatory capital, pooling internal data with external, performing Bayesian network analysis, and so on.

SPECIFICS OF OPERATIONAL LOSS DATA

The nature of operational risk is very different from that of market risk and credit risk. In fact, operational losses share many similarities with insurance claims, suggesting that most actuarial models can be a natural choice of the model for operational risk, and models well developed by the insurance industry can be almost exactly applied to operational risk. In this section we discuss some key issues characterizing operational risk that must be taken into consideration before quantitative analysis is undertaken.

Scarcity of Available Historical Data

The major obstacle banks face in developing comprehensive models for operational risk is the scarcity of available historical operational loss data. As of 2011, generally, even the largest banks have no more than 11–12 years of loss data. Shortage of relevant data means that the models and conclusions drawn from the available limited samples would lack sufficient explanatory power. This in turn means that the estimates of the expected loss and VaR may be highly volatile and unreliable. In addition, complex statistical or econometric models cannot be tested on small samples.

The problem becomes amplified when dealing with modeling extremely high operational losses: One cannot model tail events when only a few such data are present in the internal loss database. Three solutions have been proposed: (1) pooling internal and external data, (2) supplementing actual losses with near-miss losses, and (3) scenario analysis and stress tests (discussed earlier in this entry).

The idea behind pooling internal and external data is to populate a bank's existing internal database with data from outside the bank. The rationale is twofold: (1) to expand the database and hence increase the accuracy of statistical estimations and (2) to account for losses that have not occurred within the bank but that are not completely improbable based on the histories of other banks. According to BIS,

... a bank's internal measurement system must reasonably estimate unexpected losses based on the combined use of internal and relevant external loss data... (BIS, 2006, p. 150)

Baud, Frachot, and Roncalli (2002) propose a statistical methodology to pool internal and external data. Their methodology accounts for the fact that external data are truncated from below (banks commonly report their loss data to external parties in excess of \$1 million) and that bank size may be correlated with the magnitudes of losses. They showed that pooling internal and external data may help avoid underestimation of the capital charge.

Data Arrival Process

One of the difficulties that arise with modeling operational losses has to do with the irregular nature of the event arrival process. In market risk models, market positions are recorded on a frequent basis, many times daily depending on the entity, by marking to market. Price quotes are available daily or for those securities that are infrequently traded, model-based prices are available for marking a position to market. As for credit risk, credit ratings by rating agencies are available. In addition, rating agencies

provide credit watches to identify credits that are candidates for downgrades. In contrast, operational losses occur at irregular time intervals suggesting a process of a discrete nature. This makes it similar to the reduced-form models for credit risk, in which the frequency of default (i.e., failure to meet a credit agreement) is of nontrivial concern. Hence, while in market risk we need to model only the return distribution in order to obtain VaR, in operational risk both loss severity and frequency distributions are important.

Another problem is related to timing and data recording issues. In market and credit risk models, the impact of a relevant event is almost immediately reflected in the market and credit returns. In an ideal scenario, banks would know how much of the operational loss would be borne by the bank from an event at the very moment the event takes place and would record the loss at this moment. However, from the practical point of view, this appears nearly impossible to implement, because it takes time for the losses to accumulate after an event takes place. Therefore, it may take days, months, or even years for the full impact of a particular loss event to be evaluated. Hence, there is the problem of discrepancy (i.e., a time lag) between the occurrence of an event and the time at which the incurred loss is being recorded.

This problem directly affects the method in which banks choose to record their operational loss data. When banks record their operational loss data, they record (1) the amount of loss, and (2) the corresponding date. We can identify three potential scenarios for the types of date banks might use:¹⁷

1. *Date of occurrence*: the date on which the event that has led to operational losses actually took place.
2. *Date on which the existence of event has been identified*: the date when bank authorities realize that an event that has led to operational losses has taken or is continuing to take place. Recording a loss at this date may be relevant

in cases when the true date of occurrence is impossible or hard to track.

3. *Accounting date*: the date on which the total amount of operational losses due to a past event are realized and fully measured, and the state of affairs of the event is closed or assumed closed.

Depending on which of the three date types is used, the models for operational risk and conclusions drawn from them may be considerably different. For example, in the third case of accounting dates, we are likely to observe cyclical/seasonal effects in the time series of the loss data (for example, many loss events would be recorded around the end of December), while in the first and second cases such effects are much less likely to be present in the data. Fortunately, however, selection of the frequency distribution does not have a serious impact on the resulting capital charge.¹⁸

Loss Severity Process

There are three main problems that operational risk analysts must be aware of with respect to the severity of operational loss data: (1) the non-negative sign of the data, (2) the high degree of dispersion of the data, and (3) the shape of the data.

The first problem related to the loss severity data deals with the sign of the data. Depending on the movements in the interest or exchange rates, the oscillations in the market returns and indicators can take either a positive or negative sign. This is different in the credit and operational risk models—usually, only losses (i.e., negative cash flows) are assumed to take place.¹⁹ Hence, in modeling operational loss magnitudes, one should either consider fitting the loss distributions that are defined only on positive values, or should use distributions that are defined on negative and positive values, truncated at zero.

The second problem deals with the high degree of dispersion of loss data. Historical

observations suggest that the movements in the market indicators are generally of relatively low magnitude. Bigger losses are usually attributed to credit risk. Finally, although most of the operational losses occur on a daily basis and hence are small in magnitude, the excessive losses of financial institutions are in general due to the operational losses, rather than credit or market risk-related losses. Empirical evidence indicates that there is an extremely high degree of dispersion of the operational loss magnitudes, ranging from near-zero to billions of dollars. In general, this dispersion is measured by variance or standard deviation.²⁰

The third problem concerns the shape of the loss distribution. The shape of the data for operational risk is very different from that of market or credit risk. In market risk models, for example, the distribution of the market returns is often assumed to be nearly symmetric around zero. Asymmetric cases refer to the data whose distribution is either left-skewed (i.e., the left tail of the distribution is very long) or right-skewed (i.e., the right tail of the distribution is very long) and/or whose distribution has two or more peaks of different height. Operational losses are highly asymmetric, and empirical evidence on operational risk indicates that the losses are highly skewed to the right. This is in part explained by the presence of “low frequency/high severity” events. See Figure 2 for an exemplary histogram of operational losses.

As previously discussed, empirical evidence on operational losses indicates a majority of observations being located close to zero, and a small number of observations being of a very high magnitude. The first phenomenon refers to a high kurtosis (i.e., peak) of the data, and the second one indicates heavy tails (or fat tails). Distributions of such data are often described as leptokurtic.

The Gaussian (or normal) distribution is often used to model market risk and credit risk. It is characterized by two parameters, μ and σ , that are its mean and standard deviation. Figure 3 provides an example of a normal density.

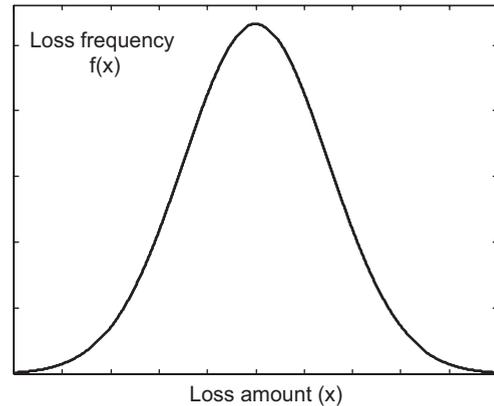


Figure 3 An Example of a Gaussian Density

Despite being easy to work with and having attractive features (such as symmetry and stability under linear transformations), the Gaussian distribution makes several critical assumptions about the loss data. They include the following:

- The Gaussian assumption is useful for modeling the distribution of events that are symmetric around their mean. It has been empirically demonstrated that operational losses are not symmetric and severely right-skewed, meaning that the right tail of the loss distribution is very long.
- In most cases (except for the cases when the mean is very high), the use of Gaussian distribution allows for the occurrence of negative values. This is not a desirable property for modeling loss severity because negative losses are usually not possible.²¹
- More importantly, the Gaussian distribution has an exponential decay in its tails (this property puts the Gaussian distribution into the class of light-tailed distributions), which means that the tail events (i.e., the events of an unusually high or low magnitude) have a near-zero probability of occurrence. However, very high-magnitude operational losses can seriously jeopardize a financial institution. Thus, it would be inappropriate to model operational losses with a distribution that essentially excludes the possibility of high-impact individual losses. Empirical evidence

strongly supports the conjecture that the distribution of operational losses is in fact very leptokurtic—that is, has a high peak and very heavy tails (i.e., very rare events are assigned a positive probability).

For the reasons presented above, it is unlikely that the Gaussian distribution would find much application for the assessment of operational risk.²² Heavier tailed distributions such as log-normal, Weibull, and even Pareto and alpha-stable, ought to be considered.

Dependence Between Business Units

In order to increase the accuracy of operational risk assessment, banks are advised to classify their operational loss data into groups of different degrees and nature of exposure to operational risk. Following this principle, the advanced measurement approaches (AMA) for the quantification of the operational risk capital charge, proposed by Basel II, suggest estimating operational risk capital separately for each “business line/event type” combination. Such a procedure is not common in market risk and credit risk models.

The most intuitive approach to combine risk measures collected from each of these “business line/event type” combinations is to add them up.²³ However, such an approach may result in overestimation of the total capital charge because it implies a perfect positive correlation between groups. To prevent this from happening, it is essential to account for dependence between these combinations. Covariance and correlation are the simplest measures of dependency, but they assume a linear type of dependence, and therefore can produce misleading results if the linearity assumption is not true. An alternative approach would involve using copulas that are more flexible with respect to the form of the dependence structure that may exist between different groups. Another attractive property of copulas is their ability to cap-

ture the tail dependence between the distributions of random variables. Both properties are preserved under linear transformations of the variables.

KEY POINTS

- Operational risk measurement models are divided into top-down and bottom-up models.
- Top-down models use a macro-level regulatory approach to assess operational risk and determine the capital charge. They include multifactor equity price models, income and expense-based models, operating leverage models, scenario analysis and stress testing models, and risk indicator models.
- Bottom-up models originate from a micro-level analysis of a bank’s loss data and consideration for the process and causes of loss events in determination of the capital charge. They include process-based models (such as causal network and Bayesian belief models, connectivity models, multifactor causal models, and reliability models), actuarial models, and proprietary models.
- Scarcity and reliability of available internal operational loss data remains a barrier preventing banks from developing comprehensive statistical models. Sufficiently large datasets are especially important for modeling low frequency high severity events. Three solutions have been put forward to help expand internal databases: pooling together internal and external data, accounting for near-misses, and stress tests.
- The nature of operational risk is fundamentally different from that of credit and market risks. Specifics of operational loss process include discrete data arrival process, delays between time of event and loss detection/accumulation, loss data taking only positive sign, high dispersion in magnitudes of loss data, distribution of loss data being severely right-skewed and heavy-tailed, and dependence between business units and event types.

- While many market and credit risk models make the convenient Gaussian assumption on the market returns or stock returns, this distribution is unlikely to be useful for the operational risk modeling because it is unable to capture the nonsymmetric and heavy-tailed nature of the loss data.

NOTES

1. An exception is the scenario analysis models in which specific events are identified and included in internal databases for stress testing. These events are, however, imaginable and do not appear in the banks' original databases.
2. Some of these models are described in Allen, Boudoukh, and Saunders (2004).
3. See Chapter 2 in Chernobai, Rachev, and Fabozzi (2007) for an example of an empirical study that utilized such models in order to evaluate the sensitivity of operational risk to macroeconomic factors.
4. See Marshall (2001).
5. The first four examples are due to Marshall (2001).
6. See Hiwatashi and Ashida (2002).
7. See Allen, Boudoukh, and Saunders (2004).
8. See van den Brink (2002).
9. See van den Brink (2002).
10. The Internal Measurement Approach (see description in BIS, 2001) combines some elements of the top-down approach and bottom-up approach: The gamma parameter in the formula for the capital charge is set externally by regulators, while the expected loss is determined based on internal data.
11. See Allen, Boudoukh, and Saunders (2004).
12. In February 2001 the Basel Committee suggested the Scorecard Approach as one possible advanced measurement approach to measure the operational risk capital charge.
13. See, for example, Marshall (2001) on the "fishbone analysis."
14. Relevant Bayesian belief models with applications to operational risk are discussed in Alexander and Pezier (2001), Neil and Tranham (2002), and Giudici (2004), among others.
15. See also Haubenstock (2003) and Cruz (2002). The empirical study by Allen and Bali (2007) investigates the sensitivity of operational VaR to macroeconomic, rather than a bank's internal, risk factors.
16. Actuarial models form the basis of the loss distribution approach, an advanced measurement approach for operational risk. See BIS (2001).
17. Identification of the three types of dates are based on discussions with Marco Moscadelli (Banking Supervision Department, Bank of Italy).
18. See Carillo Menéndez (2005).
19. Certainly, it is possible that an event due to operational risk can incur unexpected profits for a bank, but usually this possibility is not considered.
20. Some very heavy-tailed distributions, such as the heavy-tailed Weibull, Pareto, or alpha-stable, can have an infinite variance. In these situations, robust measures of spread must be used.
21. Certainly, it is possible to use a truncated (at zero) version of the Gaussian distribution to fit operational losses.
22. Of course, a special case is fitting the Gaussian distribution to the natural logarithm of the loss data. This is equivalent (in terms of obtaining the maximum likelihood parameter estimates) to fitting the lognormal distribution to the original loss data.
23. This is the approach that was proposed in BIS (2001).

REFERENCES

- Alexander, C. and Pezier, J. (2001). Taking control of operational risk. *Futures and Options World* 366: 60–65.
- Allen, L. and Bali, T. (2007). Cyclicalities in catastrophic and operational risk measurements. *Journal of Banking and Finance* 31, 4: 1191–1235.

- Allen, L., Boudoukh, J., and Saunders, A. (2004). *Understanding Market, Credit, and Operational Risk: The Value-at-Risk Approach*. Oxford: Blackwell Publishing.
- Baud, N., Frachot, A., and Roncalli, T. (2002). Internal data, external data, and consortium data for operational risk measurement: How to pool data properly? Technical report, Groupe de Recherche Opérationnelle, Crédit Lyonnais, France.
- Bank for International Settlements (2001). Consultative document: Operational risk. <http://www.bis.org>.
- Bank for International Settlements (2006). International convergence of capital measurement and capital standards. <http://www.bis.org>.
- Carillo Menéndez, S. (2005). Operational risk. Presentation at International Summer School on Risk Measurement and Control, Rome, June 2005.
- Chernobai, A., Rachev, S. T., and Fabozzi, F. J. (2007). *Operational Risk: A Guide to Basel II Capital Requirements, Models, and Analysis*. Hoboken, NJ: John Wiley & Sons.
- Cruz, M. G. (2002). *Modeling, Measuring, and Hedging Operational Risk*. Chichester, NY: John Wiley & Sons.
- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance* 25: 383–417.
- Giudici, P. (2004). Integration of qualitative and quantitative operational risk data: A Bayesian approach. In M. G. Cruz (ed.), *Operational Risk Modelling and Analysis. Theory and Practice*. London: RISK Books.
- Haubenstock, M. (2003). The operational risk management framework. In C. Alexander (ed.), *Operational Risk: Regulation, Analysis, and Management*. London: Prentice Hall.
- Hiwatashi, J., and Ashida, H. (2002). Advancing operational risk management using Japanese banking experiences. Technical report, Federal Reserve Bank of Chicago.
- Marshall, C. L. (2001). *Measuring and Managing Operational Risk in Financial Institutions: Tools, Techniques, and Other Resources*. Chichester, NY: John Wiley & Sons.
- Neil, M., and Tranham, E. (2002). Using Bayesian networks to predict op risk. *Operational Risk*, August: 8–9.
- Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance* 19: 425–442.
- Shih, J., Samad-Khan, A. J., and Medapa, P. (2000). Is the size of an operational risk related to firm size? *Operational Risk*, January.
- Taylor, D., and Hoffman, D. (1999). How to avoid signal failure. *Risk*, November: 13–15.
- van den Brink, J. (2002). *Operational Risk: The New Challenge for Banks*. New York: Palgrave.

Modeling Operational Loss Distributions

ANNA CHERNOBAI, PhD

Assistant Professor of Finance, M. J. Whitman School of Management, Syracuse University

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: A major risk faced by financial entities is operational risk. In general terms, operational risk is the risk of loss resulting from inadequate or failed internal processes, people, or systems or from external events. The two principal approaches in modeling operational loss distributions are the nonparametric approach and the parametric approach. It is important to employ a model that captures tail events and for this reason in operational risk modeling, distributions that are characterized as light-tailed distributions should be used with caution.

For financial entities, representing a stream of uncertain *operational losses* with a specified model is a difficult task: Data can be wrongly recorded, fuzzy, incomplete (e.g., truncated or censored), or simply limited. Two main approaches may be undertaken: nonparametric and parametric. In this entry, we focus on the nonparametric approach, common loss distributions, and mixture distributions. We begin by reviewing the nonparametric approach to modeling operational losses and then proceed to the parametric approach and review some common continuous distributions that can be relevant for modeling operational losses. For each of the distributions, we focus on its major characteristics that are important when using

them to model the operational loss data: density, distribution, tail behavior, mean, variance, mode, skewness, and kurtosis.

APPROACHES TO OPERATIONAL RISK MODELING

The two main approaches to *operational risk* modeling are:

1. *Nonparametric approach.* One approach would be to directly use the empirical density of the data or its smoothed curve version.¹ This nonparametric approach can be relevant in

two circumstances: first, when the available data are not believed to follow any conventional distribution,² and second, when the data set available at hand is believed to be sufficiently comprehensive.³

2. *Parametric approach.* The task is considerably simplified if we are able to fit a curve of a simple analytical form that satisfies certain properties. The general goal of this parametric approach is to find a loss distribution that would most closely resemble the distribution of the loss magnitudes of the available data sample.

Figure 1 shows a common histogram for the operational loss data with a fitted continuous curve. Visual examination suggests that magnitudes of the majority of the losses are very close to zero as is seen from the high peak around zero of the histogram; an insignificant fraction of data account for the long right tail of the histogram. Clearly, if we choose the parametric approach and if the fitted curve represents a density of some chosen parametric distribution, the loss distributions that would be adequate for modeling operational losses are those that are right-skewed, possibly leptokurtic, and have support on the positive values.

Figure 2 summarizes possible approaches to modeling operational loss severity.

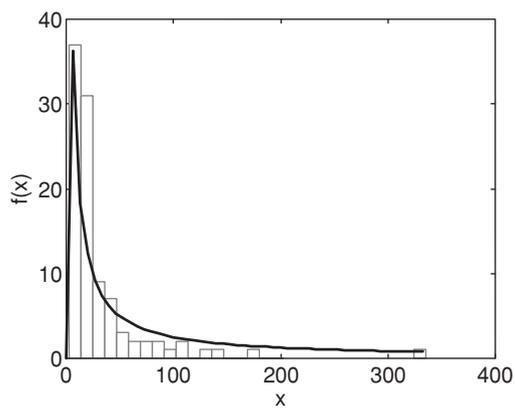


Figure 1 Illustration of a Histogram of Loss Data and Fitted Continuous Density

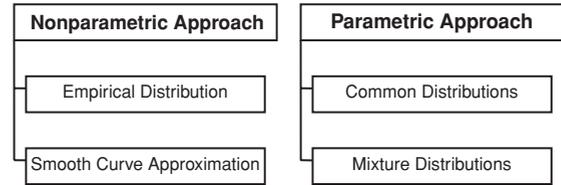


Figure 2 Approaches to Modeling Loss Severity

NONPARAMETRIC APPROACH: EMPIRICAL DISTRIBUTION FUNCTION

Modeling operational losses with their empirical distribution function is a nonparametric approach as it does not involve estimation of the parameters of a loss distribution. In this sense, it is the simplest approach. On the other hand, it makes the following two critical assumptions regarding future loss data:

- Historic loss data are sufficiently comprehensive.
- All past losses are equally likely to reappear in the future, and losses of other magnitudes (such as potential extreme events that are not a part of existent database) cannot occur.

Suppose we want to find the empirical distribution function of a random variable X . It is found by:

$$P(X \leq x) = \frac{\text{number of losses} \leq x}{\text{total number of losses}}$$

The empirical distribution function looks like a step function, with a step up occurring at each observed value of X . Figure 3 provides an illustration. The density function⁴ is simply a relative frequency histogram with a bar at each observed data value, and the height of each bar shows the proportion of losses of this magnitude out of total.

Note that the empirical distribution is often used in goodness-of-fit tests. One can compare it with a fitted loss distribution, and if the fitted loss distribution follows closely the empirical distribution, then this indicates a good fit; if it does not follow closely the empirical

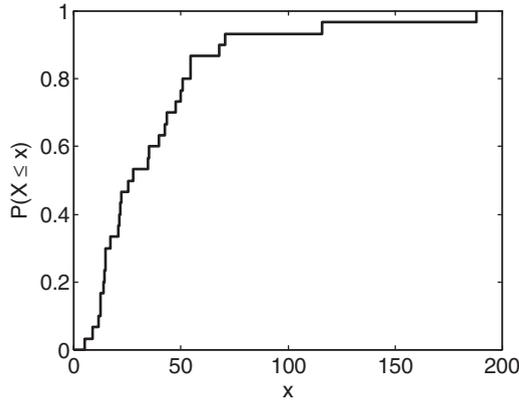


Figure 3 Illustration of Empirical Distribution Function

distribution function, then the loss distribution is not optimal.

PARAMETRIC APPROACH: CONTINUOUS LOSS DISTRIBUTIONS

In this section, we review several popular loss distributions. Certainly, a variety of additional distributions may be created by using some transformation of the original data and then fitting a distribution to the transformed data. A popular transformation involves taking the natural logarithm of the data. It is notable that if the original data are severely right-skewed, then the distribution of the log-data often becomes “bell-shaped” and nearly symmetric. For example, fitting the normal distribution to the log-data is equivalent to fitting the lognormal distribution to the original data.

Exponential Distribution

The exponential distribution for a random variable X of length n is described by its density f and distribution F of the following form:

$$f(x) = \lambda e^{-\lambda x}, \quad F(x) = 1 - e^{-\lambda x}, \quad x > 0$$

The distribution is characterized by only one parameter λ ($\lambda > 0$), which is the scale parameter.

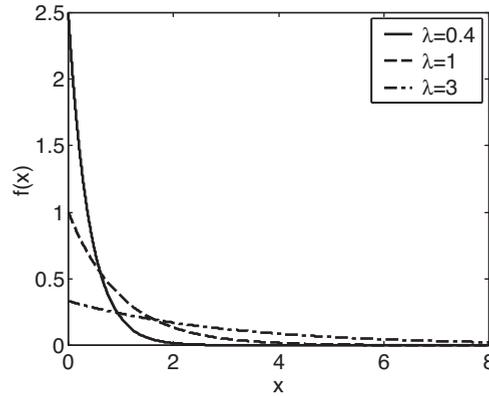


Figure 4 Illustration of Exponential Density

Examples of exponential densities are illustrated in Figure 4. The maximum likelihood estimate (MLE) for λ is

$$\hat{\lambda} = \frac{1}{\bar{x}} \quad \text{where} \quad \bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$$

Raw moments are calculated as:

$$\mathbb{E}(X^k) = \frac{k!}{\lambda^k}$$

and so the population mean and variance are

$$\text{mean}(X) = 1/\lambda, \quad \text{var}(X) = 1/\lambda^2$$

The mode of an exponential distribution is located at zero. The skewness and kurtosis coefficients are $\gamma_1 = 2$ and $\gamma_2 = 6$, respectively.

The inverse of the distribution has a simple form $F^{-1}(p) = -1/\lambda \log(1 - p)$, $p \in (0, 1)$, and so an exponential random variate can be simulated using the inverse transform method by $X = -\frac{1}{\lambda} \log U$, where U is distributed uniformly on the $(0, 1)$ interval. Another popular simulation method uses the Von Neumann algorithm.

The exponential density is monotonically decreasing toward the right and is characterized by an exponentially decaying right tail of the form $\bar{F}(x) = e^{-\lambda x}$, which means that high-magnitude events are given a near-zero probability. For this reason, it is unlikely that it would find much use in modeling operational losses, where arguably the central concern is the

losses of a very high magnitude (unless, perhaps, some generalizations of the exponential distribution or mixture models are considered).

Note that another parameterization of the exponential distribution is possible, with the density specified as $f(x) = \frac{1}{\lambda} e^{-\frac{1}{\lambda}x}$.

Lognormal Distribution

A random variable X has a lognormal distribution if its density and distribution are:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}$$

$$F(x) = \Phi\left(\frac{\log x - \mu}{\sigma}\right), \quad x > 0$$

where $\Phi(x)$ is the distribution of a standard normal, $N(0, 1)$, random variable, and can be obtained by looking up the table of the standard normal quantiles.⁵

Examples of the lognormal density are illustrated in Figure 5. The parameters μ ($-\infty < \mu < \infty$) and σ ($\sigma > 0$) are the location and scale parameters, respectively, and can be estimated with MLE as:

$$\hat{\mu} = \frac{1}{n} \sum_{j=1}^n \log x_j, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^n (\log x_j - \hat{\mu})^2 \quad (1)$$

Raw moments are calculated as:

$$\mathbb{E}(X^k) = e^{\mu k + \frac{\sigma^2 k^2}{2}}$$

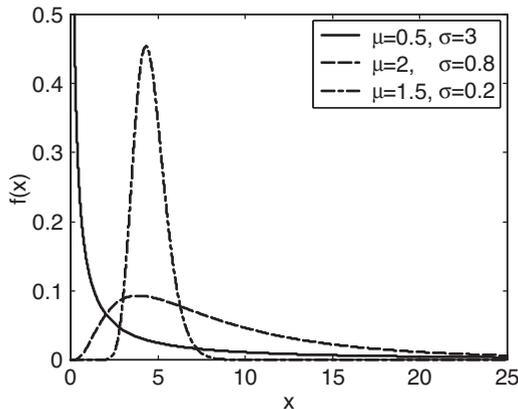


Figure 5 Illustration of Lognormal Density

and so the population mean and variance are calculated to be

$$\text{mean}(X) = e^{\mu + \frac{\sigma^2}{2}}, \quad \text{var}(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$$

The mode is located at $e^{\mu - \sigma^2}$. The skewness and kurtosis coefficients are:

$$\gamma_1 = \sqrt{e^{\sigma^2} - 1}(2 + e^{\sigma^2})$$

$$\gamma_2 = e^{4\sigma^2} + 2e^{3\sigma^2} + 3e^{2\sigma^2} - 6$$

The inverse of the distribution is $F^{-1}(p) = e^{\Phi^{-1}(p)\sigma + \mu}$, and so a lognormal random variate can be simulated by $X = e^{\Phi^{-1}(U)\sigma + \mu}$, where Φ is the standard normal distribution. Note that a lognormal random variable can be obtained from a normal random variable Y with parameters μ and σ (this is often written as $N(\mu, \sigma)$) via the transformation $X = e^Y$. Thus, if X has a lognormal distribution, then $\log X$ has a normal distribution with the same parameters.

The lognormal distribution is characterized by moderately heavy tails, with the right tail $\bar{F}(x) \sim x^{-1}e^{-\log^2 x}$. To fit a lognormal distribution to the data, one can take the natural logarithm of the dataset, and then fit to it the normal distribution. Note that the MLE will produce the same estimates, but the method of moments will produce different parameter estimates.

Weibull Distribution

The Weibull distribution is a generalization of the exponential distribution: Two parameters instead of one parameter allow for greater flexibility and heavier tails. The density and distribution are⁶

$$f(x) = \alpha\beta x^{\alpha-1} e^{-\beta x^\alpha}, \quad F(x) = 1 - e^{-\beta x^\alpha}, \quad x > 0$$

with β ($\beta > 0$) being the scale parameter and α ($\alpha > 0$) the shape parameter.

Examples of the density are illustrated in Figure 6. The MLE estimators for the parameters do not exist in closed form, and should be evaluated numerically. Raw moments are calculated as:

$$\mathbb{E}(X^k) = \beta^{-k/\alpha} \Gamma\left(1 + \frac{k}{\alpha}\right)$$

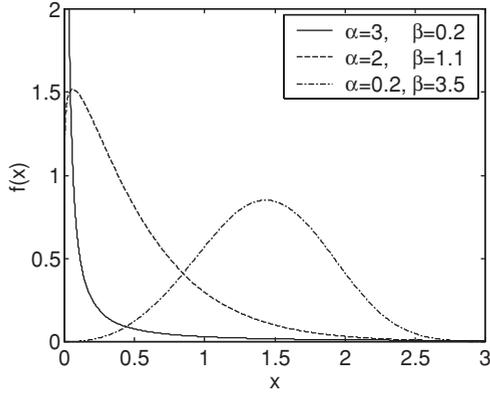


Figure 6 Illustration of Weibull Density

and so the population mean and variance are:

$$\text{mean}(X) = \beta^{-1/\alpha} \Gamma\left(1 + \frac{1}{\alpha}\right)$$

$$\text{var}(X) = \beta^{-2/\alpha} \left(\Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma^2\left(1 + \frac{1}{\alpha}\right) \right)$$

The mode is located at $\beta^{-1}(1 - \alpha^{-1})^{1/\alpha}$ for $\alpha > 0$ and at zero otherwise. The formulae for the skewness and kurtosis coefficients are:

$$\gamma_1 = \frac{2\Gamma^3(1 + \frac{1}{\alpha}) - 3\Gamma(1 + \frac{1}{\alpha})\Gamma(1 + \frac{2}{\alpha}) + \Gamma(1 + \frac{3}{\alpha})}{[\Gamma(1 + \frac{2}{\alpha}) - \Gamma^2(1 + \frac{1}{\alpha})^{3/2}]}$$

$$\gamma_2 = \frac{-6[\Gamma^4(1 + \frac{1}{\alpha}) - 12\Gamma^2(1 + \frac{1}{\alpha})\Gamma(1 + \frac{2}{\alpha}) - 3\Gamma^2(1 + \frac{2}{\alpha}) - 4\Gamma(1 + \frac{1}{\alpha})\Gamma(1 + \frac{3}{\alpha}) + \Gamma(1 + \frac{4}{\alpha})]}{[\Gamma(1 + \frac{2}{\alpha}) - \Gamma^2(1 + \frac{1}{\alpha})]^2}$$

The inverse of a Weibull random variable does not exist in a simple closed form. To generate a Weibull random variable, one can first generate an exponential random variable Y with parameter β and then follow the transformation $X = Y^{1/\alpha}$.

The right tail behavior of a Weibull random variable follows the form $\bar{F}(x) = e^{-\beta x^\alpha}$, and so the distribution is heavy-tailed for $\alpha < 1$. Weibull distribution has been found to be the optimal distribution in reinsurance models⁷ as well as in asset returns models.⁸

Note the following regarding the Weibull distribution. First, if $\alpha = 1$, then the Weibull distribution reduces to the exponential distribution. Second, other parameterizations of the Weibull distribution are possible. For example, some au-

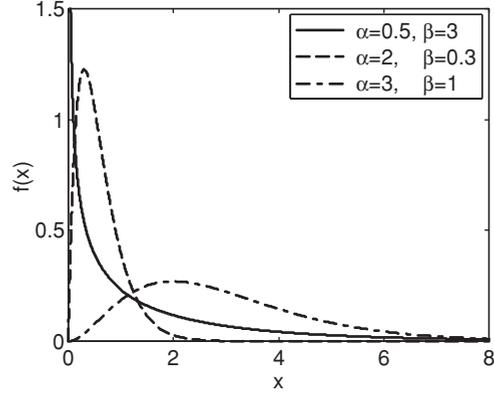


Figure 7 Illustration of Gamma Density

thors use $1/\beta$ instead of β . Sometimes $1/\beta^\alpha$ is used instead of β .

Gamma Distribution

The gamma distribution is another generalization of an exponential distribution and

is specified by its density and distribution given by⁹

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

$$F(x) = \Gamma(\alpha; \beta x), \quad x > 0$$

where the two parameters, α ($\alpha > 0$) and β ($\beta > 0$), characterize the shape and scale, respectively.

Examples of the density are illustrated in Figure 7. The MLE estimates for the parameters can be only evaluated numerically. The raw moments are found by:

$$\mathbb{E}(X^k) = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)\beta^k}$$

yielding the population mean and variance as

$$\text{mean}(X) = \frac{\alpha}{\beta}, \quad \text{var}(X) = \frac{\alpha}{\beta^2}$$

The mode is $\frac{\alpha-1}{\beta}$ for $\alpha > 1$ and zero otherwise. The skewness and kurtosis coefficients are found by

$$\gamma_1 = \frac{2}{\sqrt{\alpha}}, \quad \gamma_2 = \frac{6}{\alpha}$$

If α is an integer,¹⁰ then to generate a gamma random variable with parameters α and β one can generate a sum of α exponential random variables each with parameter β . Hence, if $U_1, U_2, \dots, U_\alpha$ are independent uniform $(0, 1)$ random variables, then $X = -1/\beta \log(\prod_{j=1}^\alpha U_j)$ has the desired distribution. A variety of methods for generation of a gamma random variable is described in Devroye (1986).

Beta Distribution

The beta distribution has density and distribution of the following form:¹¹

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$F(x) = I(x; \alpha, \beta), \quad 0 \leq x \leq 1$$

Examples of the density are illustrated in Figure 8. Note that X has a bounded support on $[0, 1]$. Certainly, operational loss data may be rescaled to fit this interval. In this case, the following version of the beta density and distri-

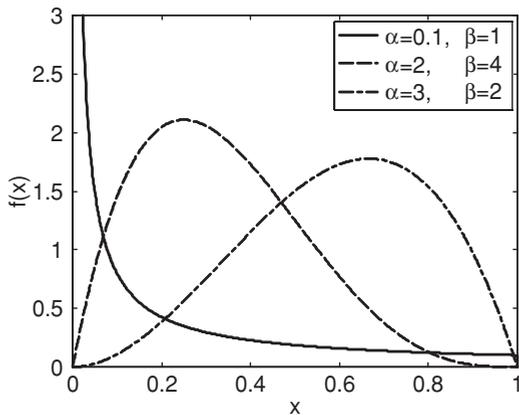


Figure 8 Illustration of Beta Density

bution is possible (the parameter θ is assumed known):

$$f(x) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \left(\frac{x}{\theta}\right)^{\alpha-1} \left(1 - \frac{x}{\theta}\right)^{\beta-1} \frac{1}{x}$$

$$F(x) = I\left(\frac{x}{\theta}; \alpha, \beta\right), \quad 0 < x < \theta, \quad \theta > 0$$

The parameters α ($\alpha > 0$) and β ($\beta > 0$) determine the shape of the distribution. The MLE estimators can be evaluated numerically. The raw moments for the regular version of the beta density can be found by

$$\mathbb{E}(X^k) = \frac{(\alpha + \beta - 1)!(\alpha + k - 1)!}{(\alpha - 1)!(\alpha + \beta + k - 1)!}$$

yielding the mean and the variance:

$$\text{mean}(X) = \frac{\alpha}{\alpha + \beta}$$

$$\text{var}(X) = \frac{\alpha\beta}{(a + \beta)^2(\alpha + \beta + 1)}$$

The mode is equal to $(\alpha - 1)/(\alpha + \beta - 2)$. The skewness and kurtosis coefficients are estimated by

$$\gamma_1 = \frac{2(\beta - \alpha)\sqrt{1 + \alpha + \beta}}{\sqrt{\alpha + \beta}(2 + \alpha + \beta)}$$

$$\gamma_2 = \frac{6[\alpha^3 + \alpha^2(1 - 2\beta) + \beta^2(1 + \beta) - 2\alpha\beta(2 + \beta)]}{\alpha\beta(\alpha + \beta + 2)(\alpha + \beta + 3)}$$

The beta random variate can be generated using an algorithm described in Ross (2001, 2002) or Devroye (1986).

Note that the beta distribution is related to the gamma distribution. Suppose we have two gamma random variables X and Y with parameters α_1, β_1 and α_2, β_2 , respectively. Then the variable $Z = X/(X+Y)$ has a beta distribution with parameters α_1, α_2 . This property can be used to generate a beta random variate from two gamma random variates.

Pareto Distribution

The Pareto distribution is characterized by its density and distribution of the form:

$$f(x) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}}, \quad F(x) = 1 - \left(\frac{\beta}{x}\right)^\alpha, \quad \beta < x < \infty$$

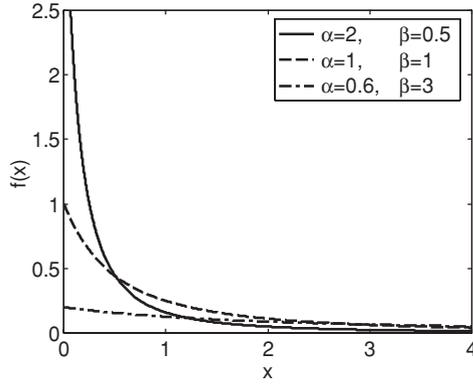


Figure 9 Illustration of Pareto Density

Note that the range of permissible values of X depends on the scale parameter β ($\beta > 0$). The parameter α ($\alpha > 0$) determines the shape.

Figure 9 illustrates some examples of the density. No closed-form expressions for the MLE estimators exist (except for the case when $\beta = 1$, in which case $\hat{\alpha} = n / \sum_{j=1}^n \log x_j$), so they have to be evaluated numerically.

The raw moments are estimated by

$$\mathbb{E}(X^k) = \frac{\alpha\beta^k}{\alpha - k}$$

from which the population mean and variance are found to be

$$\begin{aligned} \text{mean}(X) &= \frac{\alpha\beta}{\alpha - 1} \quad \text{for } \alpha > 1 \\ \text{var}(X) &= \frac{\alpha\beta^2}{(\alpha - 1)^2(\alpha - 2)} \quad \text{for } \alpha > 2 \end{aligned}$$

The mode is equal to zero. The skewness and kurtosis coefficients are:

$$\begin{aligned} \gamma_1 &= \sqrt{\frac{\alpha - 2}{\alpha} \frac{2(\alpha + 1)}{\alpha - 3}} \\ \gamma_2 &= \frac{6(\alpha^3 + \alpha^2 - 6\alpha - 2)}{\alpha(\alpha - 3)(\alpha - 4)} \end{aligned}$$

The inverse of the distribution is $F^{-1}(p) = \beta((1 - p)^{-1/\alpha} - 1)$, which can be used to generate a Pareto random variate.

The Pareto distribution is a very *heavy-tailed distribution*, as is seen from the tail behavior. α determines the heaviness of the right tail, which is monotonically decreasing for the Pareto dis-

tribution: The closer it is to zero, the thicker the tail, $\bar{F}(x) = \left(\frac{\beta}{\beta+x}\right)^\alpha$. Tails proportional to $x^{-\alpha}$ are called the power tails (as opposed to the exponentially decaying tails) because they follow a power function. The case when $\alpha \leq 1$ refers to a very heavy-tailed case, in which the mean and the variance are infinite (see the formulas for mean and variance earlier), means that losses of an infinitely high magnitude are possible.

While on one hand the Pareto distribution appears very attractive for modeling operational risk, as it is expected to capture very high-magnitude losses, on the other hand, from the practical point of view, the possibility of infinite mean and variance could pose a problem.

Note the following:

- Different versions of the Pareto distribution are possible. Occasionally a simplified, 1-parameter version of the Pareto distribution is used, with $\beta = 1$.
- A 1-parameter Pareto random variable may be obtained from an exponential random variable via a simple transformation. If a random variable Y follows an exponential distribution with parameter λ , then $X = e^Y$ has the 1-parameter Pareto distribution with the same shape parameter.
- A 2-parameter Pareto distribution may be reparameterized in such a way that we obtain the generalized Pareto distribution (GPD). The GPD can be used to model extreme events that exceed a high threshold.

Burr Distribution

The Burr distribution is a generalized three-parameter version of the Pareto distribution and allows for greater flexibility in the shape due to additional shape parameter γ ($\gamma > 0$). The density and distribution functions can be written as

$$\begin{aligned} f(x) &= \gamma\alpha\beta^\alpha \frac{x^{\gamma-1}}{(\beta + x^\gamma)^{\alpha+1}} \\ F(x) &= 1 - \left(\frac{\beta}{\beta + x^\gamma}\right)^\alpha, \quad x > 0 \end{aligned}$$

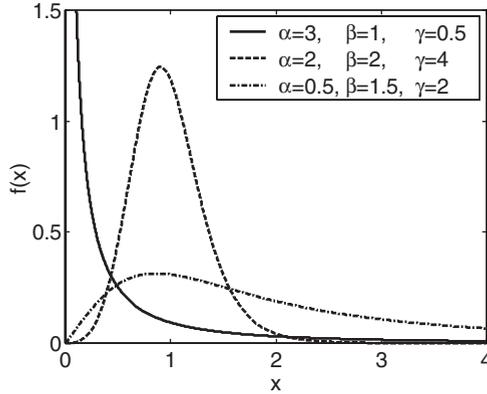


Figure 10 Illustration of Burr Density

Examples of the density are depicted in Figure 10. The MLE estimators for the parameters can generally be evaluated only numerically. The raw moments are estimated as:

$$\mathbb{E}(X^k) = \frac{\beta^{k/\gamma} \Gamma\left(1 + \frac{k}{\gamma}\right) \Gamma\left(\alpha - \frac{k}{\gamma}\right)}{\Gamma(\alpha)},$$

$$-\gamma < k < \gamma\alpha$$

from which the population mean and variance are calculated as:

$$\text{mean}(X) = \frac{\beta^{1/\gamma} \Gamma\left(1 + \frac{1}{\gamma}\right) \Gamma\left(\alpha - \frac{1}{\gamma}\right)}{\Gamma(\alpha)}$$

$$\gamma\alpha > 1$$

$$\text{var}(X) = \frac{\beta^{2/\gamma} \Gamma\left(1 + \frac{2}{\gamma}\right) \Gamma\left(\alpha - \frac{2}{\gamma}\right)}{\Gamma(\alpha)}$$

$$- \frac{\beta^{2/\gamma}}{\Gamma^2(\alpha)} \Gamma^2\left(1 + \frac{1}{\gamma}\right) \Gamma^2\left(\alpha - \frac{1}{\gamma}\right),$$

$$\gamma\alpha > 2$$

The mode is equal to $\frac{1}{\beta^{1/\gamma}} \left(\frac{\gamma-1}{\alpha\gamma+1}\right)^{1/\gamma}$ for $\gamma > 1$ and zero otherwise.

The Burr random variable can be generated by the inverse transform method, using $F^{-1}(p) = (\beta((1-p)^{-1/\alpha} - 1))^{1/\gamma}$.

The right tail has the power law property and obeys $\bar{F}(x) = \left(\frac{\beta}{\beta+x^\gamma}\right)^\alpha$. The distribution is heavy-tailed for the case $\alpha < 2$ and is very heavy-tailed when $\alpha < 1$. The Burr distribution has been used in the insurance industry, and has been found to be an optimal distribution for natural catastrophe insurance claims.¹²

Note the following two points. First, if $\gamma = 1$, then the Burr distribution reduces to the Pareto distribution. Second, other parameterizations of the Burr distribution are possible. For example, the Burr distribution with $\beta = 1$ is known as the loglogistic distribution.

EXTENSION: MIXTURE LOSS DISTRIBUTIONS

Histograms of the operational loss data often reveal a very high peak close to zero and a smaller but distinct peak toward the right tail. This may suggest that the operational loss data often do not follow a pattern of a single distribution, even for data belonging to the same loss type (such as operational losses due to business disruptions) and the same business line (such as commercial banking). One approach in modeling such losses would be to consider the GPD to model the tail events and an empirical or other distribution for the remaining lower-magnitude losses. Alternatively, one may consider a single distribution composed by a mixture of two or more loss distributions.

The density and distribution of a m -point mixture distribution can be expressed as

$$f(x) = \sum_{j=1}^m w_j f_j(x), \quad F(x) = \sum_{j=1}^m w_j F_j(x)$$

where $w_j, j = 1, 2, \dots, m$, are the positive weights attached to each member distribution, adding up to 1. It is possible to have a mixture of different types of distributions, such as exponential and Weibull, or of the same type of distribution but with different parameters.

An example of a mixture of two lognormal distributions ($\mu_1 = 0.9, \sigma_1 = 1, \mu_2 = 3, \sigma_2 = 0.5$) is depicted in Figure 11.

The MLE estimates of the parameters (including the weights) of mixture distributions can generally be evaluated only numerically. A commonly used procedure to estimate the parameters of mixture distributions is the expectation-maximization algorithm. The raw

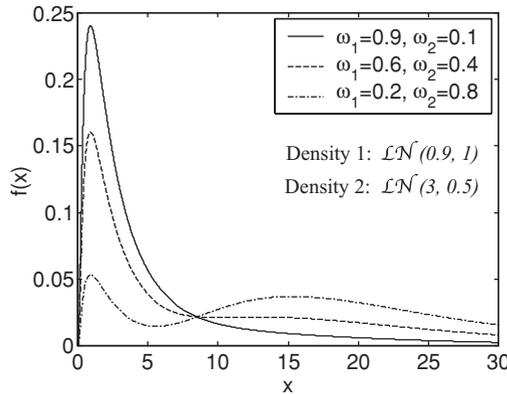


Figure 11 Illustration of 2-Point Lognormal Mixture Density

moments are found as the weighted sum of the k th moments evaluated individually for each of the m member distributions. The population mean and variance are found by

$$\text{mean}(X) = \sum_{j=1}^m w_j \mathbb{E}_j(X), \quad \text{var}(X) = \sum_{j=1}^m w_j^2 \sigma_j^2(X)$$

where the subscripts j refer to each member density. The right tail follows $\bar{F}(x) = \sum_{j=1}^m w_j \bar{F}_j(x)$.

The advantage of using mixture distributions is that they can be fitted to practically all shapes of loss distributions. On the other hand, the models may lack reliability due to a large number of parameters that need to be estimated (in particular, when the available loss data set is not large enough). For example, a 2-point mixture of exponential distributions requires only three parameters, but a 4-point mixture of exponential distributions requires seven parameters. In some cases, this problem may be overcome when certain simplifications are applied to the model. For example, it is possible to achieve a 2-point mixture of Pareto distributions with four, instead of five, unknown parameters; the following distribution has been successfully applied to liability insurance:

$$F(x) = 1 - a \left(\frac{\beta_1}{\beta_1 + x} \right)^\alpha + (1 - a) \left(\frac{\beta_2}{\beta_2 + x} \right)^{\alpha+2}$$

with the first distribution covering smaller magnitude events and having a higher weight a attached, and the second distribution covering infrequent large-magnitude events.¹³

An extension to mixture distributions may be to allow m to be a parameter, and “let the data decide” on how many distributions should enter the mixture. This, however, makes the model data-dependent and more complex.¹⁴

Note that the term mixture distribution is sometimes also used for distributions in which an unknown parameter is believed to be random and follows some distribution rather than being fixed. For example, a mixture of Poisson and gamma distributions (i.e., the parameter of the Poisson distribution follows a gamma distribution) will result in a hypergeometric distribution.

A NOTE ON THE TAIL BEHAVIOR

Operational risk managers are concerned with finding a model that would capture the “tail events.” In the context of operational losses, it is understood that *tail events* refer to the events in the upper tail of the loss distribution. A crucial task in operational risk modeling is to produce a model that would give a realistic account to the possibility of losses exceeding a very high amount (this becomes critical in the estimation of the Value-at-Risk).

In operational risk modeling, thin-tailed distributions should be used with caution. The following example illustrates the danger of fitting a *light-tailed distribution* to the data whose true distribution is heavy-tailed.¹⁵ We generated 5,000 points from the Pareto distribution (heavy-tailed) with parameters $\alpha = 1.67$ and $\beta = 0.6$. We then fitted an exponential distribution (light-tailed) to the data. The MLE procedure resulted in the exponential parameter of $\lambda = 1.61$. Figure 12 demonstrates the difference in the behavior of the tails of both distributions. In the far right, the probability of exceeding any high point is significantly lower (roughly, by

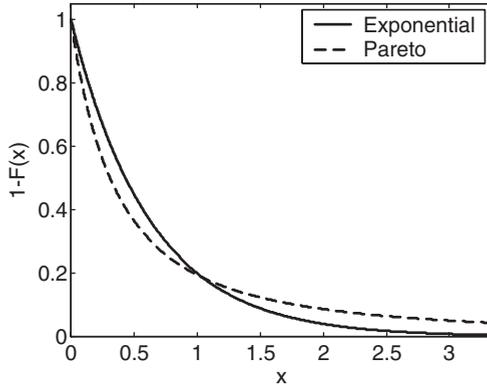


Figure 12 Tails of Pareto and Exponential Distributions Fitted to Simulated Pareto Random Variable

5%) under the exponential fit. This indicates that the probability of high-value events (and exceeding them) will be underestimated if one commits the mistake of fitting a thin-tailed loss distribution to the loss data. Such mistakes may be costly and lead to serious consequences in the operational risk management, if the potential for high-magnitude losses is being inadequately assessed.

In Table 1 common distributions are classified into two categories depending on the heaviness of the right tail. Note that the Weibull distribution can be thin-tailed or heavy-tailed depending on the value of the shape parameter. Regarding the lognormal distribution, some

literature refers to it as a thin-tailed distribution, but we follow Embrechts, Klüppelberg, and Mikosch (1997), who put it in the class of medium-tailed distributions. The beta distribution has a bounded support, which makes it a thin-tailed distribution.

EMPIRICAL EVIDENCE WITH OPERATIONAL LOSS DATA

In this section we provide results from empirical studies based on operational loss data that apply the distributions described in this entry. There are two types of studies: Those based on real operational loss data and those based on simulated data.

The empirical studies indicate that practitioners try a variety of possible loss distributions for the loss data and then determine an optimal one on the basis of goodness-of-fit tests. It is common to use the Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) tests to examine the goodness of fit of the model to the data. The two tests use different measures of the discrepancy between the fitted continuous distribution and the empirical distribution functions. The KS test better captures the discrepancy around the median of the data, while the AD test is more optimal for the tails. A smaller value of the test statistic indicates a

Table 1 Tail Behavior of Common Loss Distributions

Name	Tail $\bar{F}(x)$	Parameters
Thin-Tailed Distributions		
Normal	$\bar{F}(x) = 1 - \Phi\left(\frac{x-\mu}{\sigma}\right)$	$-\infty < \mu < \infty, \sigma > 0$
Exponential	$\bar{F}(x) = e^{-\lambda x}$	$\lambda > 0$
Gamma	$\bar{F}(x) = 1 - \Gamma(\alpha; \beta x)$	$\alpha, \beta > 0$
Weibull	$\bar{F}(x) = e^{-\beta x^\alpha}$	$\alpha \geq 1, \beta > 0$
Beta	$\bar{F}(x) = 1 - I(x; \alpha, \beta)$	$\alpha, \beta > 0$
Medium-Tailed and Heavy-Tailed Distributions		
Lognormal	$\bar{F}(x) = 1 - \Phi\left(\frac{\log x - \mu}{\sigma}\right)$	$-\infty < \mu < \infty, \sigma > 0$
Weibull	$\bar{F}(x) = e^{-\beta x^\alpha}$	$0 < \alpha < 1, \beta > 0$
Pareto	$\bar{F}(x) = \left(\frac{\beta}{\beta+x}\right)^\alpha$	$\alpha, \beta > 0$
Burr	$\bar{F}(x) = \left(\frac{\beta}{\beta+x^\gamma}\right)^\alpha$	$\alpha, \beta, \gamma > 0$

better fit. Other goodness-of-fit tests include Kuiper, Cramér-von Mises, and Pearson's χ^2 test, among others.

Studies with Real Data

We review some empirical studies based on real operational loss data from financial institutions.

Müller Study of 1950–2002 Operational Loss Data

Müller (2002) carried out empirical analysis with external operational loss data obtained from worldwide institutions in the 1950–2002 period, made available then by the IC² Operational Loss F1RST Database. Only data in U.S. dollars for the events whose state of affairs was “closed” or “assumed closed” on an indicated date were considered for the analysis. The data were available for five loss types:

- “Relationship” (such as events related to legal issues, negligence, and sales-related fraud).
- “Human” (such as events related to employee errors, physical injury, and internal fraud).
- “Processes” (such as events related to business errors, supervision, security, and transactions).
- “Technology” (such as events related to technology and computer failures and telecommunications).
- “External” (such as events related to natural and man-made disasters and external fraud).

Figure 13 shows the histograms of the five data sets. There is a clear peak in the beginning, which is captured by the excessive kurtosis; a heavy right tail is also evident and is captured by the high degree of positive skewness (see Table 2).

From the common distributions discussed in this entry, exponential, lognormal, Weibull, gamma, and Pareto distributions were used. Table 2 demonstrates the five samples' MLE parameter estimates and KS and AD statistic values for the five distributions. The center of the data is best explained by the lognormal distribution, as is concluded from the lowest KS

statistic values, for all except “Technology” type losses for which Weibull is the best. The same conclusions are drawn regarding the tails of the datasets.

Cruz Study of Legal Loss Data

Cruz (2002) applies exponential, Weibull, and Pareto distributions to a sample (in U.S. dollars) from a legal database (from an undisclosed source), consisting of 75 points.¹⁶ The sample's descriptive statistics, as well as the MLE parameters for the three distributions¹⁷ and goodness-of-fit statistics are depicted in Table 3. The data are highly leptokurtic and significantly right-skewed. Based on visual and formal tests for the goodness of fit,¹⁸ Cruz concluded that the Pareto distribution fits the data best. Nevertheless, none of the considered loss distributions is able to capture well the heaviness of the upper tail.

Moscadelli Study of 2002 LDCE Operational Loss Data

Moscadelli (2004) explores the data (in euros) collected by the Risk Management Group (RMG) of the Basel Committee in June 2002's Operational Risk Loss Data Collection Exercise (LDCE). There were 89 participating banks from 19 countries worldwide that provided their internal loss data for the year 2001. The data were classified into eight business lines and pooled together across all banks. The eight business lines are:

- BL1: Corporate Finance.
- BL2: Trading and Sales.
- BL3: Retail Banking.
- BL4: Commercial Banking.
- BL5: Payment and Settlement.
- BL6: Agency Services.
- BL7: Asset Management.
- BL8: Retail Brokerage.

The lognormal, gamma, Gumbel, Pareto, and exponential distributions were fitted to the

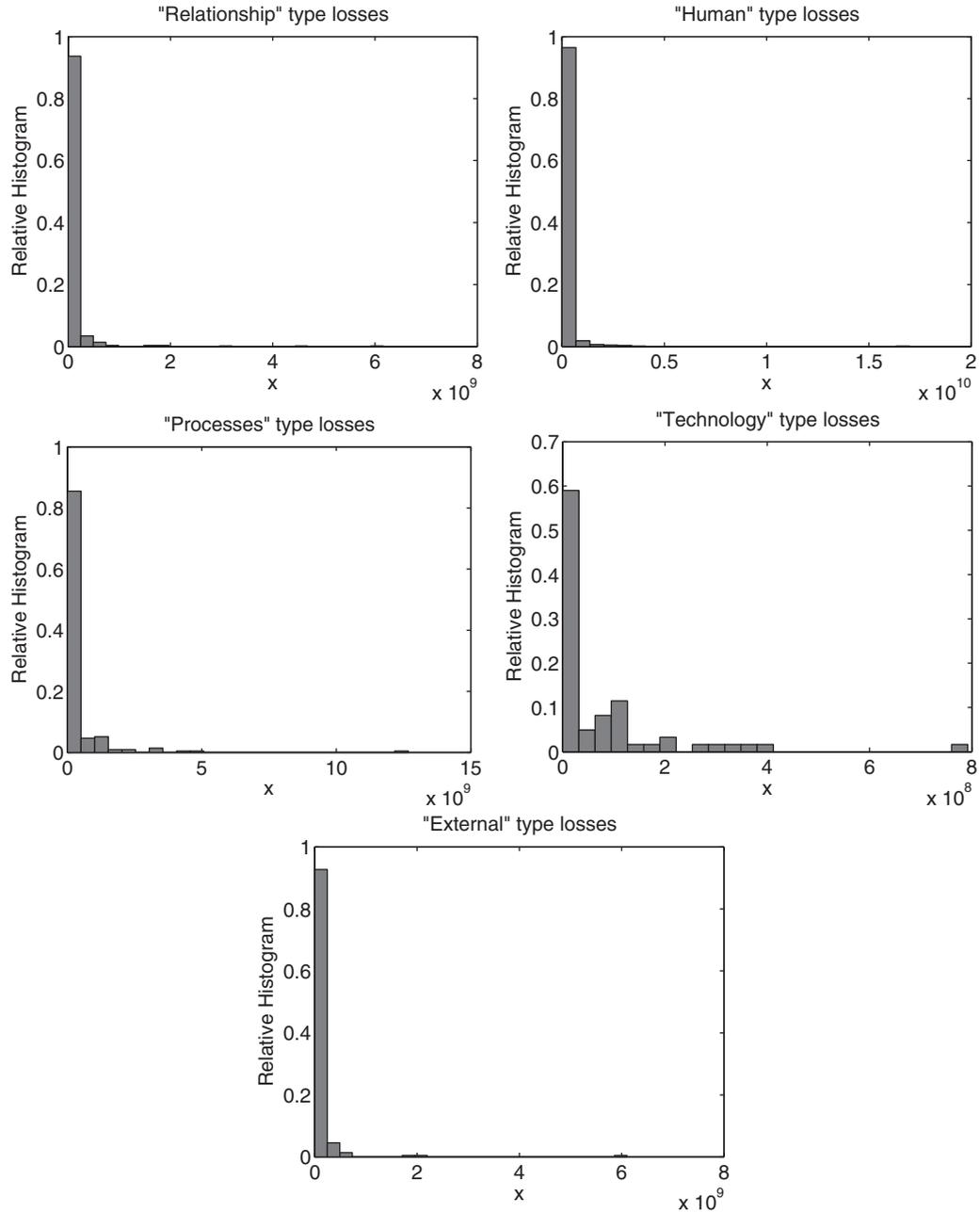


Figure 13 Relative Frequency Histograms of Operational Loss Data in Müller Study

data. The estimation procedure used in the study was somewhat simplified for two reasons. First, different banks used different minimum truncation levels for their internal data, roughly between €6,000 to €10,000. This issue was ignored in the estimation process. Sec-

ond, the data across all participating banks were pooled together without any consideration given for bank characteristics such as size.

Table 4 reproduces the sample descriptive statistic (based on 1,000 bootstrapped samples generated from the original data), MLE

Table 2 Sample Description, Parameter Estimates, and Goodness-of-Fit Tests in the Müller Study

	“Relationship”	“Human”	“Processes”	“Technology”	“External”
1. Sample Description					
# obs.	585	647	214	61	220
Mean (\$ '000,000)	0.0899	0.1176	0.3610	0.0770	0.0930
Median (\$ '000)	12.8340	6.3000	50.1708	11.0475	8.9076
St.Dev. (\$ '000,000)	0.3813	0.7412	1.0845	0.1351	0.4596
Skewness	11.1717	18.8460	7.8118	3.0699	10.9407
Kurtosis	152.2355	418.8717	81.5218	14.7173	136.9358
2. MLE Parameter Estimates and Goodness-of-Fit Test Statistics					
Exponential distribution					
λ	$9.0 \cdot 10^7$	$0.15 \cdot 10^7$	$0.36 \cdot 10^7$	$7.7 \cdot 10^7$	$9.3 \cdot 10^7$
KS test	0.4024	0.5489	0.3864	0.3909	0.4606
AD test	$1.2 \cdot 10^5$	8460	3.9185	1.9687	430.2
Lognormal distribution					
μ	16.2693	15.9525	17.6983	16.1888	15.9696
σ	2.1450	2.4551	2.2883	2.5292	2.2665
KS test	0.0301	0.0530	0.0620	0.1414	0.0449
AD test	0.0787	0.1213	0.1600	0.3043	0.1597
Weibull distribution					
α	0.0002	0.0008	0.0001	0.0003	0.0004
β	0.4890	0.4162	0.4822	0.4692	0.4527
KS test	0.0608	0.0907	0.0656	0.1179	0.0749
AD test	0.4335	0.2231	0.2247	0.2372	0.2696
Gamma distribution					
α	—	—	0.3372	0.3425	—
β	—	—	$1.07 \cdot 10^9$	$0.2 \cdot 10^9$	—
KS test	—	—	0.1344	0.1357	—
AD test	—	—	—	—	—
Pareto distribution					
α	-0.8014	-0.8936	-0.7642	-0.6326	-0.8498
β	$1.8 \cdot 10^7$	$1.6 \cdot 10^7$	$8.5 \cdot 10^7$	$2.8 \cdot 10^7$	$1.4 \cdot 10^7$
KS test	0.1296	0.1979	0.1504	0.2812	0.1783
AD test	0.4031	0.5566	0.6256	1.0918	0.4784

Table 3 Sample Descriptive Statistics, Parameter Estimates, and Goodness-of-Fit Tests in the Cruz Study

1. Sample Description			
Mean (\$)	439,725.99		
Median (\$)	252,200		
St.dev. (\$)	538,403.93		
Skewness	4.42		
Kurtosis	23.59		
2. MLE Parameter Estimates and Goodness-of-Fit Test Statistics			
Exponential	$\lambda = 440,528.63$	KS test: 0.2104	W^2 test: 1.3525
Weibull	$\alpha = 2.8312$ $\beta = 0.00263$	KS test: 0.3688	W^2 test: 4.8726
Pareto	$\alpha = 6.1737$ $\beta = 2,275,032.12$	KS test: 0.1697	W^2 test: 0.8198

Source: Cruz (2002), pp. 57, 58, and 60, with modifications.

Table 4 Sample Descriptive Statistics, Parameter Estimates, and Goodness-of-Fit Statistics in the Moscadelli Study

	BL1	BL2	BL3	BL4	BL5	BL6	BL7	BL8
1. Sample Description								
# obs.	423	5,132	28,882	3,414	1,852	1,490	1,109	3,267
Mean (€'000)	646	226	79	356	137	222	195	125
St.dev. (€'000)	6,095	1,917	887	2,642	1,320	1,338	1,473	1,185
Skewness	16	23	55	15	24	13	25	32
Kurtosis	294	674	4,091	288	650	211	713	1,232
2. MLE Parameter Estimates and Goodness-of-Fit Test Statistics								
Lognormal distribution								
μ	3.58	3.64	3.17	3.61	3.37	3.74	3.79	3.58
σ	1.71	1.27	0.97	1.41	1.10	1.28	1.28	1.08
KS test	0.18	0.14	0.18	0.16	0.15	0.12	0.11	0.12
AD test	22.52	181	1,653	174	73.74	46.33	25.68	87.67
Gumbel distribution								
μ	93.96	51.76	25.63	48.30	35.86	54.82	56.78	41.03
σ	602	185	58.80	204	110	181	154	93.51
KS test	0.43	0.37	0.34	0.37	0.36	0.35	0.32	0.31
AD test	125	1,224	6,037	831	436	333	204	577

Source: Moscadelli (2004), pp. 19 and 25.

parameter estimates (based on the original data), and goodness-of-fit test statistics¹⁹ for the lognormal and Gumbel distributions.²⁰ Other considered distributions showed a poor fit. Although lognormal and Gumbel fitted the main body of the data rather well, they performed poorly in the upper tail, according to Moscadelli. This was confirmed by the test statistic values above the 90% critical values, meaning that it is unlikely that the data come from a selected distribution at the 90% confidence level.

He further performs the analysis of the data using the extreme value theory argument for modeling high losses with the GPD, finding that GPD outperforms other considered distributions. He also confirms the findings from other empirical studies that operational losses follow a very heavy-tailed distribution.

De Fontnouvelle-Rosengren-Jordan Study of 2002 LDCE Operational Loss Data

The dataset examined in Moscadelli was also analyzed by de Fontnouvelle, Rosengren, and Jordan (2006). They limited their analysis to the data collected from six banks, and

performed the analysis on the bank-by-bank basis, rather than pooling the data as was done in Moscadelli. For confidentiality reasons, only the data belonging to the four business lines—Trading and Sales (BL1), Retail Banking (BL2), Payment and Settlement (BL3), and Asset Management (BL4)—and six loss types—Internal Fraud (LT1), External Fraud (LT2), Employment Practices and Workplace Safety (LT3), Clients, Products and Business Practices (LT4), and Execution, Delivery and Process Management (LT5)—were included in the analysis.

The following distributions were considered for the study: exponential, Weibull, lognormal, gamma, loggamma (i.e., log of data is gamma-distributed), 1-parameter Pareto, Burr, and loglogistic.²¹ The distributions were fitted using the MLE method. Overall, heavy-tailed distributions—Burr, loggamma, loglogistic, and 1-parameter Pareto—fit the data very well, while thin-tailed distributions' fit is poor, as expected. In particular, losses of LT3 are well fit by most of the heavy-tailed distributions and lognormal. In many cases, the estimated parameters would be unreasonable, for example resulting in a negative mean loss. For some

Table 5 Sample Descriptive Statistics, Parameter Estimates, and Goodness-of-Fit Tests in the Lewis Study

1. Sample Description		
Mean (£)		151,944.04
Median (£)		103,522.90
St.dev. (£)		170,767.06
Skewness		2.84
Kurtosis		12.81
2. MLE Parameter Estimates and Goodness-of-Fit Test Statistics		
Normal	$\mu = 151,944.04, \sigma = 170,767.06$	AD test: 8.090
Exponential	$\lambda = 151,944.04$	AD test: 0.392
Weibull	$\alpha = 0.95446, \beta = 0.00001$	AD test: 0.267

Source: Lewis (2004), p. 88, with modifications.

BL and LT data sets, the models failed the χ^2 goodness-of-fit test for all considered cases. Hence, de Fontnouvelle, Rosengren, and Jordan performed additional analysis using the extreme value theory and fitting the GPD to the data exceeding a high threshold.²²

Lewis Study of Legal Liability Loss Data

Lewis (2004) reports his findings for a sample (in British pounds) of legal liability losses (from an undisclosed source), consisting of 140 points.²³ He fits the normal, exponential, and Weibull distributions²⁴ to the data and compares the fit. Table 5 shows the descriptive statistics for the sample, the MLE parameters for three fitted distributions, and the values of the AD goodness-of-fit statistic. The data are highly leptokurtic and significantly right-skewed. As expected, the normal distribution results in a very poor fit, and the Weibull distribution seems the most reasonable assumption,

based on the lowest value of the AD test statistic.

Studies with Simulated Data

A number of studies on operational risk that have appeared in literature were using simulated rather than real data. We present a few examples here.

Reynolds-Syer Study

Reynolds and Syer (2003) apply a nonparametric approach to modeling operational loss severity. They use a hypothetical sample of six-year internal operational loss data of a firm, with a total of 293 observations. The summary of input data is given in Table 6. Using the sample of historic data, sampling is repeated a large number of times, and 1,000 simulated years are created. For each year, the simulated losses are summed up. The distribution of yearly aggregated

Table 6 Sample Descriptive Statistics of Loss Data in the Reynolds-Syer Study

Year	# obs.	Total (\$ '000,000)	Average (\$ '000)	St. Dev. (\$ '000)
2000	64	7.55	117.9	109.6
2001	57	6.35	111.3	106.2
2002	52	5.14	98.8	93.7
2003	55	5.29	96.1	88.0
2004	43	3.86	89.7	78.5
2005	45	3.41	75.7	68.5

Source: Reynolds and Syer (2003), p. 204.

operational losses is assumed to follow the resulting empirical distribution.

Rosenberg-Schuermann Study

Rosenberg and Schuermann (2006) use a Monte Carlo approach to generate a sample of 200,000 operational losses. For the loss distribution they consider a 1-parameter Pareto distribution with parameter $1/0.65 = 1.5385$. This parameter is based on the average of the exponential parameters²⁵ of 1/0.64 and 1/0.66, obtained for logarithmic losses from the OpRisk Analytics database and OpVantage database, respectively, in the empirical study carried out by de Fontnouvelle, DeJesus-Rueff, Jordan, and Rosengren (2003). Recall that since the shape parameter is less than one, then such Pareto distribution has a finite mean but an infinite variance. To guarantee the existence of the first two moments, Rosenberg and Schuermann set a log-loss greater than 1,000 standard deviations equal to a loss of 1,000 standard deviations.

KEY POINTS

- Broadly, one can classify the approaches to model operational loss magnitudes into two groups: nonparametric approach and parametric approach.
- Under the nonparametric approach, one can either model the losses using the empirical distribution function, or one can fit a smooth curve to the histogram of the data and analyze the properties of the curve instead.
- Under the parametric approach, one can fit one (or more) of common parametric distributions directly to the data (and compare them).
- Because of the specific nature of the operational loss data, the distributions that are most likely to find application to modeling the losses are those that are right-skewed and are defined only on the positive values of the underlying random variable. These distributions include the exponential, lognormal, Weibull, gamma, beta, Pareto, Burr, and mixture distributions.
- Operational risk managers are concerned with finding a model that would capture the “tail events.” Common distributions are classified into two categories depending on the heaviness of the right tail: light-tailed and heavy-tailed. In operational risk modeling, light-tailed distributions should be used with caution.
- There have been several empirical studies with operational loss data. Two types of empirical studies are distinctive: studies that use real loss data and studies that use simulated data. Generally, most of the studies suggest that heavy-tailed loss distributions (such as lognormal or Pareto) best describe operational loss magnitudes.

NOTES

1. An example is cubic spline approximation as is done in Rosenberg and Schuermann (2006). Useful references on this approach include Silverman (1986) and Scott (1992).
2. See Rosenberg and Schuermann (2006).
3. See Cizek, Härdle, and Weron (2005).
4. To be more precise, for a discrete random variable it is called probability mass function.
5. The lognormal distribution was proposed by the Basel Committee for the operational risk modeling in 2001.
6. $\Gamma(a)$ is the complete gamma function, $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$. When a is an integer, then $\Gamma(a) = (a - 1)!$
7. See Madan and Unal (2004) and Kremer (1998).
8. See Mittnik and Rachev (1993a, 1993b).
9. $\Gamma(a; b)$ is the incomplete gamma function defined as $\Gamma(a; b) = \frac{1}{\Gamma(a)} \int_0^b t^{a-1} e^{-t} dt$.
10. In this case, the gamma distribution is called the Erlang distribution.
11. $I(x; \alpha, \beta)$ is the regularized beta function equal to $\int_0^x u^{\alpha-1} (1-u)^{\beta-1} du \times \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.
12. See Cizek, Härdle, and Weron (2005).

13. See Klugman, Panjer, and Willmot (2004).
14. See Klugman, Panjer, and Willmot (2004).
15. In literature, thin-tailed distributions are also called light-tailed distributions, and heavy-tailed distributions are also called fat-tailed distributions. We will use the corresponding terms interchangeably.
16. Original dataset is available from Cruz (2002), Chapter 3, p. 57.
17. Note that the density specification for the exponential and Weibull distributions in Cruz (2002) are different. We report the parameter values based on the specifications of the density functions as presented in this entry.
18. The KS values reported in Table 3 should be further scaled by \sqrt{n} (n being the sample length) if we want to compare the goodness of fit across samples of different lengths.
19. The test statistics are unadjusted to the length of data.
20. The Gumbel distribution is light-tailed and has density $f(x) = \frac{1}{\sigma} \exp\left\{-\frac{x-\mu}{\sigma} - \exp\left\{-\frac{x-\mu}{\sigma}\right\}\right\}$, defined on $x \in \Re$. The support allows for negative loss values, so the Gumbel distribution is unlikely to find much application in operational risk modeling.
21. The density of the loglogistic distribution is $f(x) = ax^{1/b-1}/[b(1+ax^{1/b})^2]$.
22. For the tables with the χ^2 goodness-of-fit statistic values and other details of this empirical study we refer the reader to de Fontnouvelle, Rosengren, and Jordan (2006).
23. Original dataset is available from Lewis (2004), Chapter 7, p. 87.
24. Lewis (2004) does not report the parameter estimates for the Gaussian and Weibull cases. We computed them directly by fitting the distributions to the data.
25. We stated earlier that an exponential transformation of an exponentially distributed random variable follows a 1-parameter Pareto distribution.

REFERENCES

- Cizek, P., Härdle, W., and Weron, R. (eds.) (2005). *Statistical Tools for Finance and Insurance*. Heidelberg: Springer.
- Cruz, M. G. (2002). *Modeling, Measuring, and Hedging Operational Risk*. Chichester, NY: John Wiley & Sons.
- de Fontnouvelle, P., DeJesus-Rueff, V., Jordan, J., and Rosengren, E. (2003). Using loss data to quantify operational risk. Technical report, Federal Reserve Bank of Boston.
- de Fontnouvelle, P., Rosengren, E., and Jordan, J. (2006). Implications of alternative operational risk modelling techniques. In M. Carey and R. Stulz (eds.), *The Risks of Financial Institutions*, NBER/University of Chicago Press.
- Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York, NY: Springer-Verlag.
- Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modeling External Events for Insurance and Finance*. Berlin: Springer-Verlag.
- Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2004). *Loss Models: From Data to Decisions*, 2nd ed. Hoboken, NJ: John Wiley & Sons.
- Kremer, E. (1998). Largest claims reinsurance premiums for the Weibull model. In *Blätter der Deutschen Gesellschaft für Versicherungsmathematik*: 279–284.
- Lewis, N. (2004). *Operational Risk with Excel and VBA*. Hoboken, NJ: John Wiley & Sons.
- Madan, D. B., and Unal, H. (2004). Risk-neutralizing statistical distributions: With an application to pricing reinsurance contracts on FDIC losses. Technical report 2004-01, FDIC, Center for Financial Research.
- Mittnik, S., and Rachev, S. T. (1993a). Modelling asset returns with alternative stable distributions. *Econometric Reviews* 12: 261–330.
- Mittnik, S., and Rachev, S. T. (1993b). Reply to comments on modelling asset returns with alternative stable distributions and some extensions. *Econometric Reviews* 12: 347–389.
- Moscadelli, M. (2004). The modelling of operational risk: Experience with the analysis of the data collected by the Basel Committee. Technical report, Bank of Italy.
- Müller, H. (2002). Quantifying operational risk in a financial institution. Master's thesis, Institut für Statistik und Wirtschaftstheorie, Universität Karlsruhe.
- Reynolds, D., and Syer, D. (2003). A general simulation framework for operational loss distributions. In C. Alexander (ed.), *Operational Risk*:

- Regulation, Analysis, and Management*. London: Prentice Hall.
- Rosenberg, J. V., and Schuermann, T. (2006). A general approach to integrated risk management with skewed, fat-tailed risks. *Journal of Financial Economics* 79, 3: 569–614.
- Ross, S. M. (2001). *Simulation*, 3rd ed. Boston, MA: Academic Press.
- Ross, S. M. (2002). *Introduction to Probability Models*, 8th ed. Boston, MA: Academic Press.
- Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley & Sons.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Optimization Tools

Introduction to Stochastic Programming and Its Applications to Finance

KORAY D. SIMSEK, PhD

Associate Professor, Sabanci School of Management, Sabanci University

Abstract: Mathematical programming is one of a number of operations research techniques that employs mathematical optimization models to assist in decision making. Mathematical programming includes linear programming, integer programming, mixed-integer programming, nonlinear programming, stochastic programming, and goal programming. Mathematical programming models allow the decision maker to identify the “best” solution. This is in contrast to other mathematical tools that are in the arsenal of decision makers such as statistical models (which tell the decision maker what occurred in the past), forecasting models (which tell the decision maker what might happen in the future), and simulation models (which tell the decision maker what will happen under different conditions). The mean-variance model for portfolio selection as formulated by Markowitz is an example of an application of one type of mathematical programming (quadratic programming). However, in formulating optimization models in many applications in finance, decision makers need to take into consideration the uncertainty about the model’s parameters and the multiperiod nature of the problem faced. To deal with these situations, the technique of stochastic programming is employed.

The dynamic nature of financial decision making requires the use of tools that are capable of capturing the multiperiod nature inherent in problems faced by asset managers in portfolio selection decisions and financial managers in capital budgeting decisions. These tools should be understandable with adequate treatment of uncertainty. They should incorporate practical considerations, such as transaction costs in the case of asset managers. Stochastic programming bears these characteristics. In this entry, we discuss the basics of stochastic program-

ming, give a brief history, and emphasize its importance by comparing the approach to other tools used in finance.

WHAT IS STOCHASTIC PROGRAMMING?

Stochastic programming is nothing but a fancy name for the study of optimal decision making under uncertainty. As opposed to “deterministic,” the term “stochastic” implies that some of the parameters of the problem are random

(that is, not known with certainty); the term “programming” points to links with *mathematical programming* and *optimization algorithms*.

Uncertainty is almost always inherent in real-world decision problems (and even more so in financial planning). As an example, we may consider a bet whose outcome is determined by flipping a coin. In such problems, uncertainty of parameters may be due to the presence of uncertain events (e.g., a coin flip in the previous example) or simply due to lack of reliable data.

In the past, due to computational and informational limitations, optimal decision models were often formulated deterministically by replacing the uncertainties with expectations or best estimates. With contributions from many disciplines, including operations research, and improvements in the information technology (faster hardware and software), stochastic programming is rapidly developing today.

The main features of a stochastic program, which can be viewed as an optimal decision model with explicit consideration of uncertainties, are:

- Random parameters with known (or partially known) distributions.
- Several decision variables with many potential values.
- Discrete time periods for decisions.
- Use of expectations (or other functions of decision variables) for objectives.

The problem structure, constraints, and objectives (risk/reward) are modeled across time along with the uncertainty of events. Future uncertainty is modeled through generating scenarios over time. In other words, the realizations of the uncertain parameters may be (gradually) revealed after some or all of the decisions have been made. High-performance computers take advantage of sophisticated algorithms to determine the optimal decision that will take into account the future uncertainty. As the uncertainty is revealed after each stage, recourse decisions can be made in the light of new information.

The relative importance of these main features contrasts with other decision-making models, such as statistical decision theory, decision analysis, dynamic programming, Markov decision processes, and *stochastic control (SC)*. In contrast to statistical decision theory, stochastic programming has emphasized solution methods and analytical solution properties over procedures for constructing objectives and updating probabilities. Stochastic programs generally have higher dimensions (that is, larger problem size) than SC models, which put more emphasis on control rules and have more restrictive constraint assumptions.

We can see the first forms of decision models that involve uncertainty in the early days of the history of mathematical programming. Beale (1955), Dantzig (1955), and Charnes and Cooper (1959) were the first to propose linear programs with random parameters. Dantzig named his approach “linear programming under uncertainty,” whereas Charnes and Cooper called theirs “chance/probabilistically constrained programming.”

Subsequently, stochastic programming has become a major subfield of mathematical programming with several theoretical developments. For overviews of the literature including algorithms and applications, see Kall and Wallace (1994), Infanger (1994), Ermoliev and Wets (1988), Birge and Louveaux (1997), Wallace and Ziemba (2003), Prékopa (1995), Higle and Sen (1996), Wallace et al. (1996), Censor and Zenios (1997), Wets and Ziemba (1999), Dupačová et al. (2002), Birge et al. (2002), and Ruszczyński and Shapiro (2003).

In general, stochastic optimization models result in large-scale programs since they include a large number of scenarios to reflect all possible outcomes of future uncertainty. Therefore, efforts on the algorithmic developments focused on adaptations of large-scale linear programming methods for special classes of stochastic programs whose structures are exploitable. In other words, emphasis was placed on forming the deterministic equivalent program and

taking advantage of the structure of the resulting formulation. Dantzig and Madansky (1961) introduced Dantzig-Wolfe decomposition as a possible method. One of the most successful approaches has been the application of Benders' decomposition (Benders, 1962) method to stochastic programs, originally developed by Van Slyke and Wets (1969). Birge (1985) extended this idea for multistage stochastic programs. In general, these methods concentrate on linear models. The diagonal quadratic approximation (DQA) algorithm, originally developed for linear programs by Mulvey and Ruszczyński (1992), can handle both quadratic and general convex (or equivalently concave) objective functions with linear constraints, as shown in Berger et al. (1994). The progressive hedging algorithm, developed by Rockafellar and Wets (1991), dualizes the nonanticipativity constraints and, like DQA, iterates over scenarios to force these constraints to be equal. Unlike DQA, there is no quadratic penalty term and all scenarios are coordinated through a master processor. Mulvey and Vladimirou (1991) successfully implemented the progressive hedging algorithm in the context of stochastic networks.

Specialized software packages that employ these methods are much faster than general solvers. Combined with algebraic modeling languages, such as AMPL, these specialized stochastic programming solvers provide efficient means of tackling problems that involve high levels of uncertainty.

Stochastic Programming in Finance

Financial planning represents one of the major application areas of stochastic programming. In fact, it is a natural domain for stochastic programming, since risk needs to be incorporated into investment decisions (portfolio decisions and capital budgeting decisions) and the problem structure is amenable to algebraic constraints and relationships. Deterministic approximations would fail to see the big picture. For example, through stochastic programs, portfolio allocations that would opti-

mize an investor's risk level under several scenarios can be determined; by contrast, because they ignore risk, deterministic programs provide inadequate solutions. Static portfolio selection models, based on Markowitz's mean-variance model (1952), have been proposed in many cases; however, their implementations may result in significant transaction costs and mistimed liquidation of assets. Examples of application of stochastic programming in financial planning can be found in Ziemba and Vickson (1975) and Zenios (1992).

Within finance, stochastic programming applications have greatly increased in recent years, particularly in asset-liability management (ALM). Multistage stochastic programs take into account the dynamic aspects of ALM problems faced by institutional investors. Based on assumptions about the (joint) dynamics of risk factors that are usually described by stochastic processes, representative scenarios for investment strategies are generated. Transactions take place at discrete points in time over a finite planning horizon. Moreover, several constraints (e.g., liability considerations, liquidity restrictions, limits on risk exposure) can be taken into account.

Since multistage programs suffer from an exponential growth in problem size with respect to the number of periods under consideration, the first models for ALM that appeared in the early 1980s [see Kallberg et al. (1982), Kusy and Ziemba (1986)] were restricted to a two-stage structure due to computational limitations. Mulvey and Vladimirou (1992) looked at optimal investment strategies given liabilities in a network environment. At Fannie Mae, Holmer (1994) implemented a system to minimize investment risk while taking into account that firm's retained mortgage portfolio. Advances in computing power, paired with efficient algorithms that are specialized for stochastic programming, help researchers implement and solve very large scale stochastic programs, such as the pension fund model of Gondzio and Kouwenberg (2001), with millions of scenarios.

One of the first successful commercial multistage stochastic programming applications is the Russell-Yasuda Kasai model (see Cariño et al., 1994, 1998; and Cariño and Ziemba, 1998). The model was employed to optimize the investment decisions for a Japanese insurance company over time where investment returns and liabilities are uncertain. The problem is complicated by constraints to meet the random liabilities and legal restrictions on the use of income in Japan.

Other successful commercial applications include the Towers Perrin-Tillinghast ALM system of Mulvey et al. (2000), the fixed income portfolio management models of Zenios (1995) and Beltratti et al. (1999), and the InnoALM system of Geyer and Ziemba (2008). A good number of applications in ALM are provided in Ziemba and Mulvey (1998), Ziemba (2003), and Zenios and Ziemba (2006).

Among other areas in finance, capital budgeting and fixed income portfolio management have been researched extensively using stochastic programming methods. For the former, Lockett and Gear (1975), De et al. (1982), and Turney (1990) are the earliest applications. Bradley and Crane (1972) were the first to propose stochastic programming for bond portfolio management. Zenios and Kang (1993) developed a portfolio immunization strategy in a multi-period stochastic optimization framework. Granville et al. (1994) describe a dual method for an asset-only allocation problem. Many other applications in the fixed income literature exist, including Hiller and Eckstein (1993) and Golub et al. (1995).

STOCHASTIC PROGRAMMING VERSUS OTHER METHODS IN FINANCE

In this section, we compare stochastic programming with other methods applied to financial planning (especially to ALM). First,

we highlight the dynamic aspects of stochastic programming and show its differences with *static models*. Afterward, we briefly discuss *continuous-time models* in finance and compare these models with stochastic programming—a discrete-time approach.

Static versus Dynamic Models in Financial Planning

The most well-known static model for financial planning is, without a doubt, the mean-variance model of Markowitz (1952). In this framework, the minimum-variance portfolio that satisfies a required expected return defines the optimal portfolio. Mulvey (1989) extended this model to account for liabilities by replacing the return measure with the surplus (defined as assets minus liabilities). Others have introduced downside risk measures (e.g., conditional value-at-risk, semivariance, mean absolute deviation, to name a few) to replace variance, recognizing the fact that variance is not a good risk measure for most asset classes (such as derivatives and fixed income securities) and for long-term investors. (See, e.g., Fishburn, 1977, Worzel et al., 1994, and Rockafellar and Uryasev, 2000.) Despite being computationally attractive, static models are inappropriate for long-term investors facing sequential decisions. Single-period models, unlike *dynamic models*, fail to cope with the dynamic aspects of the problem, such as transaction costs.

Among other static models, duration-matching models seem to be interesting, especially for ALM. These models seek to protect the surplus against an interest rate uncertainty. The optimal portfolio of assets is the lowest-cost portfolio whose value and duration are equal to those of liabilities. Applications can be quite successful in certain cases, for example, when a defined benefit pension plan has been terminated and taken over by an insurance company or when the transaction costs are low. However, these models ignore the facts that individual cash inflows and outflows are not matched and that one needs to adjust to the

changes in duration at every stage (which leads to high transaction costs). Therefore, computational and structural advantages of these models are insufficient to justify their drawbacks.

Dynamic models, in contrast, provide substantial flexibility to address the issues faced by long-term investors. They are not as easy to solve and conceptualize as static models; however, as discussed earlier in this entry, the advances in technical aspects of stochastic programming and today's computational power more than make up for these incapacities. Among these methods, dynamic programming seems to be especially interesting from an ALM perspective, as the optimal decisions are obtained in feedback form. However, it suffers from *the curse of dimensionality* as the planning horizon or the uncertainty representation is extended. An alternative method to overcome this problem is to specify a decision rule within the same framework, which also helps handle the transaction costs more easily. Nevertheless, incorporating decision rules leads to nonconvex optimization models (see Mulvey and Simsek, 2002). Fleten et al. (2002) illustrate the superior performance of dynamic models over static models. In the next section, we discuss the two major types of dynamic models.

Continuous-Time Models versus Stochastic Programming

Continuous-time models were introduced to the finance literature by Merton (1969). The variables that define the states of the world are modeled through *stochastic differential equations* (SDEs). Asset prices also follow SDEs whose parameters may be state and/or time dependent. Trading is assumed to occur continuously. Under additional assumptions on investors' preferences (that is, utility functions) and the structure of the economy, an explicit analytical solution can be found for these models by SC techniques. Thus, they provide better insights than the stochastic programming solutions, which are hard to generalize. However,

as Cochrane (2001, p. 28) suggests: "... in the complexity of most practical situations, one often ends up resorting to numerical simulation of a discretized model anyway."

Although some of the SC recommendations are implementable, the model simplifications may render them ineffective. As these models cannot incorporate complex constraints imposed by realistic situations and most investors (e.g., pension funds) do not want to trade continuously, we turn to stochastic programming, which allows decisions to be made at a finite number of discrete points in time.

In most cases, stochastic programming models require the uncertainties be approximated by a scenario tree with a finite number of states of the world at each time. As Kouwenberg and Zenios (2006, p. 291) suggest: "... important practical issues such as transaction costs, multiple state variables, market incompleteness, taxes and trading limits, regulatory restrictions, and corporate policy requirements can be handled *simultaneously* within the framework." This huge practical advantage, unfortunately, comes at a significant cost: curse of dimensionality. As analytical solutions are not possible, stochastic programming models need to be solved via numerical optimization. The model size explodes as the size of the state space or the number of decision stages increases. In recent years, this drawback has been substantially overcome through the development of new algorithms and the advances in computing power. Still, one should be careful about incorporating too much detail into a stochastic programming model, not because of the computational disadvantages but mainly to avoid confusing the decision maker, since SP solutions are hard to generalize.

It is, however, interesting to note that the continuous-time models have been the focus of research in the financial economics literature, whereas models in the operation research literature are mostly stated in discrete time. As Berger (1995) points out, there have been several successful applications of SC, such as the

Black-Scholes option pricing formula (Black and Scholes, 1973) and the continuous-time capital asset pricing model (Merton, 1973). See also Constantinides (1986), Dumas and Luciano (1991), and Shreve and Soner (1991) for SC applications with practical considerations such as transaction costs.

A GENERAL MULTISTAGE STOCHASTIC PROGRAMMING MODEL FOR FINANCIAL PLANNING

To illustrate the use of stochastic programming, we provide in this section a multistage stochastic program to tackle a long-term investment problem. We formulate the deterministic equivalent of the stochastic program and we discuss the issue of modeling the uncertain parameters on *scenario generation* methods.

Model Formulation

Here, we define the multiperiod investment problem as a multistage stochastic program. The basic model is a variant of Mulvey et al. (1997), with special attention to transaction costs.

To define the model, we divide the entire planning horizon T into two discrete time intervals T_1 and T_2 , where $T_1 = \{0, 1, \dots, \tau\}$ and $T_2 = \{\tau + 1, \dots, T\}$. The former corresponds to periods in which investment decisions are made. Period τ defines the end of the planning horizon. We focus on the investor's position at the beginning of period τ . Decisions occur at the beginning of each time stage. Much flexibility exists. An active trader might see his time interval as short as minutes, whereas a pension plan adviser will be more concerned with much longer planning periods such as the dates between the annual board of directors' meetings. It is possible for the steps to vary over time—short intervals at the beginning of the planning period and longer intervals toward the end. T_2 handles the horizon at time τ by calculating economic

and other factors beyond period τ up to period T . The investor renders passive decisions after the end of period τ .

Asset classes are defined by set $A = \{1, 2, \dots, I\}$, with category 1 representing cash. The remaining asset classes can include growth and value stocks, bonds, real estate, hedge funds, or private equity. The asset classes should track well-defined market segments. Ideally, the co-movements between pairs of asset class returns would be relatively low so that diversification can be done across the asset classes.

In multiperiod models, uncertainty is represented by a set of distinct realizations, called *scenarios*, $s \in S$. The scenarios may reveal identical values for the uncertain quantities up to a certain period; that is, they share common information history up to that period. We address the representation of the information structure through nonanticipativity constraints, which require that variables sharing a common history, up to time period t , must be set equal to each other (see equation (7) below).

We assume that the portfolio is rebalanced at the beginning of each period. Alternatively, we could simply make no transaction except to reinvest any dividend and interest—a buy-and-hold strategy. For convenience, we also assume that the cash flows are reinvested in the generating asset class and all the borrowing (if any) is done on a single-period basis.

For each $i \in A$, $t \in T_1$, and $s \in S$, we define the following parameters and decision variables.

Parameters

- $r_{i,t}^s$ = $1 + \rho_{i,t}^s$, where $\rho_{i,t}^s$ is the percent return for asset i , in time period t , under scenario s (projected by a stochastic scenario generator, for example, see Mulvey et al. [2000]).
- π_s Probability that scenario s occurs, $\sum_{s \in S} \pi_s = 1$.
- w_0 Wealth at the beginning of time period 0.
- $\sigma_{i,t}$ Transaction costs incurred in rebalancing asset i at the beginning of period t (symmetric transaction costs are assumed, that is, cost of selling equals cost of buying).
- β_i^s Borrowing rate in period t , under scenario s .

Decision Variables

- $x_{i,t}^s$ Amount of money in asset class i , at the beginning of time period t , under scenario s , after rebalancing.
- $v_{i,t}^s$ Amount of money in asset class i , at the beginning of time period t , under scenario s , before rebalancing.
- w_t^s Total wealth at the beginning of time period t , under scenario s .
- $p_{i,t}^s$ Amount of asset i purchased for rebalancing in period t , under scenario s .
- $d_{i,t}^s$ Amount of asset i sold for rebalancing in period t , under scenario s .
- b_t^s Amount of money borrowed at the beginning of period t , under scenario s .

$$v_{i,t}^s = r_{i,t-1}^s x_{i,t-1}^s \quad \forall s \in S, t = 1, \dots, \tau, i \in A \quad (4)$$

$$x_{i,t}^s = v_{i,t}^s + p_{i,t}^s (1 - \sigma_{i,t}) - d_{i,t}^s \quad \forall s \in S, i \in A/\{1\}, t = 1, \dots, \tau \quad (5)$$

$$x_{1,t}^s = v_{1,t}^s + \sum_{i \neq 1} d_{i,t}^s (1 - \sigma_{i,t}) - \sum_{i \neq 1} p_{i,t}^s - b_{t-1}^s (1 + \beta_{t-1}^s) + b_t^s \quad \forall s \in S, t = 1, \dots, \tau \quad (6)$$

$$x_{i,t}^s = x_{i,t}^{s'} \quad \forall s \text{ and } s' \text{ with identical past up to time } t, t = 1, \dots, \tau \quad (7)$$

Given these definitions, we present the deterministic equivalent of the stochastic asset-only allocation problem.

$$\text{Max EU}(w_\tau^s) = \sum_{s \in S} \pi_s U(w_\tau^s) \quad (1)$$

subject to

$$\sum_{i \in A} x_{i,0}^s = w_0 \quad \forall s \in S \quad (2)$$

$$\sum_{i \in A} x_{i,\tau}^s = w_\tau \quad \forall s \in S \quad (3)$$

A generalized network investment model is presented in Figure 1. This figure depicts the flows across time for each of the asset classes. While all constraints cannot be put into a network model, the graphical form is easy for asset managers to comprehend. General linear and nonlinear programs are now readily available for solving the resulting problem. However, a network may have computational advantages for extremely large problems, such as security level models.

As with single-period models, the nonlinear objective function (1) can take several

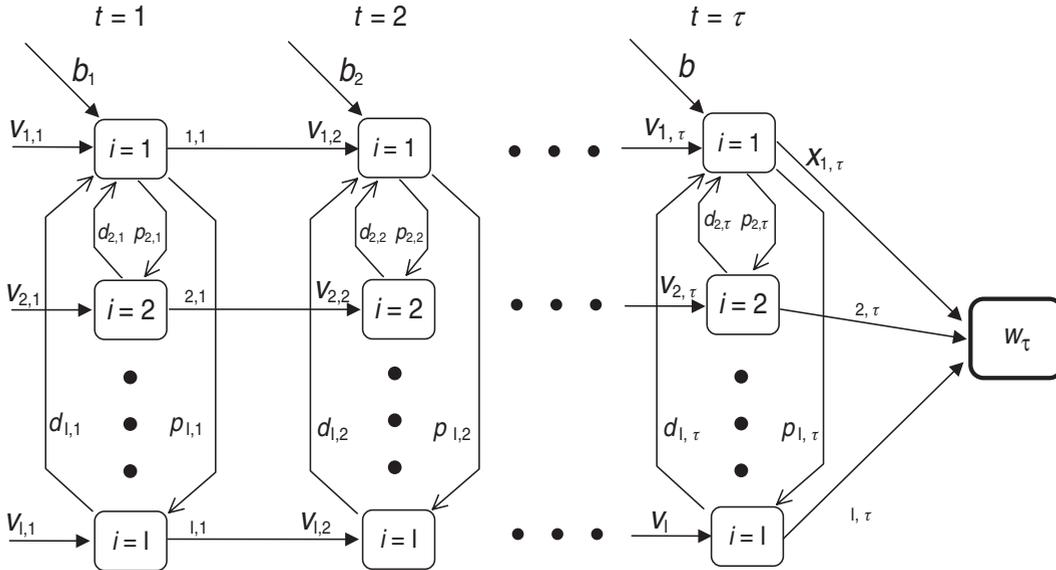


Figure 1 Network Representation for Each Scenario, $s \in S$

different forms. If the classical return-risk function is employed, then (1) becomes $\text{Max } Z = \eta \times \text{Mean}(w_\tau) - (1 - \eta) \times \text{Risk}(w_\tau)$, where $\text{Mean}(w_\tau)$ is the expected total wealth and $\text{Risk}(w_\tau)$ is the risk of the total wealth across the scenarios at the beginning of period τ . Parameter η indicates the relative importance of risk as compared with the expected value. This objective leads to an efficient frontier at period τ by allowing alternative values of η in the range $[0,1]$. It can be shown that a viable alternative to the mean-risk framework is the von Neumann-Morgenstern expected utility of wealth at the beginning of period τ .

Let's review the six constraints:

1. Constraint (2) guarantees that the total initial investment equals the initial wealth.
2. Constraint (3) represents the total wealth in the beginning of period τ . This constraint can be modified to include assets, liabilities, and investment goals, in which case the modified result is referred to as the "surplus wealth" (Mulvey, 1989). Many investors render investment decisions without reference to their liabilities or investment goals. Mulvey (1989) incorporates the notion of surplus wealth into the mean-variance and the expected utility models to address liabilities in the context of asset allocation strategies.
3. Constraint (4) depicts the wealth $v_{i,t}^s$ accumulated at the beginning of period t before rebalancing in asset i .
4. Constraint (5) gives the flow balance constraint for all assets except cash for all periods. This constraint guarantees that the amount invested in period t equals the net wealth for the asset.
5. Constraint (6) represents the flow balancing constraint for cash.
6. Constraints (7) are the nonanticipativity constraints.

The preceding constraints ensure that the scenarios with the same past will have identical decisions up to that period. While these constraints are numerous, solution algorithms take

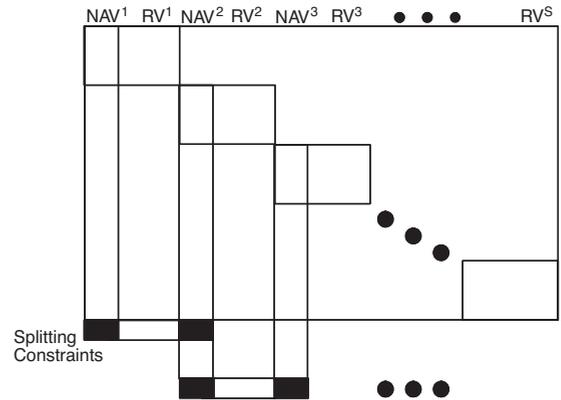


Figure 2 Split Variable Formulation (NAV: Nonanticipativity Variables, RV: Remaining Variables)

advantage of their simple structure. See Birge and Louveaux (1997), Dantzig and Infanger (1993), Kall and Wallace (1994), and Mulvey and Ruszczyński (1995), for example.

Figure 2 depicts the constraint structure for a split variable formulation of the stochastic asset allocation problem. This formulation has proven successful for solving the model using techniques such as the progressive hedging algorithm of Rockafellar and Wets (1991) and the DQA algorithm by Mulvey and Ruszczyński (1995). The split variable formulation can be beneficial for direct solvers that use the interior point method. Given today's powerful PCs, the nonlinear optimization system can be solved in a direct fashion for realistic-size implementations.

By substituting constraint (7) back into constraints (2) to (6), we obtain a standard form of the stochastic allocation problem. Constraints for this formulation exhibit a dual block diagonal structure for two-stage stochastic programs and a nested structure for general multi-stage problems. This formulation may be better for some direct solvers. The standard form of the stochastic program possesses fewer decision variables than the split variable model and is the preferred structure by many researchers in the field. This model can be solved by means of decomposition methods, for example, the

L-shaped method (a specialization of Benders' algorithm). (See Birge and Louveaux, 1997; Consigli and Dempster, 1998; Dantzig and Infanger, 1993; and Kouwenberg and Zenios, 2006.)

As shown by Consigli and Dempster (1998), Dantzig and Infanger (1993), Mulvey et al. (2000), Ziemba and Mulvey (1998), and Ziemba (2003), a multistage model can provide superior performance over single-period models.

Modeling Future Uncertainties (Scenario Generation)

To model future uncertainty in our financial planning problem, we utilize a representative set of scenarios. In this section, we review the procedures for scenario generation and give details about the approach described.

In most cases, stochastic programming models require that the future uncertainties are approximated by a scenario tree with a finite number of states of the world at each time. The planning horizon is divided into T time periods (generally years for pension planning).

In most cases, stochastic programming models require that the future uncertainties are approximated by a scenario tree with a finite number of states of the world at each time. The planning horizon is divided into T time periods (generally years for pension planning). A sample scenario tree of three periods and nine scenarios is depicted in Figure 3. The root of the tree represents the current state of the world. A scenario is defined as a single branch from the root to any leaf of the tree (e.g., the boldfaced path corresponds to scenario 4). Thus, all of the parameter uncertainties are depicted along this branch. Each node represents a state of the world under a given scenario at a given time; for instance, the boldfaced node corresponds to the set of uncertainties at the end of period 2 under scenario 4. The stochastic program will determine an optimal decision for each node of the scenario tree, given the information available at that point. As there are multiple succeeding nodes, the optimal decisions will be determined without exploiting hindsight. A stochastic programming model will find the optimal policy that will fit the current state of the world and the decision maker in each node, while anticipating the op-

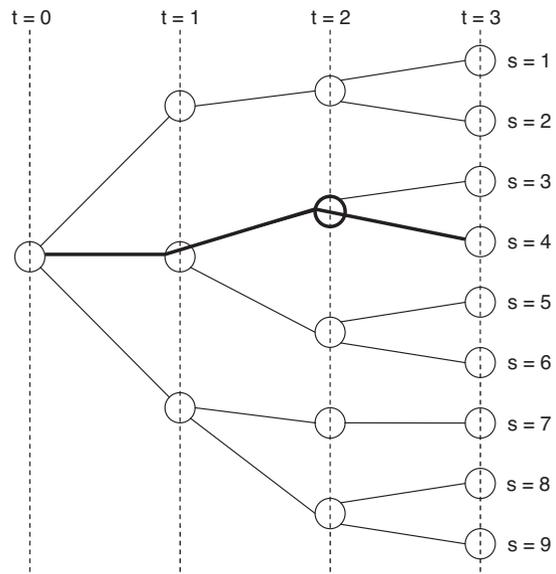


Figure 3 A Three-Period Scenario Tree

timal adjustment of the policy later on as the tree evolves and more information is revealed.

Generating scenario trees to represent the evolution of future uncertainty is a two-step process. Figure 4 depicts a diagram of the process.

The first step involves the construction of a stochastic forecasting model. This involves choosing a model that would be appropriate for the uncertain variables and calibrating the parameters of this model using historical data.

The simplest approach, bootstrapping historical data, eliminates the need for a mathematical model (see, e.g., Grauer and Hakansson, 1982). Among mathematical models, stochastic differential equations and time series analysis are two commonly used techniques to generate anticipatory scenarios. Our preference is to employ the former technique, in which a

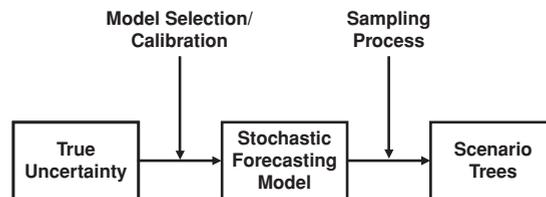


Figure 4 From Historical Data to Scenario Trees

sequence of economic factors (e.g., gross domestic product, corporate earnings, interest rates, and inflation) drive the state variables (see Mulvey [1996] for the details). The parameters of the SDEs for the economic factors and asset returns can be calibrated using historical data. A standard variance reduction method, antithetic variates, can be employed to improve the accuracy of the model's recommendations. Indirect inference methods for calibrating the parameters of the resulting stochastic system can be employed (see Gouriéroux et al., 1993).

The next step involves the discretization of the scenarios generated by the stochastic forecasting model in the first step. To avoid any computational disadvantages, this has to be done using a small number of nodes, which in turn will lead to approximation errors. There are several methods to achieve this depending on the models employed in the first step (see Hoyland and Wallace, 2001; Kouwenberg and Zenios, 2006; Grebeck, Rachev, and Fabozzi, 2009; and Ziemba, 2003). We first create discretized sample paths by moment-matching, using the cascaded SDE structure in the first step. Then, we convert these sample paths to a scenario tree by clustering (see Dupáčová et al., 2002). We begin by grouping similar first stage values of the sample paths into clusters, and then continue sequentially through each stage.

For an ALM system, one needs to generate scenarios for the liability side as well as the asset side. Obviously, both components are driven by economic factors. Liabilities are affected by actuarial predictions as well. When modeling the asset returns, one may need to use sentiment or expert judgment to improve the range of scenarios.

The future value of the liabilities can be especially tricky to project for institutions, such as pension plans, where the liabilities consist of several contracts and therefore the valuation is affected by various sources of uncertainty. For a typical pension plan, one can simulate the future status of the participants by making assumptions about the retirement rates, resigna-

tion frequency, promotion/demotion probabilities, and the mortality rate. Once this is done, the interest rates are forecasted and used to calculate the present value of the liabilities.

When modeling the asset returns, the economic factors that drive the primary asset-class returns are projected as a first step, which would then be followed by the projection of returns for these primary assets. More complex assets would be the last to be modeled in this setup. Alternatively, one can model all uncertain variables at once through one big set of multivariate time-series models.

KEY POINTS

- Stochastic programming is an operations research method for optimal decision making under uncertainty and bears suitable characteristics for modeling and solving financial planning applications, such as asset-liability management, capital budgeting, and fixed income portfolio management.
- The main features of a stochastic program are: random parameters with known (or partially known) distributions; several decision variables with many potential values; multiple discrete time periods for decisions; use of expectations (or other functions of decision variables) for objectives.
- Stochastic programs typically have larger problem size than stochastic control models, which put more emphasis on control rules and have more restrictive constraint assumptions.
- Stochastic programming models generally require that the future uncertainties are approximated by a scenario tree with a finite number of states of the world at each time.
- Multiperiod stochastic programs with a large number of parameters and scenarios result in large-scale deterministic-equivalent programs. Specialized software packages combined with algebraic modeling languages are utilized to efficiently tackle these problems.

REFERENCES

- Beale, E. M. L. (1955). On minimizing a convex function subject to linear inequalities. *Journal of Royal Statistical Society, Series B* 17: 173–184.
- Beltratti, A., Consiglio, A., and Zenios, S. A. (1999). Scenario modeling for the management of international bond portfolios. *Annals of Operations Research* 85: 227–247.
- Benders, J. F. (1962). Partitioning procedures for solving mixed variables programming problems. *Numerische Mathematik* 4: 238–252.
- Berger, A. J. (1995). Large Scale Stochastic Optimization with Applications to Finance. PhD Thesis, Princeton University.
- Berger, A. J., Mulvey, J. M., and Ruszczyński, A. (1994). An extension of the DQA algorithm to convex stochastic programs. *SIAM Journal on Optimization* 4: 735–753.
- Birge, J. R. (1985). Decomposition and partitioning methods for multistage stochastic linear programming. *Operations Research* 33: 989–1007.
- Birge, J. R., Edirisinghe, N. C. P., and Ziemba, W. T. (eds.). (2002). *Research in Stochastic Programming* (Special Issue *Annals of Operations Research*). Baltzer Science Publishers BV.
- Birge, J. R., and Louveaux, F. (1997). *Introduction to Stochastic Programming*. New York: Springer.
- Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.
- Bradley, S. P., and Crane, D. B. (1972). A dynamic model for bond portfolio management. *Management Science* 19, 2: 139–151.
- Cariño, D. R., Kent, T., Myers, D. H., et al. (1994). The Russell-Yasuda Kasai model: An asset/liability model for a Japanese insurance company using multistage stochastic programming. *Interfaces* 24: 29–49.
- Cariño, D. R., Myers, D. H., and Ziemba, W. T. (1998). Concepts, technical issues, and uses of the Russell-Yasuda Kasai financial planning model. *Operations Research* 46: 450–462.
- Cariño, D. R., and Ziemba, W. T. (1998). Formulation of the Russell-Yasuda Kasai financial planning model. *Operations Research* 46: 433–449.
- Censor, Y., and Zenios, S. A. (1997). *Optimization: Theory, Algorithms and Applications*. Series on Numerical Mathematics and Scientific Computation. New York: Oxford University Press.
- Charnes, A., and Cooper, W. W. (1959). Chance-constrained programming. *Management Science* 5: 73–79.
- Cochrane, J. H. (2001). *Asset Pricing*. Princeton, NJ: Princeton University Press.
- Consigli, G., and Dempster, M. A. H. (1998). Dynamic stochastic programming for asset-liability management. *Annals of Operations Research* 81: 131–161.
- Constantinides, G. M. (1986). Capital market equilibrium with transaction costs. *Journal of Political Economy* 94: 842–862.
- Dantzig, G. B. (1955). Linear programming under uncertainty. *Management Science* 1: 197–206.
- Dantzig, G. B., and Infanger, G. (1993). Multistage stochastic linear programs for portfolio optimization. *Annals of Operations Research* 45: 59–76.
- Dantzig, G. B., and Madansky, A. (1961). On the solution of two-stage linear programs under uncertainty. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press.
- De, K., Acharya, D., and Sahu, K. C. (1982). A chance-constrained goal programming model for capital budgeting. *Journal of the Operational Research Society* 33: 635–638.
- Dumas, B., and Luciano, E. (1991). An exact solution to a dynamic portfolio choice problem under transaction costs. *Journal of Finance* 46: 577–595.
- Dupačová, J., Hurt, J., and Stepan, J. (2002). *Stochastic Modeling in Economics and Finance*. Dordrecht, Netherlands: Kluwer Academic Publishers.
- Ermoliev, Y., and Wets, R. J-B. (eds.). (1988). *Numerical Methods for Stochastic Optimization*. Berlin: Springer-Verlag.
- Fishburn, P. C. (1977). Mean-variance risk analysis with risk associated with below target returns. *American Economic Review* 67: 116–126.
- Fleten, S.-E., Hoyland, K., and Wallace, S. W. (2002). The performance of stochastic dynamic and fixed mix portfolio models. *European Journal of Operational Research* 140, 1: 37–49.
- Geyer, A., and Ziemba, W. T. (2008). The Innovest Austrian pension fund financial planning model InnoALM. *Operations Research* 56: 797–810.
- Golub, B., Holmer, M., McKendall, R., Pohlman, L., and Zenios, S. A. (1995). A stochastic programming model for money management. *European Journal of Operational Research* 85: 282–296.
- Gondzio, J., and Kouwenberg, R. (2001). High-performance computing for asset-liability management. *Operations Research* 49: 879–891.
- Gourieroux, C., Monfort, A., and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics* 8: S85–S118.

- Grauer, R. R., and Hakansson, N. H. (1982). Higher return, lower risk: Historical returns on long-run, actively managed portfolios of stocks, bonds and bills, 1936–78. *Financial Analysts Journal* 38: 39–53.
- Granville, S., Pereira, M. V. F., and McCoy, M. (1994). Stochastic dual dynamic programming applied to financial planning problems. Technical Report PSR1, R-Alberto de Campos, Rio de Janeiro, Brazil.
- Grebeck, M. J., Rachev, S.T., and Fabozzi, F. J. (2009). Stochastic programming and stable distributions in asset liability management. *Journal of Risk* 12, 2: 29–47.
- Higle, J. L., and Sen, S. (1996). *Stochastic Decomposition: A Statistical Method for Large Scale Stochastic Programming*. Dordrecht, Netherlands: Kluwer Academic Publishers.
- Hiller, R. S., and Eckstein, J. (1993). Stochastic dedication: Designing fixed income portfolios using massively parallel Benders decomposition. *Management Science* 39, 11: 1422–1438.
- Holmer, M. R. (1994). The asset-liability management strategy system at Fannie Mae. *Interfaces* 24: 3–21.
- Hoyland, K., and Wallace, S. (2001). Generating scenario trees for multistage problems. *Management Science* 47: 295–307.
- Infanger, G. (1994). *Planning under Uncertainty*. Danvers, MA: Boyd and Fraser.
- Kall, P., and Wallace, S. W. (1994). *Stochastic Programming*. New York: John Wiley & Sons.
- Kallberg, J., White, R., and Ziemba, W. T. (1982). Short term financial planning under uncertainty. *Management Science* 28: 670–682.
- Kouwenberg, R., and Zenios, S. A. (2006). Stochastic programming models for asset liability management. In S. A. Zenios and W. T. Ziemba (eds.), *Handbook of Asset and Liability Management: Theory and Methodology* (pp. 253–303). Amsterdam: North-Holland.
- Kusy, M. I., and Ziemba, W. T. (1986). A bank asset and liability management model. *Operations Research* 34: 356–376.
- Lockett, A. G., and Gear, A. E. (1975). Multistage capital budgeting under uncertainty. *Journal of Financial and Quantitative Analysis* 10, 1: 21–36.
- Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance* 7: 77–91.
- Merton, R. C. (1969). Lifetime portfolio selection under uncertainty: The continuous time case. *Review of Economics and Statistics* 51: 247–257.
- Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica* 41: 867–887.
- Mulvey, J. M. (1989). A surplus optimization perspective. *Investment Management Review* 3: 31–39.
- Mulvey, J. M. (1996). Generating scenarios for the Towers Perrin investment system. *Interfaces* 26: 1–15.
- Mulvey, J. M., Gould, G., and Morgan, C. (2000). The asset and liability management system for Towers Perrin-Tillinghast. *Interfaces* 30: 96–114.
- Mulvey, J. M., Rosenbaum, D. P., and Shetty, B. (1997). Strategic financial risk management and operations research. *European Journal of Operational Research* 97: 1–16.
- Mulvey, J. M., and Ruszczyński, A. (1992). A diagonal quadratic approximation method for large scale linear programs. *Operations Research Letters* 12: 205–215.
- Mulvey, J. M., and Ruszczyński, A. (1995). A new scenario decomposition method for large-scale stochastic optimization. *Operations Research* 43: 477–490.
- Mulvey, J. M., and Simsek, K. D. (2002). Rebalancing strategies for long-term investors. In E. J. Kontoghiorghes, B. Rustem, and S. Siokos (eds.), *Computational Methods in Decision-Making, Economics and Finance: Optimization Models* (pp. 15–33). Netherlands: Kluwer Academic Publishers.
- Mulvey, J. M., and Vladimirov, H. (1991). Solving multistage stochastic networks: An application of scenario aggregation. *Networks* 21: 619–643.
- Mulvey, J. M., and Vladimirov, H. (1992). Stochastic network programming for financial planning problems. *Management Science* 38: 1642–1664.
- Prékopa, A. (1995). *Stochastic Programming*. Dordrecht, Netherlands: Kluwer Academic Publishers.
- Rockafellar, R. T., and Uryasev, S. (2000). Optimization of conditional Value-at-Risk. *The Journal of Risk* 2, 3: 21–41.
- Rockafellar, R. T., and Wets, R. J.-B. (1991). Scenario and policy aggregation in optimization under uncertainty. *Mathematics of Operations Research* 16: 119–147.
- Ruszczynski, A., and Shapiro, A. (eds.). (2003). Stochastic programming. In *Handbooks in Operations Research and Management Science*, Volume 10. Amsterdam: North-Holland.
- Shreve, S. E., and Soner, H. M. (1991). Optimal investment and consumption with two bonds

- and transaction costs. *Mathematical Finance* 1: 53–84.
- Turney, S. T. (1990). Deterministic and stochastic dynamic adjustment of capital investment budgets. *Mathematical Computation and Modelling* 13, 5: 1–9.
- Van Slyke, R. M., and Wets, R. J-B. (1969). L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM Journal on Applied Mathematics* 17: 638–663.
- Wallace, S., Higle, J., and Sen, S. (eds.). (1996). *Stochastic Programming, Algorithms and Models* (Special Issue Annals of Operations Research). Baltzer Science Publishers BV.
- Wallace, S. W., and Ziemba, W. T. (eds.). (2003). *Applications of Stochastic Programming*, SIAM Mathematical Series on Optimization.
- Wets, R. J-B., and Ziemba, W. T. (eds.). (1999). *Stochastic Programming—State of the Art* (Special Issue Annals of Operations Research). Baltzer Science Publishers BV.
- Worzel, K. J., Vassiadou-Zeniou, C., and Zenios, S. A. (1994). Integrated simulation and optimization models for tracking fixed-income indices. *Operations Research* 42: 223–233.
- Zenios, S. A. (1992). *Financial Optimization*. Cambridge, UK: Cambridge University Press.
- Zenios, S. A. (1995). Asset/liability management under uncertainty for fixed-income securities. *Annals of Operations Research* 59: 77–97.
- Zenios, S. A., and Kang, P. (1993). Mean-absolute deviation and portfolio optimization for mortgage-backed securities. *Annals of Operations Research* 45: 433–450.
- Zenios, S. A., and Ziemba, W. T. (2006). *Handbook of Asset and Liability Management: Theory and Methodology*. Amsterdam: North-Holland.
- Ziemba, W. T. (2003). *The Stochastic Programming Approach to Asset-Liability and Wealth Management*. AIMR-Blackwell.
- Ziemba, W. T., and Mulvey, J. (eds.). (1998). *Worldwide Asset and Liability Modeling*. Cambridge, UK: Cambridge University Press.
- Ziemba, W. T., and Vickson, R. G. (1975). *Stochastic Optimization Models in Finance*. New York: Academic Press.

Robust Portfolio Optimization

DESSISLAVA A. PACHAMANOVA, PhD

Associate Professor of Operations Research, Babson College

PETTER N. KOLM, PhD

Clinical Associate Professor and Director of the Mathematics in Finance Masters Program,
Courant Institute, New York University

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

SERGIO M. FOCARDI, PhD

Partner, The Intertek Group

Abstract: As the use of quantitative techniques has become more widespread in the investment industry, the issue of how to handle portfolio estimation and model risk has grown in importance. Robust optimization is a technique for incorporating estimation errors directly into the portfolio optimization process, and is typically applied in conjunction with robust statistical estimation methods. The robust optimization approach uses the distribution from the estimation process to find a portfolio allocation in one single optimization, while keeping the computational costs low. Robust portfolios tend to be less sensitive to estimation errors, offer some improved portfolio performance, and often have lower turnover ratios.

The concepts of *portfolio optimization* and diversification have been instrumental in the understanding of financial markets and the development of financial decision making. The major breakthrough came in 1952 with the publication of Harry Markowitz's theory of portfolio selection. Markowitz suggested that sound financial decision making is a quantitative trade-off between risk and return. His work spurred a vast amount of research on quantifying market behavior, and one of the main practical consequences of his theory was the

acceptance of the notion that diversification reduces portfolio risk.

Sixty years after Markowitz's seminal work, substantial advances have been made in the theory and practice of portfolio management. Today, quantitative techniques for forecasting asset returns, portfolio allocation, risk measurement, trading and rebalancing, to mention a few, have a major presence in the financial industry. Their proliferation has been facilitated by the decreased cost of computing power and the increased availability of

sophisticated and specialized software that allows investors to incorporate their forecasts about the future direction of markets into disciplined analytical frameworks.

As the use of quantitative techniques has become widespread in the investment industry, the consideration of estimation risk and model risk has grown in importance. For example, Bayesian techniques and robust estimation of model parameters are now common in financial applications. Most recently, practitioners have begun incorporating the uncertainty introduced by estimation errors directly into the portfolio optimization process by mathematical techniques referred to as robust optimization. Contrary to the traditional approach, in which inputs to the portfolio allocation framework are treated as deterministic, robust portfolio optimization incorporates the notion that inputs have been estimated with errors. In this case, the inputs are not the traditional forecasts, such as expected returns and asset covariances, but rather *uncertainty sets* containing these point estimates (e.g., confidence intervals around the forecasts).

In this entry, we survey the area of robust optimization and its applications in portfolio management. We begin by explaining the main ideas behind the robust optimization approach, and discuss the relationship between robust optimization and other robust methods for portfolio management. Next, we review some important developments in robust optimization applications, and conclude with a discussion of future directions in *robust portfolio management*.

THE ROBUST OPTIMIZATION APPROACH

Introduced in the operations research literature by Ben-Tal and Nemirovski (1998) and El Ghaoui and Lebret (1997), modern robust optimization allows a portfolio manager to solve a robust formulation of the portfolio optimization problem with one single call to an optimization

solver in about the same time as the classical portfolio optimization problem. The resulting optimal portfolio allocations tend to be more stable and less sensitive to changes in model parameters.

Consider the classical mean-variance portfolio allocation problem:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \boldsymbol{\mu}'\mathbf{w} - \lambda \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{1} = 1 \end{aligned}$$

where $\boldsymbol{\mu}$ is the vector of expected returns (alphas) for N assets in the investment universe, $\boldsymbol{\Sigma}$ is the asset-asset covariance matrix, \mathbf{w} is the N -dimensional vector of portfolio weights, λ is the risk aversion coefficient, and $\mathbf{1}$ is a vector of ones. This optimization problem simply states that the optimal portfolio weights should be chosen so that the expected portfolio return less the portfolio risk (scaled by the risk aversion coefficient) is as large as possible. The equality constraint ensures that the portfolio weights add up to one.

As demonstrated, for instance, by Black and Litterman (1992), a small change in the expected asset returns can result in large changes in the optimal portfolio allocation. In other words, the classical portfolio optimization problem is not robust with respect to small changes in its inputs. Since in practice expected returns and asset covariances cannot be measured exactly but have to be estimated—sometimes with large errors—it is important in applications that uncertainty resulting from estimation errors be taken into account.

One way to make the optimization problem robust with respect to estimation errors is to require that the optimal solution remains optimal for all values of the expected returns that are “close” to the estimates of expected returns $\hat{\boldsymbol{\mu}}$. We can express this requirement in the optimization problem as follows: Instead of using the estimate $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$, we consider a set of vectors that are close to the estimate $\hat{\boldsymbol{\mu}}$, and solve the optimization problem for all vectors in this set. The idea here is that the expected returns may

have been estimated with some error, but that the estimates are not too far away from the true expected returns. Mathematically, this idea is incorporated in the definition of an uncertainty set for $\hat{\boldsymbol{\mu}}$,

$$U_\delta(\hat{\boldsymbol{\mu}}) = \{\boldsymbol{\mu} \mid |\mu_i - \hat{\mu}_i| \leq \delta_i, i = 1, \dots, N\} \quad (1)$$

In words, the set $U_\delta(\hat{\boldsymbol{\mu}})$ contains all vectors $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$ such that each component μ_i is in the interval $[\hat{\mu}_i - \delta_i, \hat{\mu}_i + \delta_i]$, and is often referred to as a “box” uncertainty set. From a statistical point of view, these intervals can be chosen to be certain confidence intervals around each point estimate $\hat{\mu}_i$.

We solve a modification of the original optimization problem such that even if $\boldsymbol{\mu}$ takes its worst possible value within the uncertainty set, the allocation remains optimal. Namely, we solve the max-min portfolio optimization problem

$$\begin{aligned} \max_{\mathbf{w}} \quad & \left\{ \min_{\boldsymbol{\mu} \in U_\delta(\hat{\boldsymbol{\mu}})} \{\boldsymbol{\mu}'\mathbf{w}\} - \lambda \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} \right\} \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{t} = 1 \end{aligned}$$

At first sight, this optimization problem looks complicated, as we have to minimize the objective function with respect to $\boldsymbol{\mu}$ over the specified uncertainty set and, simultaneously, maximize the objective function with respect to \mathbf{w} to find the optimal allocation. However, as we will see shortly, this problem can be reformulated into an equivalent maximization problem with respect to only \mathbf{w} . First, let us understand what this model implies from an intuitive perspective.

Observe that this model incorporates the notion of aversion to estimation error in the following sense. When the interval $[\hat{\mu}_i - \delta_i, \hat{\mu}_i + \delta_i]$ for the expected return of the i th asset is large, meaning that the expected return has been estimated with large estimation error, then the minimization problem over $\boldsymbol{\mu}$ is less constrained. Consequently, the minimum will be smaller than it would be in situations when the interval for $\hat{\mu}_i$ is smaller. Obviously, when

the interval is small enough, the minimization problem will be so tightly constrained that it would deliver a solution that is close to the optimal solution of the classical portfolio optimization problem in which estimation errors are ignored. In other words, it is the size of the intervals (in general, the size of the uncertainty set) that controls the aversion to the uncertainty that comes from estimation errors.

The robust version of the classical portfolio optimization problem is obtained by solving the max-min problem above, and for this model is easy to derive without any involved mathematics. Namely, it is

$$\begin{aligned} \max_{\mathbf{w}} \quad & \hat{\boldsymbol{\mu}}'\mathbf{w} - \delta'|\mathbf{w}| - \lambda \mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{t} = 1 \end{aligned}$$

where $|\mathbf{w}|$ denotes the absolute value of the entries of the vector of weights \mathbf{w} . To gain some intuition, notice that if the weight of asset i in the portfolio is negative, the worst-case expected return for asset i is $\hat{\mu}_i + \delta_i$ (we lose the largest amount possible). If the weight of asset i in the portfolio is positive, then the worst-case expected return for asset i is $\hat{\mu}_i - \delta_i$ (we gain the smallest amount possible). Observe that $\hat{\mu}_i w_i - \delta_i |w_i|$ equals $(\hat{\mu}_i - \delta_i) w_i$ if the weight w_i is positive and $(\hat{\mu}_i + \delta_i) w_i$ if the weight w_i is negative. Hence, the mathematical expression in the objective agrees with our intuition: It minimizes the worst-case expected portfolio return. In this robust version of the mean-variance formulation, assets whose mean return estimates are less accurate (have a larger estimation error δ_i) are therefore penalized in the objective function, and will tend to have a smaller weight in the optimal portfolio allocation.

This optimization problem has the same computational complexity as the nonrobust mean-variance formulation—namely, it can be stated as a quadratic optimization problem. The latter can be achieved by using a standard trick that allows us to get rid of the absolute values for the weights. The idea is to introduce an N -dimensional vector of additional variables $\boldsymbol{\psi}$ to replace the absolute values, and to write an

equivalent version of the optimization problem,

$$\begin{aligned} \max_{\mathbf{w}, \boldsymbol{\psi}} \quad & \hat{\boldsymbol{\mu}}' \mathbf{w} - \boldsymbol{\delta}' \boldsymbol{\psi} - \lambda \mathbf{w}' \boldsymbol{\Sigma} \mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}' \mathbf{1} = 1 \\ & \psi_i \geq w_i; \psi_i \geq -w_i, i = 1, \dots, N \end{aligned}$$

Therefore, incorporating considerations about the uncertainty in the estimates of the expected returns in this example has virtually no computational cost.

We can view the effect of this particular “robustification” of the mean-variance portfolio optimization formulation in two different ways. On the one hand, we see that the values of the expected returns for the different assets have been adjusted downwards in the objective function of the optimization problem. That is, the robust optimization model “shrinks” the expected return of assets with large estimation error. On the other hand, we can interpret the additional term in the objective function as a “risk-like” term that represents penalty for estimation error. The size of the penalty is determined by the investor’s aversion to estimation risk, and is reflected in the magnitude of the deltas.

More complicated specifications for uncertainty sets have more involved mathematical representations, but can still be selected so that they preserve an easy computational structure for the robust optimization problem. For example, a frequently used uncertainty set is

$$U_\delta(\hat{\boldsymbol{\mu}}) = \{ \boldsymbol{\mu} | (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})' \boldsymbol{\Sigma}_\mu^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}) \leq \delta^2 \} \quad (2)$$

where $\boldsymbol{\Sigma}_\mu$ is the covariance matrix of estimation errors for the vector of expected returns $\boldsymbol{\mu}$. This uncertainty set represents the requirement that the scaled sum of squares (scaled by the inverse of the covariance matrix of estimation errors) between all elements in the set and the point estimates $\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_N$ can be no larger than δ^2 . We note that this uncertainty set cannot be interpreted as individual confidence intervals around each point estimate. Instead, those familiar with statistics will notice that this uncertainty set captures the idea of a joint confidence region used, for example, in

Wald tests. In practice, the covariance matrix of estimation errors is often assumed to be diagonal. For this particular case, the set contains all vectors of expected returns that are within a certain number of standard deviations from the point estimate of the vector of expected returns, and the resulting robust portfolio optimization problem would protect the investor if the vector of expected returns is indeed within that range.

Selecting Uncertainty Sets from Statistical Procedures

How do we select uncertainty sets for a particular application? In practice, their shape and size are usually based on statistical estimates and probabilistic guarantees. For example, uncertainty set (1) defines an N -dimensional box: It considers possible deviations of the N uncertain parameters from their expected values, and the resulting robust portfolio optimization problem protects against the worst possible realization of each individual parameter separately. Uncertainty set (2) defines an N -dimensional ellipsoid (in two dimensions, an ellipsoid is an ellipse), and is not as conservative as (1). The resulting robust portfolio optimization offers protection from the worst possible joint deviation of the actual expected returns from the forecasts, by considering the correlations between the estimation errors of the uncertain parameters through the covariance matrix $\boldsymbol{\Sigma}_\mu$.

The calibration of the parameters that enter the definition of uncertainty sets is very important. For example, the intervals for $\hat{\boldsymbol{\mu}}$ that define the uncertainty set (1) above can be matched to 95% or 99% confidence intervals for the estimates of the expected returns. The value of the parameter δ in the uncertainty set (2) can be related to probabilistic guarantees, as we will explain later.

The covariance matrix $\boldsymbol{\Sigma}_\mu$ of the errors in the estimated expected asset returns in uncertainty set (2) can be obtained using several different techniques. However, its estimation

can be problematic because of the difficulty in separating the estimation error in expected returns from the inherent variability in actual realized returns (Lee, Stefek, and Zhelenyak, 2006). Specifically, if a portfolio manager forecasts a 5% active return over the next time period, but achieves 1%, he cannot argue that there was a 4% error in his expected return, so evaluating Σ_{μ} from historical data can be tricky.

In theory, if returns in a given sample of size T are assumed to come from a normal distribution, then Σ_{μ} equals $(1/T) \cdot \Sigma$, where Σ is the covariance matrix of asset returns as before. However, experience seems to suggest that this may not be the best method in practice. One issue is that this approach applies only in a world in which returns are stationary. Another important issue is whether the estimate of the asset covariance matrix Σ itself is reliable if it is estimated from a sample of historical data. It is well-known that computing a meaningful asset return covariance matrix requires a large number of observations—many more observations than the number of assets in the portfolio—and even then the sample covariance matrix may contain large estimation errors that produce poor results in the mean-variance optimization. A fix when sufficient data are not available is to compute the estimation errors in expected returns at a factor (e.g., industry, country, sector) level, and use their variances and covariances in the estimation error covariance matrix for the individual asset returns in a manner similar to standard factor models.

Several approximate methods for estimating Σ_{μ} have also been found to work well in practice (Stubbs and Vance, 2005). For example, it has been observed that simpler estimation approaches, such as computing the diagonal matrix containing the variances of the estimates (as opposed to the complete error covariance matrix), often provide most of the benefit in robust portfolio optimization. In addition, standard approaches for estimating expected returns, such as Bayesian statistics and regression-based methods, generate estimates for the estimation

error covariance matrix in the process of generating the estimates themselves.

Uncertainty sets (1) and (2) are both symmetric, that is, the sets are symmetric around the vector of uncertain parameters $\hat{\mu}$. One can also consider asymmetric uncertainty sets that better reflect information about the probability distributions of the uncertain parameters when the probability distributions are skewed (see Natarajan, Pachamanova, and Sim, 2008). Recently, there has been also a substantial interest in developing “structured” uncertainty sets, that is, uncertainty sets that are constructed for a specific purpose. Frequently, structured uncertainty sets based on simple intersections of elementary uncertainty sets are used to minimize the “conservatism” in traditional ellipsoidal or “box” uncertainty sets. We will discuss such uncertainty sets in more detail later in this entry.

Clarifying a Misconception about Robust Optimization

Among practitioners, the notion of robust portfolio optimization is often equated with the robust mean-variance model we just discussed, with uncertainty set (1) or (2) for the expected asset returns. While robust optimization applications frequently involve one form or another of this model, the actual scope of robust optimization can be much broader. We note that the term “robust optimization” refers to the technique of incorporating information about uncertainty sets for the parameters in the optimization model, and not to the specific definitions of uncertainty sets or the choice of which parameters to model as uncertain. For example, we can use the robust optimization methodology to incorporate considerations for uncertainty in the estimate of the covariance matrix in addition to the uncertainty in expected returns, and obtain a different robust portfolio allocation formulation.

Robust optimization can be applied also to portfolio allocation models that are different from the mean-variance framework, e.g.,

Sharpe ratio and value-at-risk optimization (see, for example, Goldfarb and Iyengar, 2003 and Natarajan, Pachamanova, and Sim, 2008). There are numerous useful and reasonable robust formulations, and a complete review is beyond the scope of this entry. We refer interested readers to Fabozzi et al. (2007) for further details.

THE RELATIONSHIP TO BAYESIAN METHODS AND ECONOMIC THEORY

Critics have argued that robust optimization is not really different from shrinkage estimators that combine the minimum variance portfolio with a speculative investment portfolio. Indeed, when using a particular uncertainty set for the expected returns (assuming all other parameters in the mean-variance problem are certain), it can be shown that the optimal mean-variance portfolio weights using robust optimization are a linear combination of the weights of the minimum variance portfolio (which is independent of investor preferences or expected returns) and a mean-variance efficient portfolio. These portfolio weights can also be obtained by solving a standard mean-variance problem with expected return estimates derived from a standard shrinkage estimator with specific shrinkage parameters (see, for example, Garlappi, Uppal, and Wang, 2007 and Scherer, 2005). Robust optimization thus appears to offer a less transparent way to express investor preferences and tolerance to uncertainty than other approaches, such as *Bayesian methods*, in which the shrinkage parameters can be defined explicitly.

In general, however, robust optimization is not necessarily equivalent to shrinkage estimation. For instance, differences are apparent in the presence of additional portfolio constraints. Furthermore, as we mentioned earlier, the robust optimization methodology can be used to account for uncertainty in parameters other than expected asset returns (covariances of asset

returns, for example), making its relationship with Bayesian estimation even less obvious.

The concept of robust optimization has been criticized also from the point of view of classical economic theory (see, for example, Sims, 2001). From a behavioral and decision-making point of view, few individuals have max-min preferences. Indeed, max-min preferences describe the behavior of decision makers who face great ambiguity and thus make choices consistent with the belief that the worst possible outcomes are highly likely. This kind of conservative behavior is not typical of the average investor. The problem of overconservativeness in applying robust optimization, however, can be controlled by modifying the specification of uncertainty sets for the parameters, as we will explain in the following section.

USING ROBUST PORTFOLIO OPTIMIZATION IN PRACTICE

One of the main problems in assessing the practical benefits of the robust optimization approach is that its performance is dependent on the choice (or calibration) of the model parameters, such as the coefficient of aversion to estimation error δ . In a sense, however, this issue is no different from the calibration of standard parameters in the classical portfolio optimization framework, such as the length of the estimation period to use for forecast generation and the choice of the risk aversion coefficient. These and other parameters need to be determined empirically or subjectively.

Note also that other *robust modeling* devices such as Bayesian estimators and the Black-Litterman model (for an overview, see Fabozzi, Focardi, and Kolm, 2006) have similar issues. In particular, for shrinkage estimators, the portfolio manager needs to determine which shrinkage target to use and the size of the shrinkage parameter. In the Black-Litterman model, he needs to provide his confidence in equilibrium as well as his confidence in his views.

The values of the robust formulation parameters can sometimes be matched to probabilistic guarantees. For example, if the estimates of the expected asset returns are assumed to be normally distributed, then there is an $\omega\%$ chance that the estimates will fall in the ellipsoidal set (2) around the manager's estimates $\hat{\mu}$,

$$U_\delta(\hat{\mu}) = \{ \mu \mid (\mu - \hat{\mu})' \Sigma_\mu^{-1} (\mu - \hat{\mu}) \leq \delta^2 \}$$

if δ^2 is assigned the value of the ω th percentile of a χ^2 distribution with degrees of freedom equal to the number of assets in the portfolio. As an example, suppose that there are 15 assets in the asset universe and that all returns are normally distributed. If we choose $\delta^2 = 25$, then 95% of all expected returns will be in the set $U_\delta(\hat{\mu})$.

More generally, if the expected returns can belong to any possible probability distribution, then assigning

$$\delta = \sqrt{\frac{1 - \omega}{\omega}}$$

guarantees that the estimates will fall in the uncertainty set $U_\delta(\hat{\mu})$ with probability at least $\omega\%$ (El Ghaoui, Oks, and Oustry, 2003).

It has been observed that in practice the standard robust mean-variance formulation with the above uncertainty set specification for estimated expected returns may result in portfolio allocations that are too pessimistic. Recall that the traditional robust counterpart tries to find the optimal solution so that constraints containing uncertain coefficients are satisfied for the worst-case realizations of the uncertain parameters. Naturally, the larger the uncertainty set, the greater the chance that the optimal portfolio allocation will be conservative. Therefore, especially in situations in which the worst-case expected returns can be far away from the estimated expected returns, some portfolio performance may be sacrificed. Of course, we can always make a formulation less pessimistic by considering a smaller uncertainty set. For the uncertainty set above, we can achieve this by decreasing the parameter δ . However, there is a recent trend among practitioners to apply more

structured restrictions. We provide an example of a structured uncertainty set next.

When we formulated the robust portfolio optimization problem earlier in this entry, we made the assumption that all of the actual realizations of expected returns could be worse than their expected values. Thus, the net adjustment in the expected portfolio return will always be downwards. While this leads to a more robust problem than the original one, in many instances it may be too pessimistic to assume that all estimation errors go against us. Instead, it may be more reasonable to believe that at least some of the true realizations will be above their expected values. For example, we may make the assumption that there are approximately as many realizations above the estimated values as there are realizations below the estimated values. This condition can be incorporated in the portfolio optimization problem by adding an additional restriction to the uncertainty set (2). Ceria and Stubbs (2006) refer to this adjustment as a "zero net alpha adjustment." Instead of adjusting the alphas of the estimates, we can perform this kind of adjustment also on their standard deviations or variances. It can be shown that the effect of the zero net adjustment is equivalent to modifying the covariance matrix Σ_μ of estimation errors for the expected returns. Tests with real data indicate that robust mean-variance optimization with this kind of adjustment for expected return estimates outperforms classical mean-variance optimization 70% to 80% of the time (Ceria and Stubbs, 2006).

Other structured uncertainty sets include "tiered" uncertainty sets in which some of the uncertain parameters are modeled as "well-behaved," while others are modeled as "misbehaving." The modeler can require protection against a prespecified number of parameters that he believes will "misbehave," that is, which will deviate significantly from their expected values (see, for example, Bienstock, 2006). In the context of portfolio optimization, we would specify "misbehaving" parameters

as those realizations of expected asset returns that are likely to be lower than their estimates.

Effect of Robust Portfolio Optimization Formulations on Performance

As we mentioned earlier, some tests with simulated and real market data indicate that robust optimization, when inaccuracy is assumed in the expected return estimates, outperforms classical mean-variance optimization in terms of total excess return a large percentage (70% to 80%) of the time (Ceria and Stubbs, 2006). Other tests have not been as conclusive (Lee, Stefek, and Zhelenyak, 2006). The factor that accounts for much of the difference is how the uncertainty in parameters is modeled. Therefore, finding a suitable degree of robustness and appropriate definitions of uncertainty sets can have a significant impact on portfolio performance.

Independent tests by practitioners and academics using both simulated and market data appear to confirm that robust optimization generally results in more stable portfolio weights, that is, that it eliminates the extreme corner solutions resulting from traditional mean-variance optimization. Robust mean-variance optimization also appears to improve worst-case portfolio performance, and results in smoother and more consistent portfolio returns. Finally, by preventing large swings in positions, robust optimization frequently makes better use of the turnover budget and risk constraints.

Robust optimization, however, is not a panacea. By using robust portfolio optimization, investors are likely to trade off the optimality of their portfolio allocation in cases in which nature behaves as they predicted for protection against the risk of inaccurate estimation. Therefore, investors using the technique should not expect to do better than classical portfolio optimization when estimation errors have little impact, or when typical scenarios occur. They should, however, expect insurance in scenarios

in which their estimates deviate from the actual realized values by up to the amount they have prespecified in the modeling process.

PRACTICAL CONSIDERATIONS FOR ROBUST PORTFOLIO ALLOCATION

Which type of robust models is best for modeling financial portfolios? The short answer is: It depends. Among others, it depends on the size of the portfolio, the type of assets and their distributional characteristics, the portfolio strategies and trading styles involved, and the existing infrastructure. Sometimes it makes sense to consider a combination of several techniques, such as a blend of Bayesian estimation and robust portfolio optimization. This is an empirical question—indeed, the only way to find out which strategy performs best is through thorough research and testing. A simple step-by-step checklist for robust quantitative portfolio management could include:

1. Risk forecasting: Develop an accurate risk model
2. Return forecasting: Construct robust expected return estimates
3. Classical portfolio optimization: Start with a simple framework
4. Model risk mitigation:
 - a. Minimize estimation risk through the use of robust estimators
 - b. Improve the stability of the optimization framework through robust optimization
5. Extensions

Needless to say, by no means do we claim that this list is complete or that it has to be followed religiously—it is simply provided as a starting point for the quantitative portfolio manager.

In general, the most difficult item in this list is the calculation of robust expected return estimates. Developing profitable trading strategies (“ α generation”) is notoriously hard, but not

impossible. It is important to remember that modern portfolio optimization techniques and fancy mathematics are not going to help at all if the underlying trading strategies are poor.

Implicit in this list is that for each step one needs to perform thorough testing in order to understand the effect of changes and new additions to the model. It is not unusual that quantitative analysts and portfolio managers will have to revisit previous steps as part of the research and development process. For example, it is important to understand the interplay between forecast generation and the reliability of optimized portfolio weights. Introducing a robust optimizer may lead to more reliable, and often more stable, portfolio weights. However, how to make the optimization framework more robust depends on how expected return and risk forecasts are produced. Therefore, sometimes one has to refine or modify basic forecast generation. Identifying the individual and the combined contribution of different techniques is crucial in the development of a successful quantitative framework.

Minimizing estimation risk and improving the reliability of the optimization framework can be done in either order, or sometimes at the same time. The goal of both approaches is of course to improve the overall reliability and performance of the portfolio allocation framework. Some important questions to consider here are: When/why does the framework perform well (poorly)? How sensitive is it to changes in inputs? How does it behave when constraints change? Are portfolio weights intuitive—do they make sense? How high is the turnover of the portfolio over time?

Starting from the simple framework of classical portfolio optimization, many extensions are possible. Typical examples include the introduction of transaction costs models, more complex constraints (e.g., integer constraints such as round lotting or cardinality constraints), different risk measures (e.g., downside risk measures, higher moments), and dynamic and stochastic programming for incorporating in-

tertemporal dependencies. Often, these are problem specific and have to be dealt with on a case-by-case basis.

FUTURE DIRECTIONS

Advances in the mathematical and physical sciences have always had a major impact on finance. In particular, probability theory, statistics, econometrics, and operations research have provided the necessary tools and discipline for the development of modern financial economics and large-scale portfolio management. The substantial advances in the areas of robust estimation and robust optimization during the 1990s have proven to be of significant importance for the practical applicability and reliability of portfolio management and optimization.

From a theoretical perspective, the area of robust optimization is quite mature. By contrast, there are many unanswered questions in the practice of robust portfolio optimization. There is a need for more empirical research in order to provide better guidelines for applying robust optimization in a way that guarantees superior portfolio performance. In particular, practitioners need to understand better (1) the implications of using different types of uncertainty set, (2) the interaction between different forecast generation methods (estimation techniques) and robust optimization, (3) how to calibrate model parameters in the optimization model, and (4) how to deal with the overconservatism inherent in many robust models.

The robust optimization framework offers great flexibility and many new interesting possibilities in portfolio management. For instance, robust portfolio optimization can exploit the notion of statistically equivalent portfolios. Specifically, with robust optimization, a manager can find the best portfolio that (1) minimizes trading costs with respect to the current holdings, and (2) has an expected portfolio return and variance that are statistically equivalent to those of the classical mean-variance

portfolio. Common portfolio constraints, such as transaction cost considerations and tax implications, can be handled efficiently in the robust optimization framework.

Robust optimization has also shown promise as a computationally attractive alternative to classical optimization methods when it comes to multiperiod portfolio management. There are numerous benefits to taking a long-term view of *investment management*. Treating portfolio allocation as a multiperiod problem provides a framework for robust overall portfolio management that takes into consideration the effects of rebalancing, transaction costs, future liabilities, and taxes.

By incorporating multiperiod views on asset behavior in rebalancing models, portfolio managers may be able to reduce their transaction costs, as the portfolio will not be rebalanced unnecessarily often. As a simple example, if a portfolio manager expects asset returns to dip at the next time period, but then recover, he may choose to hold on to the assets in his portfolio in order to minimize transaction costs. However, if the net gain from realizing the tax loss is higher than the expense of the transactions, he may choose to trade for short term benefit despite believing that the portfolio value will recover after two trading periods. These trade-offs are complex to evaluate and model, and traditional optimization techniques for multistage optimization, such as dynamic programming (see, for example, Bertsekas, 1995a) and stochastic programming (see, for example, Wallace and Ziemba, 2005), have not been very successful in this context as they result in computationally intractable problems due to the “curse of dimensionality.” However, if future asset returns are treated as uncertain parameters, and the uncertainty in their estimates is modeled through appropriately chosen uncertainty sets, the resulting portfolio optimization formulations are computationally tractable.

We emphasize that while the focus of this entry has been on the application of robust optimization to portfolio construction, robust

optimization is a powerful and general tool with financial applications that extend well beyond that of portfolio allocation. The robust optimization technique appears promising in enhancing existing models for optimal trading, the computation of hedge ratios, the estimation of econometric models, and quantitative model selection—just to mention a few. Certainly, the future may bring many more.

KEY POINTS

- As the use of quantitative techniques has become widespread in the investment industry, the consideration of estimation risk and model risk has grown in importance.
- In contrast to the traditional approach in which inputs to the portfolio allocation framework are treated as deterministic, robust portfolio optimization incorporates estimation errors in input parameters directly into the optimization process.
- In robust portfolio optimization, the inputs are not the traditional forecasts, such as expected returns and risk, but rather uncertainty sets containing these point estimates (e.g., confidence intervals around the forecasts).
- The robust optimization is a general technique that leads to a more reliable portfolio allocation framework and offers greater flexibility and many new interesting possibilities for the portfolio manager.
- One of the main problems in assessing the practical benefits of the robust optimization approach is that its performance is dependent on the choice (or calibration) of the model parameters, such as the coefficient of aversion to estimation error.
- Which type of robust model is best for modeling financial portfolios depends on, among other things, the size of the portfolio, the type of assets and their distributional characteristics, the portfolio strategies and trading styles involved, and the existing infrastructure.

REFERENCES

- Black, F., and Litterman, R. (1992). Global portfolio optimization. *Financial Analysts Journal* 48, September–October: 28–43.
- Ben-Tal, A., and Nemirovski, A. (1998). Robust convex optimization. *Mathematics of Operations Research* 23: 769–805.
- Bertsekas, D. (1995a). *Dynamic Programming and Optimal Control*, vol. 1. Belmont, MA: Athena Scientific.
- Bertsekas, D. (1995b). *Dynamic Programming and Optimal Control*, vol. 2. Belmont, MA: Athena Scientific.
- Bienstock, D. (2006). Experiments with robust optimization. Presentation at the International Symposium on Mathematical Programming, Rio de Janeiro, Brazil.
- Ceria, S., and Stubbs, R. (2006). Incorporating estimation errors into portfolio selection: Robust portfolio construction. *Journal of Asset Management* 7, 2: 109–127.
- El Ghaoui, L., and Lebret, H. (1997). Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18: 1035–1064.
- El Ghaoui, L., Oks, M., and Oustry, F. (2003). Worst-case value-at-risk and robust portfolio optimization: A conic optimization approach. *Operations Research* 51, 4: 543–556.
- Fabozzi, F. J., Focardi, S. M., and Kolm, P. N. (2006). *Financial Modeling of the Equity Market: From CAPM to Cointegration*. Hoboken, NJ: John Wiley & Sons.
- Fabozzi, F. J., Kolm, P. N., Pachamanova, D. A., and Focardi, S. F. (2007). *Robust Portfolio Optimization and Management*. Hoboken, NJ: John Wiley & Sons.
- Focardi, S. M., and Fabozzi, F. J. (2004). *The Mathematics of Financial Modeling and Investment Management*. Hoboken, NJ: John Wiley & Sons.
- Garlappi, L., Uppal, R., and Wang, T. (2007). Portfolio selection with parameter and model uncertainty: A multi-prior approach. *Review of Financial Studies* 20, 1: 41–81.
- Goldfarb, D., and Iyengar, G. (2003). Robust portfolio selection problems. *Mathematics of Operations Research* 28, 1: 1–38.
- Lee, J., Stefek, D., and Zhelenyak, A. (2006). Robust portfolio optimization: A closer look. *MSCI Barra Research Insights Report*, June.
- Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance* 7, 1: 77–91.
- Natarajan, K., Pachamanova, D. A., and Sim, M. (2008). Incorporating asymmetric distributional information in robust value-at-risk optimization. *Management Science* 54, 3: 573–585.
- Rachev, S. T., Hsu, J., Bagasheva, B., and Fabozzi, F. J. (2007). *Bayesian Methods in Finance*. Hoboken, NJ: John Wiley & Sons.
- Scherer, B. (2005). *How Different is Robust Optimization Really?* New York: Deutsche Asset Management.
- Sims, C. A. (2001). Pitfalls in a mini-max approach to model uncertainty. *American Economic Review* 91, 2: 51–54.
- Stubbs, R., and Vance, P. (2005). Computing return estimation error matrices for robust optimization. Report, Axioma, Inc., April.
- Wallace, S., and Ziemba, W. (eds.) (2005). *Applications of Stochastic Programming*. Philadelphia, PA: Cambridge University Press.

Probability Theory

Concepts of Probability Theory

MARKUS HÖCHSTÖTTER, PhD

Assistant Professor, University of Karlsruhe

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair-Professor, Department of Applied Mathematics and Statistics,
Stony Brook University, and Chief Scientist, FinAnalytica

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Probability theory is the mathematical approach to formalizing the uncertainty of events. Even though a decision maker may not know which one of the set of possible events may finally occur, with probability theory, a decision maker has the means of providing each event with a certain probability. Furthermore, it provides the decision maker with the axioms to compute the probability of a composed event in a unique way. The rather formal environment of probability theory translates in a reasonable manner to the problems related to risk and uncertainty in finance such as, for example, the future price of a financial asset. Today, investors may be aware of the price of a certain asset, but they cannot say for sure what value it might have tomorrow. To make a prudent decision, investors need to assess the possible scenarios for tomorrow's price and assign to each scenario a probability of occurrence. Only then can investors reasonably determine whether the financial asset will satisfy an investment objective.

Probability theory serves as the quantification of risk in finance. To estimate probabilistic models, we have to gather and process empirical

data. In this context, we need the tools provided by statistics. In this entry, we introduce the general concepts of probability theory.

HISTORICAL DEVELOPMENT OF ALTERNATIVE APPROACHES TO PROBABILITY

Before we introduce the formal definitions, we provide a brief outline of the historical development of probability theory and the alternative approaches since probability is, by no means, unique in its interpretation. We will describe the two most common approaches: relative frequencies and axiomatic system.

Probability as Relative Frequencies

The relative frequencies approach to probability was conceived in 1928 by Richard von Mises (Mises, 1928) and as the name suggests formulates probability as the relative frequencies denoted by $f(x_i)$. This initial idea was extended by Hans Reichenbach (1935). Given large samples, it was understood that $f(x_i)$ was equal to the true probability of value x_i . For example, if $f(x_i)$ is small, then the true probability of value x_i occurring should be small, in general. However, $f(x_i)$ itself is subject to uncertainty. Thus, the relative frequencies might deviate from the corresponding probabilities. For example, if the sample is not large enough, whatever large may be, then it is likely that we obtain a rare set of observations and draw the wrong conclusion with respect to the underlying probabilities.

This point can be illustrated with a simple example. Consider throwing a six-sided dice 12 times. Intuitively, one would expect the numbers 1 through 6 to occur twice each, since this would correspond to the theoretical probabilities of $1/6$ for each number. But since so many different outcomes of this experiment are very likely possible, one might observe relative frequencies of these numbers different from $1/6$. So, based on the relative frequencies, one might draw the wrong conclusion with respect to the true underlying probabilities of the according values. However, if we increase the repetitions from 12 to 1,000, for example, with a high de-

gree of certainty, the relative frequency of each number will be pretty close to $1/6$.

The reasoning of von Mises and Reichenbach was that since extreme observations are unlikely given a reasonable sample size, the relative frequencies will portray the true probabilities with a high degree of accuracy. In other words, probability statements based on relative frequencies were justifiable since, in practice, highly unlikely events could be ruled out.

In the context of our dice example, they would consider unlikely that certain numbers appeared significantly more often than others if the series of repetitions is, say, 1,000. But still, who could guarantee that we do not accidentally end up throwing 300 1s, 300 2s, 400 3s, and nothing else?

The approach of von Mises becomes relevant, again, in the context of estimating and hypothesis testing. For now, however, we will not pay any further attention to it but turn to the alternative approach to probability theory.

Axiomatic System

Introduced by Andrei N. Kolmogorov in 1933, the axiomatic system abstracted probability from relative frequencies as obtained from observations and instead treated probability as purely mathematical. The variables were no longer understood as the quantities that could be observed but rather as some theoretical entities "behind the scenes." Strict rules were set up that controlled the behavior of the variables with respect to their likelihood of assuming values from a predetermined set. So, for example, consider the price of a stock, say General Electric (GE). GE's stock price as a variable is not what you can observe but a theoretical quantity obeying a particular system of probabilities. What you observe is merely realizations of the stock price with no implication on the true probability of the values since the latter is given and does not change from sample to sample. The

relative frequencies, however, are subject to change depending on the sample.

We illustrate the need for an axiomatic system due to the dependence of relative frequencies on samples using our dice example. Consider the chance of occurrence of the number 1. Based on intuition, since there are six different “numbers of dots” on a dice, the number 1 should have a chance of $1/6$, right? Suppose we obtain the information based on two samples of 12 repetitions each, that is, $n_1 = n_2 = 12$. In the following table, we report the absolute frequencies, a_i , representing how many times the individual numbers of dots 1 through 6 were observed.

Number of Dots	Absolute Frequencies a_i	
	Sample 1	Sample 2
1	4	1
2	1	1
3	3	1
4	0	1
5	1	1
6	3	7
Total	12	12

That is, in sample 1, 1 dot was observed 4 times while, in sample 2, 1 dot was observed only once, and so on.

From the above observations, we obtain the following relative frequencies

Number of Dots	Relative Frequencies $f(x_i)$	
	Sample 1	Sample 2
1	0.3333	0.0833
2	0.0833	0.0833
3	0.2500	0.0833
4	0.0000	0.0833
5	0.0833	0.0833
6	0.2500	0.5833
Total	1.0000	1.0000

That is, in sample 1, 1 dot was observed 33.33% of the time while in sample 2, 1 dot was observed 8.33% of the time, and so on. We see that both samples lead to completely different results about the relative frequencies for the

number of dots. But, as we will see, the theoretical probability is $1/6 = 0.1667$, for each value 1 through 6. So, returning to our original question of the chance of occurrence of 1 dot, the answer is still $1/6 = 0.1667$.

In finance, the problem arising with this concept of probability is that, despite the knowledge of the axiomatic system, we do not know for sure what the theoretical probability is for each value. We can only obtain a certain degree of certainty as to what it approximately might be. This insight must be gained from estimation based on samples and, thus, from the related relative frequencies. So, it might appear reasonable to use as many observations as possible. However, even if we try to counteract the sample-dependence of relative frequencies by using a large number of observations, there might be a change in the underlying probabilities exerting additional influence on the sample outcome. For example, during the period of a bull market, the probabilities associated with an upward movement of some stock price might be higher than under a bear market scenario.

Despite this shortcoming, the concept of probability as an abstract quantity as formulated by Kolmogorov (1933) has become the standard in probability theory.

SET OPERATIONS AND PRELIMINARIES

Before proceeding to the formal definition of probability, randomness, and random variables we need to introduce some terminology.

Set Operations

A set is a combination of *elements*. Usually, we denote a set by some capital (uppercase) letter, for example S , while the elements are denoted by lowercase letters such as a, b, c, \dots or a_1, a_2, \dots . To indicate that a set S consists of exactly the elements a, b, c , we write $S = \{a, b, c\}$. If we want to say that element a belongs to S , the notation used is that $a \in S$ where \in means “belongs

to." If, instead, a does not belong to S , then the notation used is $a \notin S$ where \notin means "does not belong to."

A type of set such as $S = \{a,b,c\}$ is said to be *countable* since we can actually count the individual elements a , b , and c . A set might also consist of all real numbers inside of and including some bounds, say a and b . Then, the set is equal to the interval from a to b , which would be expressed in mathematical notation as $S = [a,b]$. If either one bound or both do not belong to the set, then this would be written as either $S = (a,b]$, $S = [a,b)$, or $S = (a,b)$, respectively, where the parentheses denote that the value is excluded. An interval is an *uncountable* set since, in contrast to a countable set $S = \{a,b,c\}$, we cannot count the elements of an interval.¹

We now present the operators used in the context of sets. The first is *equality* denoted by $=$ and intuitively stating that two sets are equal, that is, $S_1 = S_2$, if they consist of the same elements. If a set S consists of no elements, it is referred to as an *empty set* and is denoted by $S = \emptyset$. If the elements of S_1 are all contained in S_2 , the notation used is $S_1 \subset S_2$ or $S_1 \subseteq S_2$. In the first case, S_2 also contains additional elements not in S_1 while, in the second case, the sets might also be equal. For example, let $S_1 = \{a,b\}$ and $S_2 = \{a,b,c\}$, then $S_1 \subset S_2$. The operator \subseteq would indicate that S_2 consists of, at least, a and b . Or, let $M_1 = [0,1]$ and $M_2 = [0.5,1]$, then $M_2 \subset M_1$.

If we want to join a couple of sets, we use the *union operator* denoted by \cup . For example, let $S_1 = \{a,b\}$ and $S_2 = \{b,c,d\}$, then the union would be $S_1 \cup S_2 = \{a,b,c,d\}$. Or, let $M_1 = [0,1]$ and $M_2 = [0.5,1]$, then $M_2 \cup M_1 = [0,1] = M_1$.² If we join n sets S_1, S_2, \dots, S_n with $n \geq 2$, we denote the union by $\cup_{i=1}^n S_i$.

The opposite operator to the union is the *difference* denoted by the \setminus symbol. If we take the difference between set S_1 and set S_2 , that is, $S_1 \setminus S_2$, we discard from S_1 all the elements that are common to both S_1 and set S_2 . For example, let $S_1 = \{a,b\}$ and $S_2 = \{b,c,d\}$, then $S_1 \setminus S_2 = \{a\}$.

To indicate that we want to single out elements that are contained in several sets simul-

taneously, then we use the *intersection operator* \cap . For example, with the previous sets, the intersection would be $S_1 \cap S_2 = \{b\}$. Or, let $M_1 = [0,1]$ and $M_2 = [0.5,1]$, then $M_1 \cap M_2 = [0.5,1] = M_2$. Instead of the \cap symbol, one sometimes simply writes $S_1 S_2$ to indicate intersection.

If two sets contain no common elements (i.e., the intersection is the empty set), then the sets are said to be *pairwise disjoint*. For example, the sets $S_1 = \{a,b\}$ and $S_2 = \{c,d\}$ are pairwise disjoint since $S_1 \cap S_2 = \emptyset$. Or, let $M_1 = [0,0.5)$ and $M_2 = [0.5, 1]$, then $M_1 \cap M_2 = \emptyset$. If we intersect n sets S_1, S_2, \dots, S_n with $n \geq 2$, we denote the intersection by $\cap_{i=1}^n S_i$.

The *complement* to some set S is denoted by \bar{S} . It is defined as $S \cap \bar{S} = \emptyset$ and $S \cup \bar{S} = \Omega$. That is, the complement \bar{S} is the remainder of Ω that is not contained in S .

Right-Continuous and Non-decreasing Functions

Next we introduce two concepts of functions that should be understood in order to appreciate probability theory: right-continuous function and non-decreasing function.

A function f is right-continuous at \tilde{x} if the limit from the right of the function values coincides with the actual value of f at \tilde{x} . Formally, that is $\lim_{x \rightarrow \tilde{x}} f(x) = f(\tilde{x})$. We illustrate this in Figure 1. At the abscissae x_1 and x_2 , the function f jumps to $f(x_1)$ and $f(x_2)$ respectively.³ After each jump, the function remains at the new level for some time. Hence, approaching x_1 from the right, that is, for higher x -values, the function f approaches $f(x_1)$ smoothly. This is not the case when approaching x_1 from the left since f jumps at x_1 and, hence, deviates from the left-hand limit. The same reasoning applies to f at abscissa x_2 . A function is said to be a *right-continuous function* if it is right-continuous at every value on the x -axis.

A function f is said to be a *non-decreasing function* if it never assumes a value smaller than any value to the left. We demonstrate this using Figure 2. We see that while, in the different sections

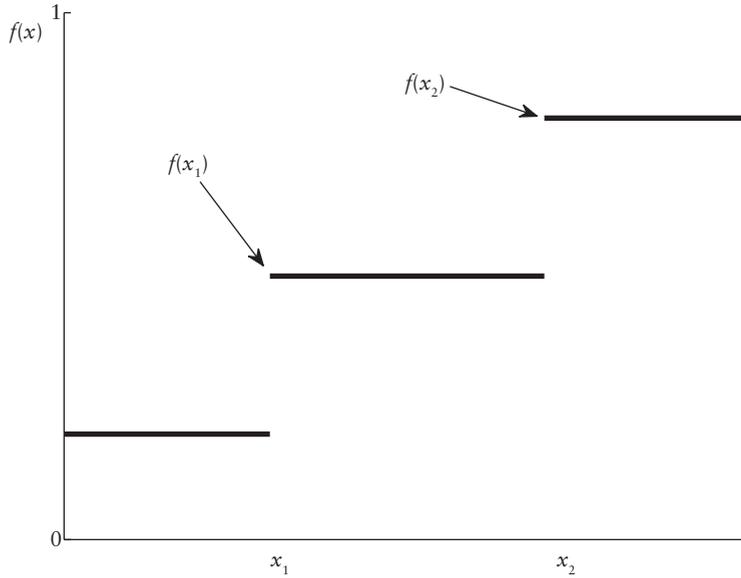


Figure 1 Demonstration of Right-Continuity of Some Hypothetical Function f at Values x_1 and x_2

A, B, and C, f might grow at different rates, it never decreases. Even for x -values in section B, f has zero and thus a nonnegative slope.

Outcome, Space, and Events

Before we dive into the theory, we will use examples that help illustrate the concept be-

hind the definitions that follow later in this entry.

Let us first consider again the number of dots of a dice. If we throw it once, we observe a certain value, that is, a realization of the abstract number of dots, say 4. This is one particular outcome of the random experiment. We will denote the outcomes by ω and a particular

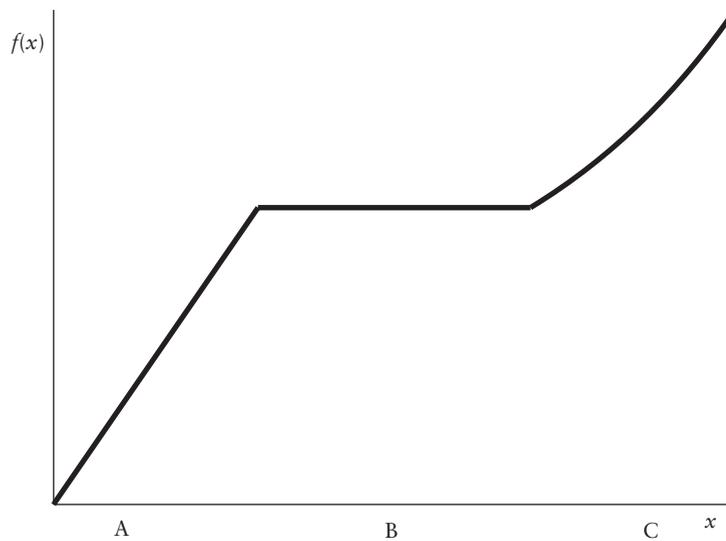


Figure 2 Hypothetical Non-decreasing Function f

outcome i will be denoted by ω_i . We might just as well have realized 2, for example, which would represent another outcome. All feasible outcomes, in this experiment, are given by

$$\omega_1 = 1 \quad \omega_2 = 2 \quad \omega_3 = 3 \quad \omega_4 = 4 \quad \omega_5 = 5 \quad \omega_6 = 6$$

The set of all feasible outcomes is called *space* and is denoted by Ω . In our example, $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$.

Suppose that we are not interested in the exact number of points but care about whether we obtain an odd or an even number, instead. That is, we want to know whether the outcome is from $A = \{\omega_1, \omega_3, \omega_5\}$ —that is, the set of all odd numbers—or $B = \{\omega_2, \omega_4, \omega_6\}$ —the set of all even numbers. The sets A and B are both contained in Ω ; that is, both sets are *subsets* of Ω . Any subsets of Ω are called *events*. So, we are interested in the events “odd” and “even” number of dots. When individual outcomes are treated as events, they are sometimes referred to as *elementary events* or *atoms*.

All possible subsets of Ω are given by the so-called *power set* 2^Ω of Ω . A power set of Ω is a set containing all possible subsets of Ω including the empty set \emptyset and Ω , itself.⁴

For our dice example, the power set is given in Table 1. With the aid of this power set, we are able to describe all possible events such as, for example, the number of dots less than 3 (i.e., $\{\omega_1, \omega_2\}$) or the number of dots either 1 or greater than or equal to 4 (i.e., $\{\omega_1, \omega_4, \omega_5, \omega_6\}$).

The power set has an additional pleasant feature. It contains any union of arbitrarily many events as well as any intersection of arbitrarily many events. Because of this, we say that 2^Ω is *closed under countable unions* and *closed under countable intersections*. Unions are employed to express that at least one of the events has to occur. We use intersections when we want to express that the events have to occur simultaneously. The power set also contains the complements to all events.

As we will later see, all these properties of the power set are features of a σ -*algebra* (in words: sigma-algebra), often denoted by \mathbb{A} .

Now consider an example where the space Ω is no longer countable. Suppose that we are analyzing the daily logarithmic returns for a common stock or common stock index. Theoretically, any real number is a feasible outcome for a particular day’s return.⁵ So, events are characterized by singular values as well as closed or open intervals on the real line. For example, we might be interested in the event E that the S&P 500 stock index return is “at least 1%.” Using the notation introduced earlier, this would be expressed as the half-open interval $E = [0.01, \infty)$.⁶ This event consists of the uncountable union of all outcomes between 0.01 and ∞ . Now, as the sets containing all feasible events, we might take, again, the power set of the real numbers, that is, 2^Ω with $\Omega = (-\infty, \infty) = \mathbb{R}$.⁷ But, for theoretical reasons

Table 1 The Power Set of the Example Number of Dots of a Dice

$$2^\Omega = \{\emptyset, \{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}, \{\omega_5\}, \{\omega_6\}, \{\omega_2, \omega_3\}, \{\omega_2, \omega_4\}, \{\omega_2, \omega_5\}, \{\omega_2, \omega_6\}, \dots$$

$$\{\omega_3, \omega_4\}, \{\omega_3, \omega_5\}, \{\omega_3, \omega_6\}, \{\omega_4, \omega_5\}, \{\omega_4, \omega_6\}, \{\omega_5, \omega_6\}, \{\omega_1, \omega_2, \omega_3\},$$

$$\{\omega_1, \omega_2, \omega_4\}, \{\omega_1, \omega_2, \omega_5\}, \{\omega_1, \omega_2, \omega_6\}, \{\omega_1, \omega_3, \omega_4\}, \{\omega_1, \omega_3, \omega_5\}, \{\omega_1, \omega_3, \omega_6\},$$

$$\{\omega_1, \omega_4, \omega_5\}, \{\omega_1, \omega_4, \omega_6\}, \{\omega_1, \omega_5, \omega_6\}, \{\omega_2, \omega_3, \omega_4\}, \{\omega_2, \omega_3, \omega_5\}, \{\omega_2, \omega_3, \omega_6\}, \dots$$

$$\{\omega_2, \omega_4, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}, \{\omega_2, \omega_5, \omega_6\}, \{\omega_3, \omega_4, \omega_5\}, \{\omega_3, \omega_4, \omega_6\}, \{\omega_3, \omega_5, \omega_6\},$$

$$\{\omega_4, \omega_5, \omega_6\}, \{\omega_1, \omega_2, \omega_3, \omega_4\}, \{\omega_1, \omega_2, \omega_3, \omega_5\}, \{\omega_1, \omega_2, \omega_3, \omega_6\}, \{\omega_1, \omega_2, \omega_4, \omega_5\},$$

$$\{\omega_1, \omega_2, \omega_4, \omega_6\}, \{\omega_1, \omega_2, \omega_5, \omega_6\}, \{\omega_1, \omega_3, \omega_4, \omega_5\}, \{\omega_1, \omega_3, \omega_4, \omega_6\},$$

$$\{\omega_1, \omega_3, \omega_5, \omega_6\}, \{\omega_1, \omega_4, \omega_5, \omega_6\}, \{\omega_2, \omega_3, \omega_4, \omega_5\}, \{\omega_2, \omega_3, \omega_4, \omega_6\},$$

$$\{\omega_2, \omega_3, \omega_5, \omega_6\}, \{\omega_2, \omega_4, \omega_5, \omega_6\}, \{\omega_3, \omega_4, \omega_5, \omega_6\}, \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\},$$

$$\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_6\}, \{\omega_1, \omega_2, \omega_3, \omega_5, \omega_6\}, \{\omega_1, \omega_2, \omega_4, \omega_5, \omega_6\},$$

$$\{\omega_1, \omega_3, \omega_4, \omega_5, \omega_6\}, \{\omega_1, \omega_3, \omega_4, \omega_5, \omega_6\}, \Omega\}$$

Note: The notation $\{\omega_i\}$ for $i = 1, 2, \dots, 6$ indicates that the outcomes are treated as events.

beyond the scope of this entry, that might cause trouble.

Instead, we take a different approach. To design our set of events of the uncountable space Ω , we begin with the inclusion of the events “any real number,” which is the space Ω , itself, and “no number at all,” which is the empty set \emptyset . Next, we include all events of the form “less than or equal to a ,” for any real number a , that is, we consider all half-open intervals $(-\infty, a]$, for any $a \in \mathbb{R}$. Now, for each of these $(-\infty, a]$, we add its complement $\overline{(-\infty, a]} = \Omega \setminus (-\infty, a] = (a, \infty)$, which expresses the event “greater than a .” So far, our set of events contains \emptyset, Ω , all sets $(-\infty, a]$, and all the sets (a, ∞) . Furthermore, we include all possible unions and intersections of everything already in the set of events as well as of the resulting unions and intersections themselves. By doing this, we guarantee that any event of practical relevance of an uncountable space is considered by our set of events.

With this procedure, we construct the Borel σ -algebra, \mathbb{B} . This is the collection of events we will use any time we deal with real numbers.

The events from the respective σ -algebra of the two examples can be assigned probabilities in a unique way, as we will see.

The Measurable Space

Let us now express the ideas from the previous examples in a formal way. To describe a random experiment, we need to formulate

1. Outcomes ω
2. Space Ω
3. σ -algebra \mathbb{A}

Definition 1—Space: The space Ω contains all outcomes. Depending on the outcomes ω , the space Ω is either countable or uncountable.

Definition 2— σ -algebra: The σ -algebra \mathbb{A} is the collection of events (subsets of Ω) with the following properties:

- a. $\Omega \in \mathbb{A}$ and $\emptyset \in \mathbb{A}$.
- b. If event $E \in \mathbb{A}$ then $\bar{E} \in \mathbb{A}$.

- c. If the countable sequence of events $E_1, E_2, E_3, \dots \in \mathbb{A}$ then $\cup_{i=1}^{\infty} E_i \in \mathbb{A}$ and $\cap_{i=1}^{\infty} E_i \in \mathbb{A}$.

Definition 3—Borel σ -algebra: The σ -algebra formed by $\emptyset, \Omega = \mathbb{R}$, intervals $(\infty, a]$ for some real a , and countable unions and intersections of these intervals is called a *Borel σ -algebra* and denoted by \mathbb{B} .

Note that we can have several σ -algebrae for some space Ω . Depending on the events we are interested in, we can think of a σ -algebra \mathbb{A} that contains fewer elements than 2^Ω (for countable Ω), or the Borel σ -algebra (for uncountable Ω). For example, we might think of $\mathbb{A} = \{\emptyset, \Omega\}$, that is, we only want to know whether any outcome occurs or nothing at all.⁸ It is easy to verify that this simple \mathbb{A} fulfills all requirements a, b, and c of Definition 2.

Definition 4—Measurable space: The tuple (Ω, \mathbb{A}) with \mathbb{A} being a σ -algebra of Ω is a *measurable space*.

A tuple is the combination of several components. For example, when we combine two values a and b , the resulting tuple is (a, b) , which we know to be a pair. If we combine three values a, b , and c , the resulting tuple (a, b, c) is known as a triplet.

Given a measurable space, we have enough to describe a random experiment. All that is left is to assign probabilities to the individual events. We will do so next.

PROBABILITY MEASURE

We start with a brief discussion of what we expect of a probability or *probability measure*; that is, the following properties:

Property 1: A probability measure should assign each event E from our σ -algebra a nonnegative value corresponding to the chance of this event occurring.

Property 2: The chance that the empty set occurs should be zero since, by definition, it is the improbable event of “no value.”

Property 3: The event that “any value” might occur (i.e., Ω) should be 1 or, equivalently, 100% since some outcome has to be observable.

Property 4: If we have two or more events that have nothing to do with one another that are pairwise disjoint or *mutually exclusive*, and create a new event by uniting them, the probability of the resulting union should equal the sum of the probabilities of the individual events.

To illustrate, let:

- The first event state that the S&P 500 log return is “maximally 5%,” that is, $E_1 = (-\infty, 0.05]$.
- The second event state that the S&P 500 log return is “at least 10%,” that is, $E_2 = [0.10, \infty)$.

Then, the probability of the S&P log return either being no greater than 5% or no less than 10% should be equal to the probability of E_1 plus the probability of E_2 .

Let’s proceed a little more formally. Let (Ω, \mathbb{A}) be a measurable space. Moreover, consider the following definition.

Definition 5—Probability measure: A function P on the σ -algebra \mathbb{A} of Ω is called a *probability measure* if it satisfies:

- a. $P(\emptyset) = 0$ and $P(\Omega) = 1$.
- b. For a countable sequence of events E_1, E_2, \dots in \mathbb{A} that are pairwise disjoint (i.e., $E_i \cap E_j = \emptyset \dots, i \neq j$), we have

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

This property is referred to as *countable additivity*.

Then we have everything we need to model randomness and chance, that is, we have the space Ω , the σ -algebra \mathbb{A} of Ω , and the probability measure P . This triplet (Ω, \mathbb{A}, P) forms the so called *probability space*.

At this point, we introduce the notion of *P-almost surely* (*P-a.s.*) occurring events. It is imaginable that even though $P(\Omega) = 1$, not all of the outcomes in Ω contribute positive prob-

ability. The entire positive probability may be contained in a subset of Ω while the remaining outcomes form the *unlikely* event with respect to the probability measure P . The event accounting for the entire positive probability with respect to P is called the *certain event with respect to P*. If we denote this event by E_{as} , then we have $P(E_{as}) = 1$ yielding $P(\Omega \setminus E_{as}) = 0$.

There are certain peculiarities of P depending on whether Ω is countable or not. It is essential to analyze these two alternatives since this distinction has important implications for the determination of the probability of certain events. Here is why.

Suppose, first, that Ω is countable. Then, we are able to assign the event $\{\omega_i\}$ associated with an individual outcome, ω_i , a nonnegative probability $p_i = P(\{\omega_i\})$, for all $\omega_i \in \Omega$. Moreover, the probability of any event E in the σ -algebra \mathbb{A} can be computed by adding the probabilities of all outcomes associated with E . That is,

$$P(E) = \sum_{\omega_i \in E} p_i$$

In particular, we have

$$P(\Omega) = \sum_{\omega_i \in \Omega} p_i = 1$$

Let us resume the six-sided dice tossing experiment. The probability of each number of dots 1 through 6 is 1/6 or formally,

$$P(\{\omega_1\}) = P(\{\omega_2\}) = \dots = P(\{\omega_6\}) = 1/6$$

or equivalently,

$$p_1 = p_2 = \dots = p_6 = 1/6$$

Suppose, instead, we have $\Omega = \mathbb{R}$. That is, Ω is uncountable and our σ -algebra is given by the Borel σ -algebra, \mathbb{B} . To give the probability of the events E in \mathbb{B} , we need an additional device, given in the next definition.

Definition 6—Distribution function: A function F is a *distribution function of the probability measure P* if it satisfies the following properties:

- a. F is right-continuous.
- b. F is non-decreasing.

- c. $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$.
- d. For any $x \in \mathbb{R}$, we have $F(x) = P((-\infty, x])$.

It follows that, for any interval $(x, y]$, we compute the associated probability according to

$$F(y) - F(x) = P((x, y]) \tag{1}$$

So, in this case we have a function F uniquely related to P from which we derive the probability of any event in \mathbb{B} . Note that in general F is only right-continuous, that is the limit of $F(y)$, when $y > x$ and $y \rightarrow x$, is exactly $F(x)$. At point x , we might have a jump of the distribution $F(x)$. The size of this jump equals $P(\{x\})$. This distribution function can be interpreted in a similar way to the relative empirical cumulative distribution function. That is, we state the probability of our quantity of interest being less than or equal to x .

To illustrate, the probability of the S&P 500 log return being at most 1%, $E = (-\infty, 0.01]$, is given by $F^{\text{S\&P 500}}(0.01) = P((-\infty, 0.01])$,⁹ while the probability of it being between -1% and 1% is

$$F^{\text{S\&P 500}}(0.01) - F^{\text{S\&P 500}}(-0.01) = P((-0.01, 0.01])$$

RANDOM VARIABLE

Now the time has come to introduce the concept of a *random variable*. When we refer to some quantity as being a random variable, we want to express that its value is subject to uncertainty, or randomness. Technically, the variable of interest is said to be *stochastic*. In contrast to a deterministic quantity whose value can be determined with certainty, the value of a random variable is not known until we can observe a realized outcome of the random experiment. However, since we know the probability space (Ω, \mathbb{A}, P) , we are aware of the possible values it can assume.

One way we can think of a random variable denoted by X is as follows. Suppose we have a random experiment where some outcome ω from the space Ω occurs. Then, depending on

this ω , the random variable X assumes some value $X(\omega) = x$, where ω can be understood as input to X . What we observe, finally, is the value x , which is only a consequence of the outcome ω of the underlying random experiment.

For example, we can think of the price of a 30-year Treasury bond as a random variable assuming values at random. However, expressed in a somewhat simple fashion, the 30-year Treasury bond depends completely on the prevailing market interest rate (or yield) and, hence, is a function of it. So, the underlying random experiment concerns the prevailing market interest rate with some outcome ω while the price of the Treasury bond, in turn, is merely a function of ω .

Consequently, a random variable is a function that is completely deterministic and depends on the outcome ω of some random experiment. In most applications, random variables have values that are real numbers.

So, we understand random variables as functions from some space into an image or state space. We need to become a little more formal at this point. To proceed, we will introduce a certain type of function, the *measurable function*, in the following

Definition 7—Measurable function: Let (Ω, \mathbb{A}) and (Ω', \mathbb{A}') be two measurable spaces. That is Ω, Ω' are spaces and \mathbb{A}, \mathbb{A}' their σ -algebrae, respectively. A function $X: \Omega \rightarrow \Omega'$ is \mathbb{A} - \mathbb{A}' -measurable if, for any set $E' \in \mathbb{A}'$, we have

$$X^{-1}(E') \in \mathbb{A}$$

In words, this means that a function from one space to another is measurable if the origin with respect to this function of each image in the σ -algebra of the state space can be traced in the σ -algebra of the domain space. Instead of \mathbb{A} - \mathbb{A}' -measurable, we will, henceforth, use simply *measurable* since, in our statements, it is clear which σ -algebrae are being referred to.

We illustrate this in Figure 3. Function X creates images in Ω' by mapping outcomes ω from

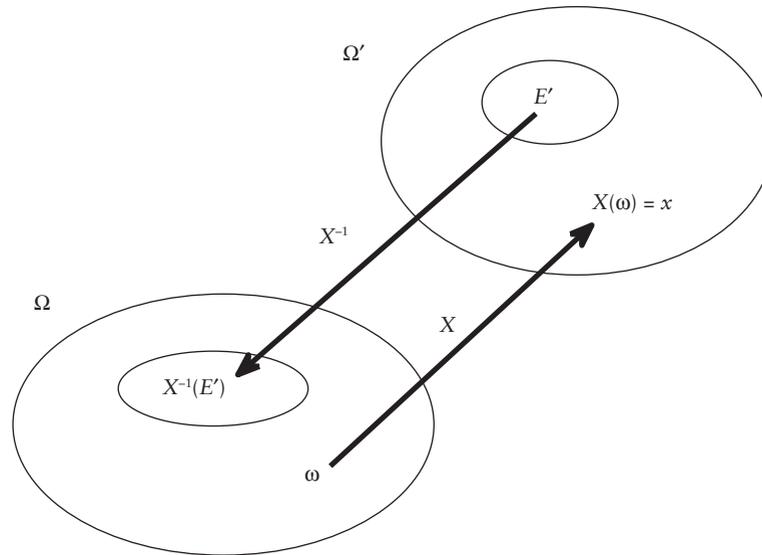


Figure 3 Relationship between Image E' and $X^{-1}(E')$ through the Measurable Function X

Ω with values $X(\omega) = x$ in Ω' . In reverse fashion, for each event E' in the state space with σ -algebra \mathbb{A}' , X^{-1} finds the corresponding origin of E' in σ -algebra \mathbb{A} of the probability space.

Now, we define a random variable X as a measurable function. That means for each event in the state space σ -algebra, \mathbb{A}' , we have a corresponding event in the σ -algebra of the domain space, \mathbb{A} .

To illustrate this, let us consider the example with the dice. Now we will treat the “number of points” as a random variable X . The possible outcome values of X are given by the state space Ω' , namely, $\Omega' = \{1,2,3,4,5,6\}$.¹⁰ The origin or domain space is given by the set of outcomes $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$. Now, we can think of our random variable X as the function $X: \Omega \rightarrow \Omega'$ with the particular map $X(\omega_i) = i$ with $i = 1, 2, \dots, 6$.

Random Variables on a Countable Space

We will distinguish between random variables on a countable space and on an uncountable space. We begin with the countable case.

The random variable X is a function mapping the countable space Ω into the state space Ω' . The state space Ω' contains all outcomes or values that X can obtain.¹¹ Thus, all outcomes in Ω' are countable images of the outcomes ω in Ω . Between the elements of the two spaces, we have the following relationship.

Let x be some outcome value of X in Ω' . Then, the corresponding outcomes from the domain space Ω are determined by the set

$$X^{-1}(\{x\}) = \{\omega : X(\omega) = x\}$$

In words, we look for all outcomes ω that are mapped to the outcome value x .

For events, in general, we have the relationship

$$X^{-1}(E') = \{\omega : X(\omega) \in E'\}$$

which is the set of all outcomes ω in the domain space that are mapped by X to the event E' in the state space. That leads us to the following definition:

Definition 8—Random variable on a countable space: Let (Ω, \mathbb{A}) and (Ω', \mathbb{A}') be two measurable spaces with countable Ω and Ω' . Then the mapping $X: \Omega \rightarrow \Omega'$ is a *random variable on a countable space* if, for any event $E' \in \mathbb{A}'$

composed of outcomes $x \in \Omega'$, we have

$$\begin{aligned} P^X(E') &= P(\{\omega : X(\omega) \in E'\}) = P(X^{-1}(E')) \\ &= P(X \in E') \end{aligned} \quad (2)$$

We can illustrate this with the following example from finance referred to as the “binomial stock price model.” The random variable of interest will be the price of some stock. We will denote the price of the stock by S . Suppose at the beginning of period t , the price of the stock is \$20 (i.e., $S_t = \$20$). At the beginning of the following period, $t + 1$, the stock price is either $S_{t+1} = \$18$ or $S_{t+1} = \$22$. We model this in the following way.

Let:

- (Ω, \mathbb{A}) and (Ω', \mathbb{A}') be two measurable spaces with $\Omega' = \{\$18, \$22\}$ (i.e., the state space of the period $t + 1$ stock price) and \mathbb{A} (i.e., the corresponding σ -algebra of all events with respect to the stock price in $t + 1$).
- Ω be the space consisting of the outcomes of some random experiment completely influencing the $t + 1$ stock price.
- \mathbb{A} be the corresponding σ -algebra of Ω with all events in the origin space.

Now, we can determine the origin of the event that

$$S_{t+1} = \$18 \text{ by } E_{\text{down}} = \{\omega : S(\omega) = \$18\}$$

and

$$S_{t+1} = \$22 \text{ by } E_{\text{up}} = \{\omega : S(\omega) = \$22\}$$

Thus, we have partitioned Ω into the two events, E_{down} and E_{up} , related to the two period $t + 1$ stock prices. With the probability measure P on Ω , we have the probability space (Ω, \mathbb{A}, P) . Consequently, due to equation (2), we are able to compute the probability $P^S(\$18) = P(E_{\text{down}})$ and $P^S(\$22) = P(E_{\text{up}})$, respectively.

Random Variables on an Uncountable Space

Now let’s look at the case when the probability space (Ω, \mathbb{A}, P) is no longer countable. Recall

the particular way in which events are assigned probabilities in this case.

While for a countable space any outcome ω can have positive probability, that is, $p_\omega > 0$, this is not the case for individual outcomes of an uncountable space. On an uncountable space, we can have the case that only events associated with intervals have positive probability. These probabilities are determined by the distribution function $F(x) = P(X < x) = P(X \leq x)$ according to equation (1).

This brings us to the following definition:

Definition 9—Random variable on a general possibly uncountable space: Let (Ω, \mathbb{A}) and (Ω', \mathbb{A}') be two measurable spaces with, at least, Ω uncountable. The map $X: \Omega \rightarrow \Omega'$ is a *random variable on the uncountable space* (Ω, \mathbb{A}, P) if it is measurable. That is, if, for any $E' \in \mathbb{A}'$, we have

$$X^{-1}(E') \in \mathbb{A}$$

induce probability from (Ω, \mathbb{A}, P) on (Ω', \mathbb{A}') by

$$\begin{aligned} P^X(E') &= p(\{\omega : X(\omega) \in E'\}) = P(X^{-1}(E')) \\ &= P(X \in E') \end{aligned}$$

We call this the *probability law* or distribution of X . Typically, the probability of $X \in E'$ is written using the following notation:

$$P^X(E') = P(X \in E')$$

Very often, we have the random variable X assume values that are real numbers (i.e., $\Omega' = \mathbb{R}$ and $\mathbb{B}' = \mathbb{B}$). Then, the events in the state space are characterized by countable unions and intersections of the intervals $(-\infty, a]$ corresponding to the events $\{X \leq a\}$, for real numbers a . In this case, we require that to be a random variable, X satisfies

$$\{\omega : X(\omega) \leq a\} = X^{-1}((-\infty, a]) \in \mathbb{B}$$

for any real a .

To illustrate, let’s use a call option on a stock. Suppose in period t we purchase a call option on a certain stock expiring in the next period $T = t + 1$. The strike price, denoted by K , is \$50.

Then as the buyer of the call option, in $t + 1$ we are entitled to purchase the stock for \$50 no matter what the market price of the stock (S_{t+1}) might be. The value of the call option at time $t + 1$, which we denote by C_{t+1} , depends on the market price of the stock at $t + 1$ relative to the strike price (K). Specifically,

- If S_{t+1} is less than K , then the value of the option is zero, that is, $C_{t+1} = 0$
- If S_{t+1} is greater than K , then the value of the option is equal to $S_{t+1} - K$

Let (Ω, \mathbb{A}, P) be the probability space with the stock price in $t + 1$; that is, $S_{t+1} = s$ representing the uncountable real-valued outcomes. So, we have the uncountable probability space $(\Omega, \mathbb{A}, P) = (\mathbb{R}, \mathbb{B}, P)$. Assume that the price at $t + 1$ can take any nonnegative value. Assume further that the probability of exactly s is zero (i.e., $P(S_{t+1} = s) = 0$), that is, the distribution function of the price at $T = 1$ is continuous. Let the value of the call option in $T = t + 1$, C_{t+1} , be our random variable mapping from Ω to Ω' . Since the possible values of the call option at $t + 1$ are real numbers, the state space is uncountable as well. Hence, we have $(\Omega', \mathbb{A}') = (\mathbb{R}, \mathbb{B})$. C_{t+1} , to be a random variable, is a \mathbb{B} - \mathbb{B}' -measurable function.

Now, the probability of the call becoming worthless is determined by the event in the origin space that the stock price falls below K . Formally, that equals

$$\begin{aligned} P^{C_{t+1}}(0) &= P(C_{t+1} \leq 0) = P(S_{t+1} \leq K) \\ &= P((-\infty, K]) \end{aligned}$$

since the corresponding event in \mathbb{A} to a 0 value for the call option is $(-\infty, K]$. Equivalently, $C_{t+1}^{-1}(\{0\}) = (-\infty, K]$. Any positive value c of C_{t+1} is associated with zero probability since we have

$$P^{C_{t+1}}(c) = P(C_{t+1} = c) = P(S_{t+1} = c + K) = 0$$

due to the relationship $C_{t+1} = S_{t+1} - K$ for $S_{t+1} > K$.

KEY POINTS

- Events in a mathematical probabilistic sense represent sets of values. They are used to describe a certain situation such as an asset price being below some benchmark value.
- A probability measure is a function that assigns each event a unique probability between zero and one. With respect to this probability measure an event is P-almost sure if it is assigned probability one, while an unlikely event is one with zero probability.
- A random variable is a function assuming values from a given set of values at random. The probability of the random variables assuming certain values is determined by the probability measure.
- A distribution function is uniquely related to the probability measure. It assigns real numbers values between zero and one. At any real number, it represents the probability that a random variable assumes values of at most this number.
- Stochastic is the Greek term for random. It is often used in probability theory to describe that something is not deterministic, that is, known with certainty in advance.

NOTES

1. Suppose we have the interval $[1,2]$, that is all real numbers between 1 and 2. We cannot count all numbers inside of this interval since, for any two numbers such as, for example, 1 and 1.001, 1.0001, or even 1.000001, there are always infinitely many more numbers that lie between them.
2. Note that in a set, we do not consider an element more than once.
3. By *abscissa* we mean a value on the horizontal x -axis.
4. For example, let $\Omega = \{1,2,3\}$, then the power set $2^\Omega = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \Omega\}$. That is, we have included all possible combinations of the original elements of Ω .

5. Let us assume, for now, that we are not restricted to a few digits due to measurement constraints or quotes conventions in the stock market. Instead, we consider being able to measure the returns to any degree of precision.
6. By convention, we never include ∞ since it is not a real number.
7. The symbol \mathbb{R} is just a mathematical abbreviation for the real numbers.
8. The empty set is interpreted as the *improbable event*.
9. We use the index in $F^{\text{S\&P } 500}$ to emphasize that this distribution function is unique to the probability of events related to the S\&P 500 log returns.
10. Note that we do not define the outcomes of number of dots as nominal or even rank

data anymore, but as numbers. That is 1 is 1, 2 is 2, and so on.

11. Theoretically, Ω' does not have to be countable; that is, it could contain more elements than X can assume values. But we restrict ourselves to countable state spaces Ω' consisting of exactly all the values of X .

REFERENCES

- Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer.
- Mises, R. von. (1928). *Wahrscheinlichkeit, Statistik und Wahrheit*. Wien: Springer.
- Reichenbach, H. (1935). *Wahrscheinlichkeitslehre: eine Untersuchung über die logischen und mathematischen Grundlagen der Wahrscheinlichkeitsrechnung*. Leiden: Sijthoff.

Discrete Probability Distributions

MARKUS HÖCHSTÖTTER, PhD

Assistant Professor, University of Karlsruhe

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair-Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Discrete probability distributions are needed whenever the random variable is to describe a quantity that can assume values from a countable set, either finite or infinite. A discrete probability distribution (or law) is quite intuitive in that it assigns certain values positive probabilities adding up to one, while any other value automatically has zero probability. In general, neglecting some of the mathematical rigor, discrete distributions can be understood from the insight gained from descriptive statistics. For example, the random number of defaults in a bond portfolio inside of a given period of time can be modeled with a discrete probability distribution. Another example is given by sampling when we are interested in whether an observation belongs to a certain group. Also, simple stock price models are based on discrete laws where the stock price can only change to one of a finite number of possible values.

Discrete random variables are random variables on the countable space. We present the most important discrete random variables used in finance and their *probability distribution* (also called *probability law*): Bernoulli, binomial, hypergeometric, multinomial, Poisson, and discrete uniform.

Appendix A provides a summary of the discrete distributions covered.

DISCRETE LAW

In order to understand the distributions discussed in this entry, we will explain the general concept of a *discrete law*. Based on the knowledge of countable probability spaces, we introduce the random variable on the countable space as the discrete random variable. To fully comprehend the discrete random variable, it is necessary to become familiar with the

process of assigning probabilities to events in the countable case. Furthermore, the cumulative distribution function will be presented as an important representative of probability. It is essential to understand the mean and variance parameters. Wherever appropriate, we draw analogies to descriptive statistics for a facilitation of the learning process.

Random Variable on the Countable Space

Recall the probability space (Ω, \mathbb{A}, P) where Ω is a countable space. The probability of any event E is given by

$$P(E) = \sum_{\omega_i \in E} p_i$$

with the p_i being the probabilities of the individual outcomes ω_i in the event E . Remember that the random variable X is the mapping from Ω into Ω' such that the state space Ω' is countable. (We denote random variables by capital letters, such as X , whereas the outcomes are denoted by small letters, such as x_i .) Thus, the probability of any event E' in the state space has probability

$$P(X \in E') = P^X(E') = \sum_{\omega_i: X(\omega_i) \in E'} p_i$$

since E' is associated with the set

$$\{\omega_i : X(\omega_i) \in E'\}$$

through X . The probability of each individual outcome of X yields the discrete probability law of X . It is given by $P(X = x_i) = p_i^X$, for all $x_i \in \Omega'$.

Only for individual discrete values x is the probability p^X positive. This is similar to the empirical frequency distribution with positive relative frequency f_i at certain observed values. If we sort the $x_i \in \Omega$ in ascending order, analogous to the empirical relative cumulative frequency distribution

$$F_{emp}^f(x) = \sum_{x_i \leq x} f_i$$

we obtain the *discrete cumulative distribution (cdf)* of X ,

$$F^X(x) = P(X \leq x) = \sum_{x_i \leq x} p_i^X$$

That is, we express the probability that X assumes a value no greater than x .

Suppose we want to know the probability of obtaining at most 3 dots when throwing a dice. That is, we are interested in the *cdf* of the random variable number of dots, at the value $x = 3$. We obtain it by

$$F^X(3) = p_1 + p_2 + p_3 = 1/6 + 1/6 + 1/6 = 0.5$$

where the p_i denote the respective probabilities of the number of dots less than or equal to 3. A graph of the *cdf* is shown in Figure 1.

Mean and Variance

The sample mean and variance are sample dependent statistics. Here we present the mean and variance of the distribution as parameters where the probability space can be understood as the analog to the population.

To illustrate, we use the random variable number of dots obtained by tossing a dice. Since we treat the numbers as numeric values, we are able to perform transformations and computations with them. By throwing a dice several times, we would be able to compute a sample average based on the respective outcome. So, a question could be: What number is theoretically expected? In our discussion below, we see how to answer that question.

Mean

The mean is the population equivalent to the sample average of a quantitative variable. In order to compute the sample average, we sum up all observations and divide the resulting value by the number of observations, which we will denote by n . Alternatively, we sum over all values weighted by their relative frequencies.

This brings us to the mean of a random variable. For the mean of a random variable,

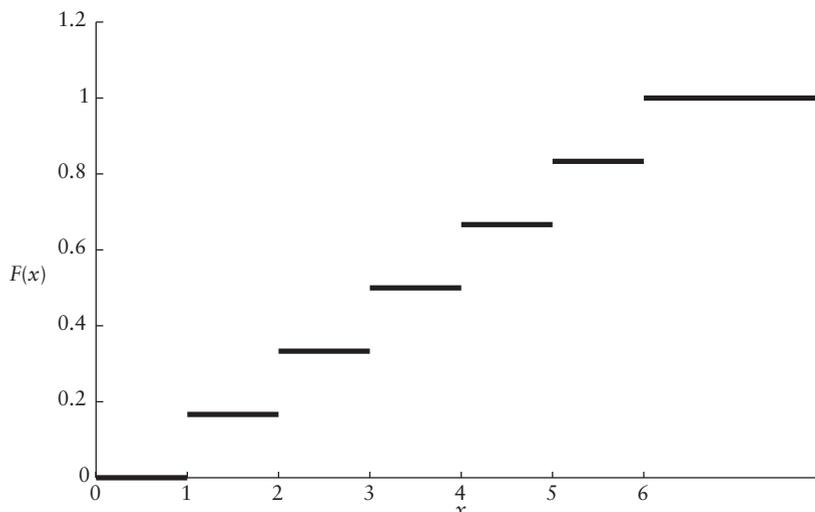


Figure 1 Cumulative Distribution Function of Number of Dots Appearing from Tossing a Dice

we compute the accumulation of the outcomes weighted by their respective probabilities; that is,

$$E(X) = \sum_{x_i \in \Omega} x_i \cdot p_i^X \quad (1)$$

given that equation (1) is finite. (Often, the mean is denoted as the parameter μ .) If the mean is not finite, then the mean is said to not exist. The mean equals the expected value of the random variable X . However, as we will see in the following examples, the mean does not actually have to be equal to one of the possible outcomes.

For the number of dots on the dice example, the expected value is

$$E(X) = \sum_{i=1}^6 i \cdot p_i \frac{1}{6} \sum_{i=1}^6 i = 21/6 = 3.5$$

So, on average, one can expect a value of 3.5 for the random variable, despite the fact this is not an obtainable number of dots. How can we interpret this? If we were to repeat the dice tossing many times, record for each toss the number of dots observed, then, if we averaged over all numbers obtained, we would end up with an average very close if not identical to 3.5.

Let's move from the dice tossing example to look at a binomial stock price model. With the

stock price S at the end of period 1 being either $S_1 = \$18$ or $S_1 = \$22$, we have only these two outcomes with positive probability each. We denote the probability measure of the stock price at the end of period 1 by $P^S(\cdot)$. At the beginning of the period, we assume the stock price to be $S_0 = \$20$. Furthermore, suppose that up- and down-movements are equally likely; that is, $P^S(18) = 1/2$ and $P^S(22) = 1/2$. So we obtain

$$E(S) = 1/2 \cdot \$18 + 1/2 \cdot \$22 = \$20$$

This means on average, the stock price will remain unchanged even though \$20 is itself not an obtainable outcome.

We can think of it this way. Suppose we observed some stock over a very long period of time and the probabilities for up- and down-movements did not change. Furthermore suppose that each time the stock price was \$20 at the beginning of some period, we recorded the respective end-of-period price. Then, we would finally end up with an average of these end-of-period stock prices very close to if not equal to \$20.

Variance

Just like in the realm of descriptive statistics, we are interested in the dispersion or spread of the data. For this, we introduce the *variance* as

a measure. Our focus is on the variance as a parameter of the random variable's distribution.

A sample measure of spread gives us information on the average deviation of observations from their sample mean. With the help of the variance, we intend to determine the magnitude we have to theoretically expect of the squared deviation of the outcome from the mean. Again, we use squares to eliminate the effect from the signs of the deviations as well as to emphasize larger deviations compared to smaller ones, just as we have done with the sample variance.

For the computation of the expected value of the squared deviations, we weight the individual squared differences of the outcomes from the mean with the probability of the respective outcome. So, formally, we define the variance of some random variable X to be

$$\sigma_X^2 = \text{Var}(X) = \sum_{x_i \in \Omega} (x_i - E(X))^2 p_i^X \quad (2)$$

For example, for the number of dots obtained from tossing a dice, we obtain the variance

$$\begin{aligned} \sigma_X^2 &= \text{Var}(X) = \sum_{i=1}^6 (i - E(X))^2 p_i^X \\ &= \frac{1}{6} [(1 - 3.5)^2 + (2 - 3.5)^2 + \dots + (6 - 3.5)^2] \\ &= 2.9167 \end{aligned}$$

Thus, on average, we have to expect a squared deviation from the mean by roughly 2.9.

The *standard deviation* is simply the square root of the variance. Formally, the standard deviation is given by

$$\sigma_X = \sqrt{\text{Var}(X)}$$

The standard deviation appeals to intuition because it is a quantity that is of the same scale as the random variable X . In addition, it helps in assessing where the probability law assigns its probability mass. A rule of thumb is that at least 75% about the probability mass is assigned to a vicinity of the mean that extends two standard deviations in each direction from the mean. Furthermore, this rule states that in at least 89% of the times, a value will occur that lies in a vicinity of the mean of three standard deviations in each direction.

For the number of dots obtained from tossing a dice, since the variance is 2.9167, the standard deviation is

$$\sigma_X = \sqrt{2.9167} = 1.7078$$

In Figure 2, we display all possible outcomes 1 through 6 indicated by the \circ symbol, including the mean of $E(X) = 3.5$. We extend a vicinity about the mean of length $\sigma_X = 1.7078$, indicated by the "+" symbol, to graphically

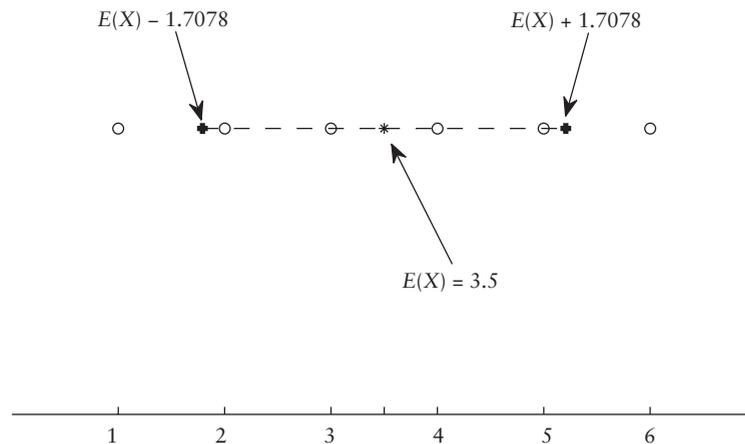


Figure 2 Relation Between Standard Deviation ($\sigma = 1.7078$) and Scale of Possible Outcomes $1, 2, \dots, 6$ Indicated by the \circ Symbol

relate the magnitude of the standard deviation to the possible values of X .

BERNOULLI DISTRIBUTION

In the remainder of this entry, we introduce the most common discrete distributions used in finance. We begin with the simplest one, the *Bernoulli distribution*.

Suppose we have a random variable X with two possible outcomes. That is, we have the state space $\Omega' = \{x_1, x_2\}$. The distribution of X is given by the probability for the two outcomes, that is,

$$p_1^X = p \text{ and } p_2^X = 1 - p$$

Now, to express the random experiment of drawing a value for X , all we need to know is the two possible values in the state space and parameter p representing the probability of x_1 . This situation is represented concisely by the Bernoulli distribution. This distribution is denoted $B(p)$ where p is the probability parameter.

Formally, the Bernoulli distribution is associated with random variables that assume the values $x_1 = 1$ and $x_2 = 0$, or $\Omega' = \{0, 1\}$. That is why this distribution is sometimes referred to as the “zero-one distribution.” One usually sets the parameter p equal to the probability of x_1 such that

$$p = P(X = x_1) = P(X = 1)$$

The mean of a Bernoulli distributed random variable is

$$E(X) = 0 \cdot (1 - p) + 1 \cdot p = p \quad (3)$$

and the variance is

$$\begin{aligned} \text{Var}(X) &= (0 - p)^2 \cdot (1 - p) + (1 - p)^2 \cdot p \\ &= p \cdot (1 - p) \end{aligned} \quad (4)$$

The Bernoulli random variable is commonly used when one models the random experiment where some quantity either satisfies a certain criterion or not. For example, it is employed when it is of interest whether an item is intact or broken. In such applications, we assign the

outcome “success” the numerical value 1 and the outcome “failure” the numerical value 0, for example. Then, we model the random variable X describing the state of the item as Bernoulli distributed.

Consider the outcomes when flipping a coin: head or tail. Now we set head equal to the numerical value 0 and tail equal to 1. We take X as the Bernoulli distributed random variable describing the side of the coin that is up after the toss. What should be considered a fair coin? It would be one where in 50% of the tosses, head should be realized and in the remaining 50% of the tosses, tail should be realized. So, a fair coin yields

$$p = 1 - p = 0.5$$

According to equation (3), the mean is then $E(X) = 0.5$ while, according to equation (4), the variance is $\text{Var}(X) = 0.25$. Here, again, the mean does not represent a possible value x from the state space Ω' . We can interpret it in the following way: Since 0.5 is halfway between one outcome (0) and the other outcome (1), the coin is fair because the mean is not inclined to either outcome.

As another example, we will take a look at credit risk modeling by considering the risk of default of a corporation. Default occurs when the corporation is no longer able to meet its debt obligations. a priori, default occurring during some period is uncertain and, hence, is treated as random. Here, we view the corporation’s failure within the next year as a Bernoulli random variable X . When the corporation defaults, $X = 0$ and in the case of survival, $X = 1$. For example, a corporation may default within the next year with probability

$$P(X = 0) = 1 - p = 1 - e^{-0.04} = 0.0392$$

and survive with probability

$$P(X = 1) = p = e^{-0.04} = 0.9608$$

We can, of course, extend the prerequisites of the Bernoulli distribution to a more general case; that is, we may choose values for the two

outcomes, x_1 and x_2 , of the random variable X different from 0 and 1. Then, we set the parameter p equal to either one of the probabilities $P(X = x_1)$ or $P(X = x_2)$. The distribution yields mean

$$E(X) = x_1 \cdot p + x_2 \cdot (1 - p)$$

and variance

$$\text{Var}(X) = (x_1 - E(X))^2 \cdot p + (x_2 - E(X))^2 \cdot (1 - p)$$

where we set $p = P(X = x_1)$.

We illustrate this generalization of the Bernoulli distribution in the case of the binomial stock price model. Again, we denote the random stock price at time period 1 by S_1 . Recall that the state space $\Omega' = \{\$18, \$22\}$ containing the two possible values for S_1 . The probability of S_1 assuming value \$18 can be set to

$$P(S_1 = \$18) = p$$

so that

$$P(S_1 = \$22) = 1 - p$$

Hence, we have an analogous situation to a Bernoulli random experiment; however, with $\Omega' = \{\$18, \$22\}$ instead of $\Omega' = \{0, 1\}$.

Suppose that

$$P(S_1 = \$18) = p = 0.4 \text{ and}$$

$$P(S_1 = \$22) = 1 - p = 0.6$$

Then, the mean is

$$E(S_1) = 0.4 \cdot \$18 + 0.6 \cdot \$22 = \$20.4$$

and the variance

$$\begin{aligned} \text{Var}(S_1) &= (\$18 - \$20.4)^2 \cdot 0.4 \\ &\quad + (\$22 - \$20.4)^2 \cdot 0.6 = (\$3.84)^2 \end{aligned}$$

BINOMIAL DISTRIBUTION

Suppose that we are no longer interested in whether merely one single item satisfies a particular requirement such as success or failure. Instead, we want to know the number of items satisfying this requirement in a sample of n items. That is, we form the sum over all items

in the sample by adding 1 for each item that is success and 0 otherwise. For example, it could be the number of corporations that satisfy their debt obligation in the current year from a sample of 30 bond issues held in a portfolio. In this case, a corporation would be assigned 1 if it satisfied its debt obligation and 0 if it did not. We would then sum up over all 30 bond issues in the portfolio.

Now, one might realize that this is the linking of n single Bernoulli trials. In other words, we perform a random experiment with n "independent" and identically distributed Bernoulli random variables, which we denote by $B(p)$. Note that we introduced two important assumptions: independent random variables and identically distributed random variables. Independent random variables or independence is an important statistical concept that requires a formal definition. We will not provide one here. Instead, we will simply relate independence to an intuitive interpretation such as uninfluenced by another factor or factors. So in the Bernoulli trials, we assume independence, which means that the outcome of a certain item does not influence the outcome of any others. By identical distribution we mean that the two random variables' distributions are the same. In our context, it implies that for each item, we have the same $B(p)$ distribution.

This experiment is as if one draws an item from a bin and replaces it into the bin before drawing the next item. Thus, this experiment is sometimes referred to as *drawing with replacement*. All we need to know is the number of trials, n , and the parameter p related to each single drawing. The resulting sum of the Bernoulli random variables is distributed as a *binomial distribution* with parameters n and p and denoted by $B(n, p)$.

Let X be distributed $B(n, p)$. Then, the random variable X assumes values in the state space $\Omega' = \{0, 1, 2, \dots, n\}$. In words, the total X is equal to the number of items satisfying the particular requirement (i.e., having a value of 1). X has some integer value i of at least 0 and at most n .

To determine the probability of X being equal to i , we first need to answer the following question: How many different samples of size n can yield a total of i hits (i.e., realizations of the outcome i)? The notation to represent realizing i hits out of a sample of size n is

$$\binom{n}{i} \tag{5}$$

The expression in equation (5) is called the *binomial coefficient* and is explained in Appendix B of this entry.

Since in each sample the n individual $B(p)$ distributed items are drawn independently, the probability of the sum over these n items is the product of the probabilities of the outcomes of the individual items. We illustrate this in the next example.

Suppose we flip a fair coin 10 times (i.e., $n = 10$) and denote by Y_i the result of the i -th trial. We denote by $Y_i = 1$ that the i -th trial produced head and by $Y_i = 0$ that it produced tail. Assume we obtain the following result

Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
1	1	0	0	0	1	0	1	1	0

So, we observe $X = 5$ times head. For this particular result that yields $X = 5$, the probability is

$$\begin{aligned} P(Y_1 = 1, Y_2 = 1, \dots, Y_{10} = 0) &= P(Y_1 = 1) \cdot P(Y_2 = 1) \cdot \dots \cdot P(Y_{10} = 0) \\ &= p \cdot p \cdot \dots \cdot (1 - p) \\ &= p^5 \cdot (1 - p)^5 \end{aligned}$$

Since we are dealing with a fair coin (i.e., $p = 0.5$), the above probability is

$$\begin{aligned} P(Y_1 = 1, Y_2 = 1, \dots, Y_{10} = 0) &= 0.5^5 \cdot 0.5^5 \\ &= 0.5^{10} \approx 0.0010 \end{aligned}$$

With

$$\binom{10}{5} = 252$$

different samples leading to $X = 5$, we compute the probability for this value of the total as

$$\begin{aligned} P(X = 5) &= \binom{10}{5} p^5 \cdot (1 - p)^5 \\ &= 252 \cdot 0.5^{10} = 0.2461 \end{aligned}$$

So, in roughly one fourth of all samples of $n = 10$ independent coin tosses, we obtain a total of $X = 5$ 1s (or heads).

From the example, we see that the exponent for p is equal to the value of the total X (i.e., $i = 5$), and the exponent for $1 - p$ is equal to $n - i = 5$.

Let p be the parameter from the related Bernoulli distribution (i.e., $P(X = 1) = p$). The probability of the $B(n, p)$ random variable X being equal to some $i \in \Omega'$ is given by

$$P(X = i) = \binom{n}{i} \cdot p^i \cdot (1 - p)^{n-i}, i = 1, 2, \dots, n \tag{6}$$

For a particular selection of parameters, the probability distribution at certain values can be found in the four tables in Appendix A.

The mean of a $B(n, p)$ random variable is

$$E(X) = n \cdot p \tag{7}$$

and its variance is

$$Var(X) = n \cdot p \cdot (1 - p) \tag{8}$$

Below we will apply what we have just learned to be the binomial stock price model and two other applications.

Application to the Binomial Stock Price Model

Let's extend the binomial stock price model in the sense that we link T successive periods during which the stock price evolves. (The entire time span of length T is subdivided into the adjacent period segments $(0,1], (1,2], \dots, (T - 1, T].$) In each period $(t, t + 1]$, the price either increases or decreases by, say, 10%. The 10% can be intuitively thought of as the variability of the stock price S . Thus, the

corresponding factor by which the price will change from the previous period is 0.9 (down movement) and 1.1 (up movement). Based on this assumption about the price movement for the stock each period, at the end of the period $(t, t + 1]$, the stock price is

$$S_{t+1} = S_t \cdot Y_{t+1}$$

where the random variable Y_{t+1} assumes a value from $\{0.9, 1.1\}$, with 0.9 representing a price decrease of 10% and 1.1 a price increase of 10%. Consequently, in the case of $Y_{t+1} = 1.1$, we have

$$S_{t+1} = S_t \cdot 1.1$$

while, in case of $Y_{t+1} = 0.9$, we have

$$S_{t+1} = S_t \cdot 0.9$$

For purposes of this illustration, let's assume the following probabilities for the down movement and up movement, respectively,

$$P(Y_{t+1} = 1.1) = p = 0.6$$

and

$$P(Y_{t+1} = 0.9) = 1 - p = 0.4$$

After T periods, we have a random total of X up movements; that is, for all periods $(0,1]$, $(1,2]$, \dots , and $(T - 1, T]$, we increment X by 1 if the period related factor $Y_{t+1} = 1.1$, $t = 0, 1, \dots, T - 1$. So, the result is some $x \in \{1, 2, \dots, T\}$. The total number of up movements, X , is a binomial distributed $B(T, p)$ random variable on the probability space $(\Omega', \mathbb{A}', P^X)$ where

1. The state space is $\Omega' = \{1, 2, \dots, T\}$.
2. σ -algebra \mathbb{A}' is given by the power set $2^{\Omega'}$ of Ω' .
3. P^X is denoted by the binomial probability distribution given by

$$P(X = k) = \binom{T}{k} p^k (1 - p)^{T-k}, k = 1, 2, \dots, T$$

with $p = 0.6$.

Consequently, according to equations (7) and (8), we have

$$E(X) = 2 \cdot 0.6 = 1.2$$

and

$$\text{Var}(X) = 2 \cdot 0.6 \cdot 0.4 = 0.48$$

By definition of S_T and X , we know that the evolution of the stock price is such that

$$S_T = S_0 \cdot 1.1^X \cdot 0.9^{T-X}$$

Let us next consider a random variable that is not binomial itself, but related to a binomial random variable. Now, instead of considering the $B(T, p)$ distributed total X , we could introduce, as a random variable, the stock price at T (i.e., S_T). Using an illustration, we will derive the stock price independently of X and, then, emphasize the relationship between S_T and X . Note that S_T is not a binomial random variable.

Let us set $T = 2$. We may start with an initial stock price of $S_0 = \$20$. At the end of the first period, that is, $(0,1]$, we have

$$S_1 = S_0 \cdot Y_1$$

either equal to

$$S_1 = \$20 \cdot 1.1 = \$22$$

or

$$S_1 = \$20 \cdot 0.9 = \$18$$

At the end of the second period, that is, $(1,2]$, we have

$$S_2 = S_1 \cdot Y_2 = \$22 \cdot 1.1 = \$24.20$$

or

$$S_2 = S_1 \cdot Y_2 = \$22 \cdot 0.9 = \$19.80$$

in the case where $S_1 = \$22$, and

$$S_2 = S_1 \cdot Y_2 = \$18 \cdot 1.1 = \$19.80$$

or

$$S_2 = S_1 \cdot Y_2 = \$18 \cdot 0.9 = \$16.20$$

in the case where $S_1 = \$18$.

That is, at time $t + 1 = T = 2$, we have three possible values for S_2 , namely, \$24.20, \$19.80,

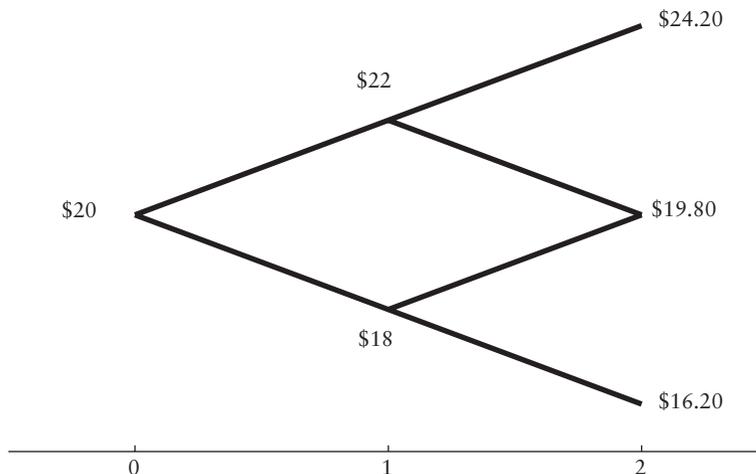


Figure 3 Binomial Stock Price Model with Two Periods

Note: Starting price $S_0 = \$20$. Upward factor $u = 1.1$, downward $d = 0.9$.

and \$16.20. Hence, we have a new state space that we will denote by $\Omega'_S = \{\$16.2, \$19.8, \$24.2\}$. Note that $S_2 = \$19.80$ can be achieved in two different ways: (1) $S_1 = S_0 \cdot 1.1 \cdot 0.9$ and (2) $S_1 = S_0 \cdot 0.9 \cdot 1.1$. The evolution of this pricing process, between time 0 and $T = 2$, can be demonstrated using the *binomial tree* given in Figure 3.

As σ -algebra, we use $\mathbb{A} = 2^{\Omega'_S}$, which is the power set of the state space Ω'_S . It includes events such as, for example, “stock price in $T = 2$ no greater than \$19.80,” defined as $E' = \{S_2 \leq \$19.80\}$.

The probability distribution of S_2 is given by the following

$$\begin{aligned} P(S_2 = \$24.20) &= P(Y_1 = 1.1) \cdot P(Y_2 = 1.1) \\ &= \binom{2}{2} p^2 = 0.6^2 = 0.36 \end{aligned}$$

$$\begin{aligned} P(S_2 = \$19.80) &= P(Y_1 = 0.9) \cdot P(Y_2 = 1.1) \\ &\quad + P(Y_1 = 1.1) \cdot P(Y_2 = 0.9) \\ &= 2(1 - p)p = \binom{2}{1} \cdot 0.4 \cdot 0.6 \\ &= 0.48 \end{aligned}$$

$$\begin{aligned} P(S_2 = \$16.20) &= P(Y_1 = 0.9) \cdot P(Y_2 = 0.9) \\ &= \binom{2}{0} (1 - p)^2 = 0.4^2 = 0.16 \end{aligned}$$

We now have the complete probability space of the random variable S_2 . One can see the connection between S_2 and X by the congruency of the probabilities of the individual outcomes, that is,

$$\begin{aligned} P(S_2 = \$24.20) &= P(X = 2) \\ P(S_2 = \$19.80) &= P(X = 1) \\ P(S_2 = \$16.20) &= P(X = 0) \end{aligned}$$

From this, we derive, again, the relationship

$$S_2 = S_0 \cdot 1.1^X \cdot 0.9^{2-X}$$

Thus, even though S_2 , or, generally S_T , is not distributed binomial itself, its probability distribution can be derived from the related binomial random variable X .¹

Application to the Binomial Interest Rate Model

We next consider a binomial interest rate model of short rates, that is, one-period interest rates. Starting in $t = 0$, the short rate evolves over the subsequent two periods as depicted in Figure 4. In $t = 0$, we have $r_0 = 4\%$, which is the short rate for period 1. For the following period, period 2, the short rate is r_1 while finally, r_2 is valid for

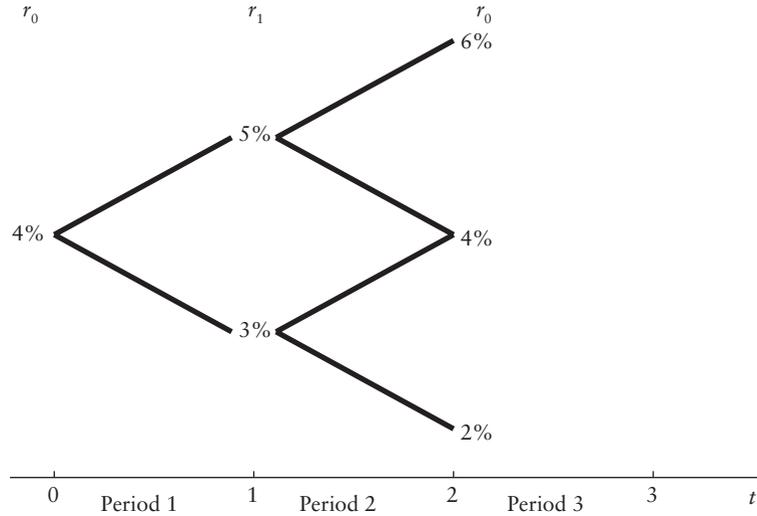


Figure 4 Binomial Interest Rate Model

period 3, from $t = 2$ through $t = 3$. Both r_1 and r_2 are unknown in advance and assume values at random.

As we see, in each of the successive periods, the short rate either increases or decreases by 1% (i.e., 100 basis points). Each movement is assumed to occur with a probability of 50%. So, in period i , $i = 1, 2$, the change in interest rate, Δr_i , has $P(\Delta r_i = 1\%) = p = 0.5$ for an up-movement and $P(\Delta r_i = -1\%) = 1 - p = 0.5$ for a down-movement. For each period, we may model the interest rate change by some Bernoulli random variable where X_1 denotes the random change in period 1 and X_2 that of period 2. The $X_i = 1$ in case of an up-movement and $X_i = 0$ otherwise. The sum of both (i.e., $Y = X_1 + X_2$) is a binomially distributed random variable, precisely $Y \sim B(2, 0.5)$, thus, assuming values 0, 1, or 2.

To be able to interpret the outcome of Y in terms of interest rate changes, we perform the following transformations. A value of $X_i = 1$ yields $\Delta r_i = 1\%$ while $X_i = 0$ translates into $\Delta r_i = -1\%$. Hence, the relationship between Y and r_2 is such that when $Y = 0$, implying two down-movements in a row, $r_2 = r_0 - 2\% = 2\%$. When $Y = 1$, implying one up- and down-movement each, $r_2 = r_0 + 1\% - 1\% = 4\%$. And finally, $Y = 2$ corresponds to two up-movements such

that $r_2 = r_0 + 2\% = 6\%$. So, we obtain the probability distribution:

r_2	$P(r_2)$
2%	$\binom{2}{0} 0.5^0 \cdot 0.5^2 = 0.25$
4%	$\binom{2}{1} 0.5^1 \cdot 0.5^1 = 0.5$
6%	$\binom{2}{2} 0.5^2 \cdot 0.5^0 = 0.25$

HYPERGEOMETRIC DISTRIBUTION

Recall that the prerequisites to obtain a binomial $B(n, p)$ random variable X is that we have n identically distributed random variables Y_i , all following the same Bernoulli law $B(p)$ of which the sum is the binomial random variable X . We referred to this type of random experiment as “drawing with replacement” so that for the sequence of individual drawings Y_i , we always have the same conditions.

Suppose instead that we do not “replace.” Let’s consider the distribution of “drawing without replacement.” This is best illustrated with an urn containing N balls, K of which are

black and $N - K$ are white. So, for the initial drawing, we have the chance of drawing a black ball equal to K/N , while we have the chance of drawing a white ball equal to $(N - K)/N$. Suppose the first drawing yields a black ball. Since we do not replace it, the condition before the second drawing is such that we have $(K - 1)$ black balls and still $(N - K)$ white balls. Since the number of black balls has been reduced by one and the number of white balls is unchanged, the chance of drawing a black ball has been reduced compared to the chance of drawing a white ball; the total is also reduced by one. Hence, the condition is different from the first drawing. It would be similar if instead we had drawn a white ball in the first drawing, however, with the adverse effect on the chance to draw a white ball in the second drawing.

Now suppose in the second drawing another black ball is selected. The chances are increasingly adverse against drawing another black ball in the third trial. This changing environment would be impossible in the binomial model of identical conditions in each trial.

Even if we had drawn first a black ball and then a white ball, the chances would not be the same as at the outset of the experiment before any balls were drawn because the total is now reduced to $N - 2$ balls. So, the chance of obtaining a black ball is now $(K - 1)/(N - 2)$, and that of obtaining a white ball is $(N - K - 1)/(N - 2)$. Mathematically, this is not the same as the original K/N and $(N - K)/N$. Hence, the conditions are altering from one drawing (or trial) to the next.

Suppose now that we are interested in the sum X of black balls drawn in a total of n trials. Let's look at this situation. We begin our reasoning with some illustration given specific values, that is,

$$\begin{aligned} N &= 10 \\ K &= 4 \\ n &= 5 \\ k &= 3 \end{aligned}$$

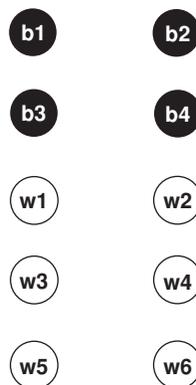


Figure 5 Drawing $n = 5$ Balls without Replacement
Note: $N = 10$, $K = 4$ (black), $n = 5$, and $k = 3$ (black).

The urn containing the black and white balls is depicted in Figure 5. Let's first compute the number of different outcomes we have to consider when we draw $n = 5$ out of $N = 10$ balls regardless of any color. We have 10 different options to draw the first ball; that is, $b1$ through $w6$ in Figure 5. After the first ball has been drawn without replacement, the second ball can be drawn from the urn consisting of the remaining nine balls. After that, the third ball is one out of the remaining eight, and so on until five balls have been successively removed. In total, we have

$$10 \times 9 \times 8 \times 7 \times 6 = 10!/5! = 30,240$$

alternative ways to withdraw the five balls. For example, we may draw $b4, b2, b1, w3$, and $w6$. However, this is the same as $w6, w3, b4, b2$, and $b1$ or any other combination of these five balls. Since we do not care about the exact order of the balls drawn, we have to account for that in that we divide the total number of possibilities (i.e., 30,240) by the number of possible combinations of the very same balls drawn. The latter is equal to

$$5 \times 4 \times 3 \times 2 \times 1 = 5! = 120$$

Thus, we have $30,240/120 = 252$ different nonredundant outcomes if we draw five out of

10 balls. Alternatively, this can be written as

$$252 = \frac{10!}{5! \times 5!} = \binom{10}{5} \quad (9)$$

Consequently, the chance of obtaining exactly this set of balls (i.e., $\{b1, b2, b4, w3, w6\}$) in any order is given by the inverse of equation (9) which is

$$\frac{1}{252} = \frac{1}{\binom{10}{5}} = 0.004 \quad (10)$$

Now recall that we are interested in the chance of obtaining a certain number k of black balls in our sample. So, we have to narrow down the number of possible outcomes given by equation (9) to all samples of size 5 that yield that number k which, here, is equal to 3. How do we do this?

We have a selection of four black balls (i.e., $b1, b2, b3$, and $b4$) to draw from. That gives us a total of $4 \times 3 \times 2 = 4! = 24$ different possibilities to recover $k = 3$ black balls out of the urn consisting of four balls. Again, we do not care about the exact order in which we draw the black balls. To us, it is the same whether we select them, for example, in the order $b1 - b2 - b4$ or $b2 - b4 - b1$, as long as we obtain the set $\{b1, b2, b4\}$. So, we correct for this by dividing the total of 24 by the number of combinations to order these particular black balls; that is,

$$3 \times 2 \times 1 = 3! = 6$$

Hence, the number of combinations of drawing $k = 3$ black balls out of four is

$$24/6 = 4!/3! = 4$$

Next we need to consider the previous number of possibilities of drawing $k = 3$ black balls in combination with drawing $n - k = 2$ white balls. We apply the same reasoning as before to obtain two white balls from the collection of six (i.e., $\{w1, w2, w3, w4, w5, w6\}$). That gives us $6 \times 5/2 = 6!/(2! \times 4!) = 15$ nonredundant options to recover two white balls, in our example.

In total, we have

$$\begin{aligned} 4 \times 15 &= \frac{4 \times 3 \times 2 \times 1}{3 \times 2 \times 1} \times \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 4 \times 3 \times 2 \times 1} \\ &= \frac{4!}{3! \times 1!} \times \frac{6!}{2! \times 4!} = \binom{4}{3} \times \binom{6}{2} = 60 \end{aligned}$$

different possibilities to obtain three black and two white balls in a sample of five balls. All these 60 samples have the same implication for us (i.e., $k = 3$). Combining these 60 possibilities with a probability of 0.004 as given by equation (10), we obtain as the probability for a sum of $k = 3$ black balls in a sample of $n = 5$

$$60/252 = 0.2381$$

Formally, we have

$$P(X = 3) = \frac{\binom{4}{3} \binom{6}{2}}{\binom{10}{5}} = 0.2381$$

Then, for our example, the probability distribution of X is

$$P(X = k) = \frac{\binom{4}{k} \binom{6}{n-k}}{\binom{10}{5}}, k = 1, 2, 3, 4 \quad (11)$$

(Note that we cannot draw more than four black balls from $b1, b2, b3$, and $b4$.)

Let's advance from the special conditions of the example to the general case; that is, (1) at the beginning, some nonnegative integer N of black and white balls combined, (2) the overall number of black balls $0 \leq K \leq N$, (3) the sample size $0 \leq n \leq N$, and (4) the number $0 \leq k \leq n$ of black balls in the sample.

In equation (11), we have the probability of k black balls in the sample of $n = 5$ balls. We dissect equation (9) into three parts: the denominator and the two parts forming the product in the numerator. The denominator gives the number of possibilities to draw a sample of $n = 5$ balls out of $N = 10$ balls, no matter what the combination of black and white. In other words, we choose $n = 5$ out of $N = 10$. The

resulting number is given by the binomial coefficient. We can extend this to choosing a general sample of n drawings out of a population of an arbitrary number of N balls. Analogous to equation (9), the resulting number of possible samples of length n (i.e., n drawings) is then given by

$$\binom{N}{n} \quad (12)$$

Next, suppose we have k black balls in this sample. We have to consider that in equation (11), we chose k black balls from a population of $K = 4$ yielding as the number of possibilities for this the binomial coefficient on the left-hand side in the numerator. Now we generalize this by replacing $K = 4$ by some general number of black balls ($K \leq N$) in the population. The resulting number of choices for choosing k out of the overall K black balls is then,

$$\binom{K}{k} \quad (13)$$

And, finally, we have to draw the remaining $n - k$ balls, which have to be white, from the population of $N - K$ white balls. This gives us

$$\binom{N - K}{n - k} \quad (14)$$

different nonredundant choices for choosing $n - k$ white balls out of $N - K$.

Finally, all we need to do is to combine equations (12), (13), and (14) in the same fashion as equation (11). By doing so, we obtain

$$P(X = k) = \frac{\binom{K}{k} \binom{N - K}{n - k}}{\binom{N}{n}}, \quad k = 1, 2, \dots, n \quad (15)$$

as the probability to obtain a total of $X = k$ black balls in the sample of length n *without replacement*.

Importantly, here, we start out with N balls of which K are black and, after each trial, we do not replace the ball drawn, so that the population

is different for each trial. The resulting random variable is *hypergeometric* distributed with parameters (N, K, n) ; that is, $Hyp(N, K, n)$, and probability distribution given by equation (15).

The mean of a random variable X following a hypergeometric probability law is given by

$$E(X) = n \cdot \frac{K}{N}$$

and the variance of this $X \sim Hyp(N, K, n)$ is given by

$$Var(X) = \sigma^2 = n \cdot \frac{K}{N} \cdot \frac{N - K}{N} \cdot \frac{N - n}{N - 1}$$

The hypergeometric and the binomial distributions are similar, though not equivalent. However, if the population size N is large, the hypergeometric distribution is often approximated by the binomial distribution with equation (6) causing only little deviation from the true probabilities of equation (15).

Application

Let's see how the hypergeometric distribution has been applied in a Federal Reserve Bank of Cleveland study by Humpage (1998) to assess whether U.S. exchange-rate intervention resulted in a desired depreciation of the dollar.

Consider the following scenario. The U.S. dollar is appreciating against a certain foreign currency. This might hurt U.S. exports to the country whose sovereign issues the particular foreign currency. In response, the U.S. Federal Reserve might be inclined to intervene by purchasing that foreign currency to help depreciate the U.S. dollar through the increased demand for foreign currency relative to the dollar. This strategy, however, may not necessarily produce the desired effect. That is, the dollar might continue to appreciate relative to the foreign currency. Let's let an intervention by the Federal Reserve be defined as the purchase of that foreign currency. Suppose that we let the random variable X be number of interventions that lead to success (i.e., depreciation of the

dollar). Given certain conditions beyond the scope of this book, the random variable X is approximately distributed hypergeometric.

This can be understood by the following slightly simplified presentation. Let the number of total observations be N days of which K is the number of days with a dollar depreciation (with or without intervention), and $N - K$ is the number of days where the dollar appreciated or remained unchanged. The number of days the Federal Reserve intervenes is given by n . Furthermore, let k equal the number of days the interventions are successful so that $n - k$ accounts for the unsuccessful interventions. The Federal Reserve could technically intervene on all N days that would yield a total of K successes and $N - K$ failures. However, the actual number of occasions n on which there are interventions might be smaller. The n interventions can be treated as a sample of length n taken from the total of N days without replacement.

The model can best be understood as follows. The observed dollar appreciations, persistence, or depreciations are given observations. The Federal Reserve can merely decide to intervene or not. Consequently, if it took action on a day with depreciation, it would be considered a success and the number of successes available for future attempts would, therefore, be diminished by one. If, on the other hand, the Federal Reserve decided to intervene on a day with appreciation or persistence, it would incur a failure that would reduce the number of available failures left by one. The $N - n$ days there are no interventions are treated as not belonging to the sample.

The randomness is in the selection of the days on which to intervene. The entire process can be illustrated by a chain with N tags attached to it containing either a $+$ or $-$ symbol. Each tag represents one day. A $+$ corresponds to an appreciation or persistence of the dollar on the associated day, while a $-$ to a depreciation. We assume that we do not know the symbol behind each tag at this point.

In total, we have K tags with a $+$ and $N - K$ with a $-$ tag. At random, we flip n of these tags, which is equivalent to the Federal Reserve taking action on the respective days. Upon turning the respective tag upside right, the contained symbol reveals immediately whether the associated intervention resulted in a success or not.

Suppose we have $N = 3,072$ total observations of which $K = 1,546$ represents the number of days with a dollar depreciation, while on $N - K = 1,508$ days the dollar either became more valuable or remained steady relative to the foreign currency.

Again, let X be the hypergeometric random variable describing successful interventions. On $n = 138$ days, the Federal Reserve saw reason to intervene, that is, purchase foreign currency to help bring down the value of the dollar which was successful on $k = 51$ days and unsuccessful on the remaining $n - k = 87$ days. Concisely, the values are given by $N = 3,072$, $K = 1,546$, $N - K = 1,508$, $n = 138$, $k = 51$, and $n - k = 87$.

So, the probability for this particular outcome $k = 51$ for the number of successes X given $n = 138$ trials is

$$P(X = 51) = \frac{\binom{1546}{51} \binom{1508}{87}}{\binom{3072}{138}} = 0.00013429$$

which is an extremely small probability.

Suppose we state the simplifying hypothesis that the Federal Reserve is overall successful if most of the dollar depreciations have been the result of interventions (i.e., purchase of foreign currency). Then, this outcome with $k = 51$ successful interventions given a total of $N - K$ depreciations shows that the decline of the dollar relative to the foreign currency might be the result of something other than a Federal Reserve intervention. Hence, the Federal Reserve intervention might be too vague a forecast of a downward movement of the dollar relative to the foreign currency.

MULTINOMIAL DISTRIBUTION

For our next distribution, the *multinomial distribution*, we return to the realm of drawing with replacement so that for each trial, there are exactly the same conditions. That is, we are dealing with independent and identically distributed random variables. (Once again we note that we are still short of a formal definition of independence in the context of probability theory. We use the term in the sense of “uninfluenced by.”) However, unlike the binomial distribution, let’s change the population so that we have not only two different possible outcomes for one drawing, but a third or possibly more outcomes.

We extend the illustration where we used an urn containing black and white balls. In our extension, we have a total of N balls with three colors: K_w white balls, K_b black balls, and $K_r = N - K_w - K_b$ red balls. The probability of each of these colors is denoted by

$$\begin{aligned} P(Y = \text{white}) &= p_w \\ P(Y = \text{black}) &= p_b \\ P(Y = \text{red}) &= p_r \end{aligned}$$

with each of these probabilities representing the population share of the respective color: $p_i = K_i/N$, for $i = \text{white, black, and red}$. Since all shares combined have to account for all N , we set

$$p_r = 1 - p_b - p_w$$

For purposes of this illustration, let $p_w = p_b = 0.3$ and $p_r = 0.4$. Suppose that in a sample of $n = 10$ trials, we obtain the following result: $n_w = 3$ white, $n_b = 4$ black, and $n_r = n - n_w - n_b = 3$ red. Furthermore, suppose that the balls were drawn in the following order

Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
r	w	b	b	w	r	r	b	w	b

where the random variable Y_i represents the outcome of the i -th trial. (We denote $w = \text{white}$,

$b = \text{black}$, and $r = \text{red}$.) This particular sample occurs with probability

$$\begin{aligned} P(Y_1 = r, Y_2 = w, \dots, Y_{10} = b) &= p_r \cdot p_w \cdot \dots \cdot p_b \\ &= p_r^3 \cdot p_w^3 \cdot p_b^4 \end{aligned}$$

The last equality indicates that the order of appearance of the individual values, once again, does not matter.

We introduce the random variable X representing the number of the individual colors occurring in the sample. That is, X consists of the three components $X_w, X_b,$ and X_r or, alternatively, $X = (X_w, X_b, X_r)$. Analogous to the binomial case of two colors, we are not interested in the order of appearance, but only in the respective numbers of occurrences of the different colors (i.e., $n_w, n_b,$ and n_r). Note that several different sample outcomes may lead to $X = (n_w, n_b, n_r)$. The total number of different nonredundant samples with $n_w, n_b,$ and n_r is given by the multinomial coefficient introduced in Appendix B, which here yields

$$\binom{n}{n_w \ n_b \ n_r} = \binom{10}{3 \ 3 \ 4} = 4,200$$

Hence, the probability for this value of $X = (k_w, k_b, k_r) = (3, 4, 3)$ is then

$$\begin{aligned} P(X = (3, 4, 3)) &= \binom{10}{3 \ 3 \ 4} \cdot p_w^3 \cdot p_b^4 \cdot p_r^3 \\ &= 4,200 \cdot 0.3^3 \cdot 0.3^4 \cdot 0.4^3 \\ &= 0.0588 \end{aligned}$$

In general, the probability distribution of a multinomial random variable X with k components X_1, X_2, \dots, X_k is given by

$$\begin{aligned} P(X_1 = n_1, X_2 = n_2, \dots, X_k = n_k) \\ = \binom{n}{n_1 \ n_2 \ \dots \ n_k} \cdot p_1^{n_1} \cdot p_2^{n_2} \cdot \dots \cdot p_k^{n_k} \end{aligned} \quad (16)$$

where, for $j = 1, 2, \dots, k$, n_j denotes the outcome of component j and the p_j the corresponding probability.

The means of the k components X_1 through X_k are given by

$$\begin{aligned} E(X_1) &= p_1 \cdot n \\ &\vdots \\ E(X_k) &= p_k \cdot n \end{aligned}$$

and their respective variances by

$$\begin{aligned} \text{Var}(X_1) &= \sigma_1^2 = p_1 \cdot (1 - p_1) \cdot n \\ &\vdots \\ \text{Var}(X_k) &= \sigma_k^2 = p_k \cdot (1 - p_k) \cdot n \end{aligned}$$

Multinomial Stock Price Model

We can use the multinomial distribution to extend the binomial stock price model described earlier. Suppose we are given a stock with price S_0 , in $t = 0$. In $t = 1$, the stock can have either price

$$\begin{aligned} S_1^{(u)} &= S_0 \cdot u \\ S_1^{(l)} &= S_0 \cdot l \\ S_1^{(d)} &= S_0 \cdot d \end{aligned}$$

Let the three possible outcomes be a 10% increase in price ($u = 1.1$), no change in price ($l = 1.0$), and a 10% decline in price ($d = 0.9$). That is, the price either goes up by some factor, remains steady, or drops by some factor. Therefore,

$$\begin{aligned} S_1^{(u)} &= S_0 \cdot 1.1 \\ S_1^{(l)} &= S_0 \cdot 1.0 \\ S_1^{(d)} &= S_0 \cdot 0.9 \end{aligned}$$

Thus, we have three different outcomes of the price change in the first period. Suppose the price change behaved the same in the second period, from $t = 1$ until $t = 2$. So, we have

$$\begin{aligned} S_2^{(u)} &= S_1 \cdot 1.1 \\ S_2^{(l)} &= S_1 \cdot 1.0 \\ S_2^{(d)} &= S_1 \cdot 0.9 \end{aligned}$$

at time $t = 2$ depending on

$$S_1 \in \left\{ S_1^{(u)}, S_1^{(l)}, S_1^{(d)} \right\}$$

Let's denote the random price change in the first period by Y_1 and the price change in the second period by the random variable Y_2 . So, it is obvious that Y_1 and Y_2 independently assume some value in the set $\{u, l, d\} = \{1.1, 1.0, 0.9\}$. After two periods (i.e., in $t = 2$), the stock price is

$$S_2 = S_0 \cdot Y_1 \cdot Y_2 \in \left\{ S_2^{(uu)}, S_2^{(ul)}, S_2^{(ld)} \right\}$$

Note that the random variable S_2 is not multinomially distributed itself. However, as we will see, it is immediately linked to a multinomial random variable.

Since the initial stock price S_0 is given, the random variable of interest is the product $Y_1 \cdot Y_2$, which is in a one-to-one relationship with the multinomial random variable $X = (n_u, n_l, n_d)$ (i.e., the number of up-, zero-, and down-movements, respectively). The state space of $Y_1 \cdot Y_2$ is given by $\{uu, ul, ud, ll, ld, dd\}$. This corresponds to the state space of X , which is given by

$$\Omega' = \{(2, 0, 0), (0, 2, 0), (0, 0, 2), (1, 1, 0), (1, 0, 1), (0, 1, 1)\}$$

Note that since $Y_1 \cdot Y_2$ is a product, we do not consider, for example, ($Y_1 = u, Y_2 = d$) and ($Y_1 = d, Y_2 = u$) separately. With

$$\begin{aligned} P(Y_i = u) &= p^u = 0.25 \\ P(Y_i = l) &= p^l = 0.50 \\ P(Y_i = d) &= p^d = 0.25 \end{aligned}$$

the corresponding probability distribution of X is given in the first two columns of Table 1. We use the multinomial coefficient

$$\binom{n}{n_u \quad n_l \quad n_d}$$

where

$$\begin{aligned} n &= \text{the number of periods} \\ n_u &= \text{the number of up-movements} \\ n_l &= \text{number of zero movements} \\ n_d &= \text{number of down-movements} \end{aligned}$$

Now, if $S_0 = \$20$, then we obtain the probability distribution of the stock price in $t = 2$ as shown in columns 2 and 3 in Table 1. Note that

Table 1 Probability Distribution of the Two-Period Stock Price Model

$X = (n_u, n_l, n_d)$	$P(X = \cdot)$	$S_2 = \cdot$
(2,0,0)	$\binom{2}{2 \ 0 \ 0} p^u p^u = 0.0625$	$S_0 \cdot u^2 = 20 \cdot 1.1^2 = 24.2$
(1,1,0)	$\binom{2}{1 \ 1 \ 0} p^u p^l = 2 \cdot 0.25 \cdot 0.5 = 0.25$	$S_0 \cdot u \cdot l = 20 \cdot 1.1 \cdot 1.0 = 22$
(1,0,1)	$\binom{2}{1 \ 0 \ 1} p^u p^d = 2 \cdot 0.25^2 = 0.125$	$S_0 \cdot u \cdot d = 20 \cdot 1.1 \cdot 0.9 = 19.8$
(0,2,0)	$\binom{2}{0 \ 2 \ 0} p^l p^l = 0.5^2 = 0.25$	$S_0 \cdot l \cdot l = 20 \cdot 1.0^2 = 20$
(0,1,1)	$\binom{2}{0 \ 1 \ 1} p^l p^d = 2 \cdot 0.5 \cdot 0.25 = 0.25$	$S_0 \cdot l \cdot d = 20 \cdot 1.0 \cdot 0.9 = 18$
(0,0,2)	$\binom{2}{0 \ 0 \ 2} p^d p^d = 0.25^2 = 0.0625$	$S_0 \cdot d^2 = 20 \cdot 0.9^2 = 16.2$

In the first and second columns, we have the probability distribution of the two period stock price changes $X = Y_1 \cdot Y_2$ in the multinomial stock price model. In the third column, we have the probability distribution of the stock price S_2 .

the probabilities of the values of S_2 are associated with the corresponding price changes X and, hence, listed on the same lines of Table 1. It is now possible to evaluate the probability of events such as, "a stock price S_2 of, at most, \$22," from the σ -algebra \mathbb{A}' of the multinomial probability space of X . This is given by

$$\begin{aligned}
 P(S_2 \leq \$22) &= P(S_2 = \$16.2) + P(S_2 = \$18) + P(S_2 = \$19.8) \\
 &\quad + P(S_2 = \$20) + P(S_2 = \$22)
 \end{aligned}$$

$$\begin{aligned}
 &= 0.25 + 0.125 + 0.25 + 0.25 + 0.0625 \\
 &= 1 - P(S_2 = \$24.2) \\
 &= 0.9375
 \end{aligned}$$

where the second line is the result of the fact that the sum of the probabilities of all disjoint events has to add up to one. That follows since any event and its complement account for the entire state space Ω' .

In Figure 6, we can see the evolution of the stock price along the different paths.

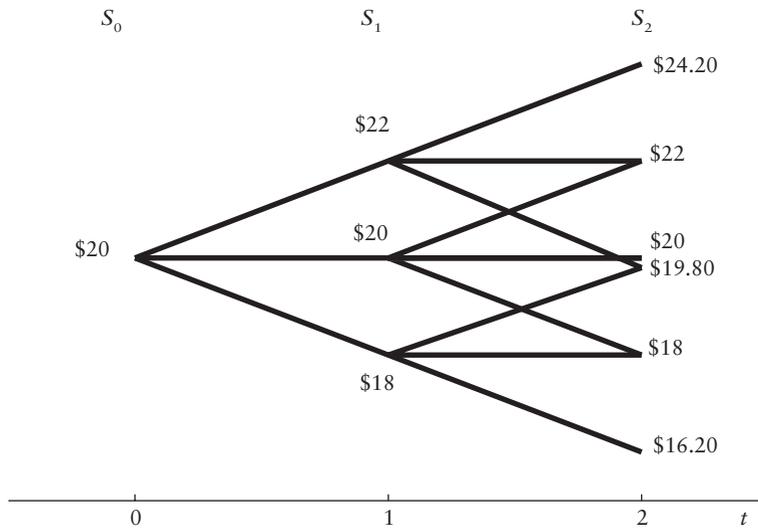


Figure 6 Multinomial Stock Price Model: Stock Price S_2 , in $t = 2$

From equation (1), the expected stock price in $t = 2$ is computed as

$$\begin{aligned} E(S_2) &= \sum_{s \in \Omega'} s \cdot P(S_2 = s) \\ &= \$16.2 \cdot 0.0625 + \$18 \cdot 0.25 + \$19.8 \cdot 0.125 \\ &\quad + \$20 \cdot 0.25 + \$22 \cdot 0.25 + \$24.2 \cdot 0.0625 \\ &= \$20 \end{aligned}$$

So, on average, the stock price will remain unchanged.

POISSON DISTRIBUTION

To introduce our next distribution, consider the following situation. A property and casualty insurer underwrites a particular type of risk, say, automotive damage. Overall, the insurer is interested in the total annual dollar amount of the claims from all policies underwritten. The total is the sum of the individual claims of different amounts. The insurer has to have enough equity as risk guarantee. In a simplified way, the sufficient amount is given by the number of casualties N times the average amount per claim.

In this situation, the insurer's interest is in the total number of claims N within one year. Note that there may be multiple claims per policy. This number N is random because the insurer does not know its exact value at the beginning of the year. The insurer knows, however, that the minimum number of casualties possible is zero. Theoretically, although it is unlikely, there may be infinitely many claims originating from the year of interest.

So far, we have considered the number of claims over the period of one year. It could be of interest to the insurer, however, to know the behavior of the random variable N over a period of different length, say five years, or even the number of casualties related to one month could be of interest. It might be reasonable to assume that there will probably be fewer claims in one month than in one year or five years.

The number of claims, N , as a random variable should follow a probability law that accounts for the length of the period under analysis. In other words, the insurers want to assure that the probability distribution of N gives credit to N being proportional to the length of the period in the sense that if a period is n times as long as another, then the number of claims expected over the longer period should be n times as large, as well.

As a candidate that satisfies these requirements, we introduce the *Poisson distribution* with parameter λ formally expressed as $Poi(\lambda)$. We define that the parameter is a positive real number (i.e., $\lambda > 0$). A Poisson random variable N —that is, $X \sim Poi(\lambda)$ —assumes nonnegative integer values. Formally, N is a function mapping the space of outcomes, Ω , into the state space

$$\Omega' = \{0, 1, 2, \dots\}$$

which is the set \mathbb{N} of the nonnegative integer numbers.

The probability measure of a Poisson random variable N for nonnegative integers $k = 0, 1, 2, \dots$ is defined as

$$P(N = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad (17)$$

where $e = 2.7183$ is the Euler constant. Here, we have unit period length.

The mean of a Poisson random variable with parameter λ is

$$E(N) = \lambda$$

while its variance is given by

$$Var(N) = \sigma^2 = \lambda \quad (18)$$

So, both parameters, mean and variance, of $N \sim Poi(\lambda)$ are given by the parameter λ .

For a period of general length t , equation (17) becomes

$$P(N = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad (19)$$

We can see that the new parameter is now λ_t , accounting for the time proportionality of the distribution of N , that is, $N = N(t)$ is the number of jumps of size 1 in the interval $(0, t)$. The mean changes to

$$EN(t) = \lambda t \quad (20)$$

and analogous to the variance given by (18) is now

$$\text{Var}(N(t)) = \sigma^2(t) = \lambda t \quad (21)$$

We can see by equation (20) that the average number of occurrences is the average per unit of time, λ , times the length of the period, t , in units of time. The same holds for the variance given by equation (21).

The Poisson distribution serves as an approximation of the hypergeometric distribution when certain conditions are met regarding sample size and parameter p .

Application to Credit Risk Modeling for a Bond Portfolio

The Poisson distribution is typically used in finance for credit risk modeling. For example, suppose we have a pool of 100 bonds issued by different corporations. By experience or empirical evidence, we may know that each quarter of a year the expected number to default is two; that is, $\lambda = 2$. Moreover, from prior research, we can approximate the distribution of N by the Poisson distribution, even though, theoretically, the Poisson distribution admits values k greater than 100. What is the number of bonds to default within the next year, on average? According to equation (3), since the mean is $E_{quarter}(N) = \lambda = 2$ per quarter, the mean per year ($t = 4$) is

$$E_{year}(N) = \lambda t = 2 \cdot 4 = 8$$

By equation (20), the variance is 8, from equation (19), the probability of, at most, 10 bonds

to default is given by

$$\begin{aligned} P(N \leq 10) &= P(N = 0) + P(N = 1) + \dots \\ &\quad + P(N = 10) \\ &= e^{-2 \times 4} \cdot \frac{(2 \times 4)^0}{0!} + e^{-2 \times 4} \cdot \frac{(2 \times 4)^1}{1!} + \dots \\ &\quad + e^{-2 \times 4} \cdot \frac{(2 \times 4)^{10}}{10!} \\ &= 0.8159 \end{aligned}$$

DISCRETE UNIFORM DISTRIBUTION

Consider a probability space $(\Omega', \mathbb{A}', P)$ where the state space is a finite set of, say n , outcomes, that is, $\Omega' = \{x_1, x_2, \dots, x_n\}$. The σ -algebra \mathbb{A}' is given by the power set of Ω' .

So far we have explained how drawings from this Ω' may be modeled by the multinomial distribution. In the multinomial distribution, the probability of each outcome may be different. However, suppose that for our random variable X , we have a constant $P(X = x_j) = 1/n$, for all $j = 1, 2, \dots, n$. Since all values x_j have the same probability (i.e., they are equally likely), the distribution is called the *discrete uniform distribution*. We denote this distribution by $X \sim DU_{\Omega'}$. We use the specification Ω' to indicate that X is a random variable on this particular state space.

The mean of a discrete, uniformly distributed random variable X on the state space $\Omega' = \{x_1, x_2, \dots, x_n\}$ is given by

$$E(X) = \sum_{i=1}^n p_i \cdot x_i = \frac{1}{n} \sum_{i=1}^n x_i \quad (22)$$

Note that equation (22) is equal to the arithmetic mean. The variance is

$$\begin{aligned} \text{Var}(X) &= \sum_{i: x_i \in \Omega'} p_i \cdot (x_i - E(X))^2 \\ &= \frac{1}{n} \sum_{i: x_i \in \Omega'} (x_i - E(X))^2 \end{aligned}$$

with $E(X)$ from equation (22).

A special case of a discrete uniform probability space is given when $\Omega' = \{1, 2, \dots, n\}$. The resulting mean, according to equation (22), is then,

$$\begin{aligned} E(X) &= \sum_{i=1}^n p_i \cdot x_i = \frac{1}{n} \sum_{i=1}^n i \\ &= \frac{1}{n} \times \frac{n(n+1)}{2} = \frac{n+1}{2} \end{aligned} \quad (23)$$

For this special case of discrete uniform distribution of a random variable X , we use the notation $X \sim DU(n)$ with parameter n .

Let's once more consider the outcome of a toss of a dice. The random variable number of dots, X , assumes one of the numerical outcomes 1, 2, 3, 4, 5, 6 each with a probability of $1/6$. Hence, we have a uniformly distributed discrete random variable X with the state space $\Omega' = \{1, 2, 3, 4, 5, 6\}$. Consequently, we express this as $X \sim DU(6)$.

Next, we want to consider several independent trials, say $n = 10$, of throwing the dice. By n_1, n_2, n_3, n_4, n_5 , and n_6 , we denote the number of occurrence of the values 1, 2, 3, 4, 5, and 6, respectively. With constant probability $p_1 = p_2 = \dots = p_6 = 1/6$, we have a discrete uniform distribution, that is, $X \sim DU(6)$. Thus, the probability of obtaining $n_1 = 1, n_2 = 2, n_3 = 1, n_4 = 3, n_5 = 1$, and $n_6 = 2$, for example, is

$$\begin{aligned} P(X_1 = 1, X_2 = 1, \dots, X_6 = 2) &= \binom{10}{1 \ 2 \dots 2} \left(\frac{1}{6}\right)^{10} \\ &= \frac{10!}{1! \times 2! \times \dots \times 2!} \cdot \left(\frac{1}{6}\right)^{10} \\ &= 151200 \cdot 0.00000016538 \\ &= 0.0025 \end{aligned}$$

Application to the Multinomial Stock Price Model

Let us resume the stock price model where in $t = 0$ we have a given stock price, say $S_0 = \$20$, where there are three possible outcomes at the end of the period. In the first period, the stock

price either increases to

$$S_1^{(u)} = S_0 \cdot 1.1 = \$22$$

remains the same at

$$S_1^{(l)} = S_0 \cdot 1.0 = \$20$$

or decreases to

$$S_1^{(d)} = S_0 \cdot 0.9 = \$18$$

each with probability $1/3$. Again, we introduce the random variable Y assuming the values $u = 1.1, l = 1.0$, and $d = 0.9$ and, thus, representing the percentage change of the stock price between $t = 0$ and $t + 1 = 1$. The stock price in $t + 1 = 1$ is given by the random variable S_1 on the corresponding state space

$$\Omega_S = \{S_1^{(u)}, S_1^{(l)}, S_1^{(d)}\}$$

Suppose we have $n = 10$ successive periods in each of which the stock price changes by the factors u, l , or d . Let the multinomial random variable $X = (X_1, X_2, X_3)$ represent the total of up-, zero-, and down-movements, respectively. Suppose, after these n periods, we have $n_u = 3$ up-movements, $n_l = 3$ zero-movements, and $n_d = 4$ down-movements. According to equation (16), the corresponding probability is

$$\begin{aligned} P(X_1 = 3, X_2 = 3, X_3 = 4) &= \binom{10}{3 \ 3 \ 4} \left(\frac{1}{3}\right)^{10} \\ &= 4200 \cdot 0.00001935 \\ &= 0.0711 \end{aligned}$$

This probability corresponds to a stock price in $t = 10$ of

$$S_{10} = S_0 \cdot u^3 \cdot l^3 \cdot d^4 = \$20 \cdot 1.1^3 \cdot 1 \cdot 0.9^4 = \$17.47$$

This stock price is a random variable given by

$$S_{10} = S_0 \cdot Y_1 \cdot Y_2 \cdot \dots \cdot Y_{10}$$

where the Y_i are the corresponding relative changes (i.e., factors) in the periods $i = 1, 2, \dots, 10$. Note that S_{10} is not uniformly distributed even though it is a function of the random variables Y_1, Y_2, \dots, Y_{10} because its possible outcomes do not have identical probability.

APPENDIX A LIST OF DISCRETE DISTRIBUTIONS

Name	Probability Law	Mean	Variance	Description
Bernoulli	$P(X=1) = p$ $P(X=0) = 1 - p$	p	$p \cdot (1 - p)$	One drawing from $\Omega' = \{1,0\}$
Binomial	$P(X=k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$	$n \cdot p$	$n \cdot p \cdot (1-p)$	n drawings with replacement from $\Omega' = \{1,0\}$
Hypergeometric	$Hyp(N, K, n)$ $P(X=k) = \frac{\binom{K}{k} \binom{N-k}{n-k}}{\binom{N}{n}}$	$\frac{K}{N} \cdot n$	$\frac{K}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1}$	n drawings without replacement from $\Omega' = \{1,0\}$
Multinomial	$P(X_1 = n_1, \dots, X_k = n_k)$ $= \binom{n}{n_1 \dots n_k} \cdot p_1^{n_1} \cdot \dots \cdot p_k^{n_k}$	$p_1 \cdot n$ \vdots $p_k \cdot n$	$p_1 \cdot (1 - p_1) \cdot n$ \vdots $\sigma_k^2 = p_k \cdot (1 - p_k) \cdot n$	n drawings with replacement from $\Omega' = \{x_1, \dots, x_k\}$
Poisson	$P(N=k) = \frac{\lambda^k}{k!} e^{-\lambda}$	λ	λ	One drawing from $\Omega' = \{0, 1, 2, \dots\} = \mathbb{N}$. Unit period length
	$P(N=k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$	λt	λt	One drawing from $\Omega' = \{0, 1, 2, \dots\} = \mathbb{N}$. Period length t
Discrete Uniform	$DU_{\Omega'}$ $P(X=x_1) = \dots = P(X=x_n) = \frac{1}{n}$	(8.1)	(8.2)	One drawing from $\Omega' = \{x_1, \dots, x_k\}$ Equal probability
	$DU(n)$ $P(X=1) = \dots = P(X=n) = \frac{1}{n}$	$\frac{n+1}{2}$	(8.2)	One drawing from $\Omega' = \{1, \dots, k\}$

Note: For the k components of a multinomial random variable, we have k means and variances.

$B(n, p)$, Binomial Probability Distribution

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k} \text{ for } n = 5$$

k	p	0.1	0.2	0.5	0.8	0.9
1		0.3281	0.4096	0.1563	0.0064	0.0005
2		0.0729	0.2048	0.3125	0.0512	0.0081
3		0.0081	0.0512	0.3125	0.2048	0.0729
4		0.0005	0.0064	0.1563	0.4096	0.3281
5		0	0.0003	0.0313	0.3277	0.5905

$B(n, p)$, Binomial Probability Distribution

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k} \text{ for } n = 10$$

k	p	0.1	0.2	0.5	0.8	0.9
1		0.3874	0.2684	0.0098	0	0
2		0.1937	0.3020	0.0439	0.0001	0
3		0.0574	0.2013	0.1172	0.0008	0
4		0.0112	0.0881	0.2051	0.0055	0.0001
5		0.0015	0.0264	0.2461	0.0264	0.0015
6		0.0001	0.0055	0.2051	0.0881	0.0112
7		0	0.0008	0.1172	0.2013	0.0574
8		0	0.0001	0.0439	0.3020	0.1937
9		0	0	0.0098	0.2684	0.3874
10		0	0	0.0010	0.1074	0.3487

$B(n, p)$, Binomial Probability Distribution

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k} \text{ for } n = 50$$

k	p	0.1	0.2	0.5	0.8	0.9
1		0.0286	0.0002	0	0	0
2		0.0779	0.0011	0	0	0
3		0.1386	0.0044	0	0	0
4		0.1809	0.0128	0	0	0
5		0.1849	0.0295	0	0	0
6		0.1541	0.0554	0	0	0
7		0.1076	0.0870	0	0	0
8		0.0643	0.1169	0	0	0
9		0.0333	0.1364	0	0	0

k	p	0.1	0.2	0.5	0.8	0.9
10		0.0152	0.1398	0	0	0
20		0	0.0006	0.0419	0	0
30		0	0	0.0419	0.0006	0
40		0	0	0	0.1398	0.0152
41		0	0	0	0.1364	0.0333
42		0	0	0	0.1169	0.0643
43		0	0	0	0.0870	0.1076
44		0	0	0	0.0554	0.1541
45		0	0	0	0.0295	0.1849
46		0	0	0	0.0128	0.1809
47		0	0	0	0.0044	0.1386
48		0	0	0	0.0011	0.0779
49		0	0	0	0.0002	0.0286
50		0	0	0	0	0.0052

$B(n, p)$, Binomial Probability Distribution

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k} \text{ for } n = 100$$

k	p	0.1	0.2	0.5	0.8	0.9
1		0.0003	0	0	0	0
2		0.0016	0	0	0	0
3		0.0059	0	0	0	0
4		0.0159	0	0	0	0
5		0.0339	0	0	0	0
6		0.0596	0.0001	0	0	0
7		0.0889	0.0002	0	0	0
8		0.1148	0.0006	0	0	0
9		0.1304	0.0015	0	0	0
10		0.1319	0.0034	0	0	0
20		0.0012	0.0993	0	0	0
30		0	0.0052	0	0	0
40		0	0	0.0108	0	0
50		0	0	0.0796	0	0
60		0	0	0.0108	0	0
70		0	0	0	0.0052	0
80		0	0	0	0.0993	0.0012
90		0	0	0	0.0034	0.1319
91		0	0	0	0.0015	0.1304
92		0	0	0	0.0006	0.1148
93		0	0	0	0.0002	0.0889
94		0	0	0	0.0001	0.0596
95		0	0	0	0	0.0339
96		0	0	0	0	0.0159
97		0	0	0	0	0.0059
98		0	0	0	0	0.0016
99		0	0	0	0	0.0003
100		0	0	0	0	0

$Poi(\lambda)$, Poisson Probability Distribution

$$P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!} \text{ for Several Values of Parameter } \lambda$$

k	λ 0.1	0.5	1	2	5	10
1	0.0905	0.3033	0.3679	0.2707	0.0337	0.0005
2	0.0045	0.0758	0.1839	0.2707	0.0842	0.0023
3	0.0002	0.0126	0.0613	0.1804	0.1404	0.0076
4	0	0.0016	0.0153	0.0902	0.1755	0.0189
5	0	0.0002	0.0031	0.0361	0.1755	0.0378
6	0	0	0.0005	0.0120	0.1462	0.0631
7	0	0	0.0001	0.0034	0.1044	0.0901
8	0	0	0	0.0009	0.0653	0.1126
9	0	0	0	0.0002	0.0363	0.1251
10	0	0	0	0	0.0181	0.1251
11	0	0	0	0	0.0082	0.1137
12	0	0	0	0	0.0034	0.0948
13	0	0	0	0	0.0013	0.0729
14	0	0	0	0	0.0005	0.0521
15	0	0	0	0	0.0002	0.0347
16	0	0	0	0	0	0.0217
17	0	0	0	0	0	0.0128
18	0	0	0	0	0	0.0071
19	0	0	0	0	0	0.0037
20	0	0	0	0	0	0.0019
50	0	0	0	0	0	0
100	0	0	0	0	0	0

N that is $k = 1, 2, 3, \dots$ as

$$k! = k \cdot (k - 1) \cdot (k - 2) \cdot \dots \cdot 1$$

For $k = 0$, we define $0! \equiv 1$.

Derivation of the Binomial Coefficient

In the context of the binomial distribution, we form the sum X of n independent and identically distributed Bernoulli random variables Y_i with parameter p or, formally, $Y_i \stackrel{iid}{\sim} B(p)$, $i = 1, 2, \dots, n$. The random variable is then distributed binomial with parameters n and p , i.e., $X \sim B(n, p)$. Since the random variables Y_i have either value 0 or 1, the resulting binomial random variable (i.e., the sum X) assumes some integer value between 0 and n . Let $X = k$ for $0 \leq k \leq n$. Depending on the exact value k , there may be several alternatives to obtain k since, for the sum X , it is irrelevant in which order the individual values of the Y_i appear.

Special Case $n = 3$

We illustrate the special case where $n = 3$ using a $B(3,0.4)$ random variable X ; that is, X is the sum of three independent $B(0.4)$ distributed random variables Y_1, Y_2 , and Y_3 . All possible values for X are contained in the state space $\Omega' = \{0, 1, 2, 3\}$. As we will see, some of these $k \in \Omega'$ can be obtained in different ways.

We start with $k = 0$. This value can only be obtained when all Y_i are 0, for $i = 1, 2, 3$. So, there is only one possibility.

Next we consider $k = 1$. A sum of $X = 1$ can be the result of one $Y_i = 1$ while the remaining two Y_i are 0. We have three possibilities for $Y_i = 1$ since it could be either the first, the second, or the third of the Bernoulli random variables. Then we place the first 0. For this, we have two possibilities since we have two Y_i left that are not equal to 1. Next, we place the second 0, which we have to assign to the remaining Y_i . As an intermediate result, we have $3 \cdot 2 \cdot 1 = 6$

APPENDIX B BINOMIAL AND MULTINOMIAL COEFFICIENTS

In this appendix, we explain the concept of the binomial and multinomial coefficients used in discrete probability distributions.

BINOMIAL COEFFICIENT

The *binomial coefficient* is defined as

$$\binom{n}{k} = \frac{n!}{k!(n - k)!}$$

for some nonnegative integers k and n with $0 \leq k \leq n$. For the binomial coefficient, we use the *factorial* operator denoted by the “!” symbol. A factorial is defined in the set of natural numbers

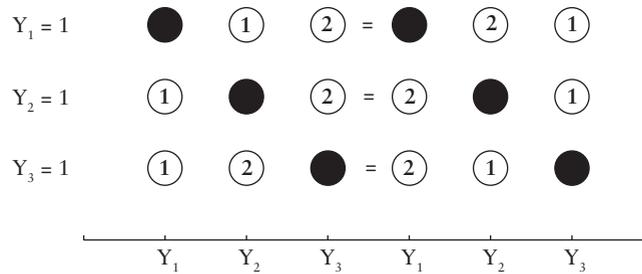


Figure B.1 Three Different Ways to Obtain a Total of $X = \sum_{i=1}^3 Y_i = 1$

Note: The alternatives matched by the = symbol lead to the same outcome, respectively.

possibilities. However, we do not need to differentiate between the two 0 values because it does not matter which of the zeros is assigned first and which second. So, we divide the total number of options by the number of possibilities to place the 0 values (i.e., 2). The resulting number of possible ways to end up with $X = 1$ is

$$\frac{3 \cdot 2 \cdot 1}{2} = \frac{3!}{2! \cdot 1!} = 3$$

For reasons we will make clear later, we introduced the middle term in the above equation.

Let us illustrate this graphically. In Figure B.1, a black ball represents a value $Y_i = 1$ at the i -th drawing while the white numbered circles represent a value of $Y_i = 0$ at the respective i -th drawing with i matching the number in the circle.

Now let $k = 2$. To yield the sum $X = 2$, we need two $Y_i = 1$ and one $Y_i = 0$. So, we have three different positions to place the 0, while the remaining two Y_i have to be equal to 1 automatically. Analogous to the prior case, $X = 1$, we do not need to differentiate between the two 1 values, once the 0 is positioned.

Finally, let $k = 3$. This is accomplished by all three $Y_i = 1$. So, there is only one possibility to obtain $X = 3$.

We summarize these results in Table B.1.

Special Case $n = 4$

We extend the prior case to the case where the random variable X is the sum of four Bernoulli distributed random variables—that is, $Y_i \stackrel{iid}{\sim} B(p), i = 1, 2, 3, 4$ —assuming either value 0 or 1 for each. The resulting sum X is then binomial distributed $B(4, p)$ assuming values k in the state space $\Omega' = \{0,1,2,3,4\}$. Again, we will analyze how the individual values of the sum X can be obtained.

To begin, let us consider the case $k = 0$. As in the prior case $n = 3$, we have only one possibility (i.e., all four Y_i equal to 0, that is, $Y_1 = Y_2 = Y_3 = Y_4 = 0$). This can be seen from the following. Technically, we have four positions to place the first 0. Then, we have three choices to place the second 0. For the third 0, we have two positions available, and one for the last 0. In total, we have

$$4 \times 3 \times 2 \times 1 = 24$$

Table B.1 Different Choices to Obtain $X = k$ when $n = 3$

$k = 0$	$k = 1$	$k = 2$	$k = 3$
$1 = \frac{3!}{0! \times 3!} = \binom{3}{0}$	$3 = \frac{3!}{1! \times 2!} = \binom{3}{1}$	$3 = \frac{3!}{2! \times 1!} = \binom{3}{2}$	$1 = \frac{3!}{3! \times 0!} = \binom{3}{3}$

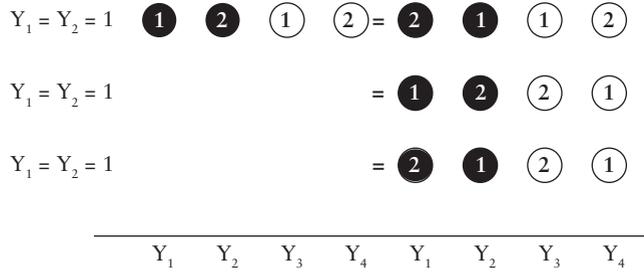


Figure B.2 Four Different Ways to Obtain $Y_1 = Y_2 = 1$

But due to the fact that we do not care about the order of the 0 values, we divide by the total number of options (i.e., 24) and then obtain

$$\frac{4 \times 3 \times 2 \times 1}{4 \times 3 \times 2 \times 1} = \frac{4!}{4!} = 1$$

Next, we derive a sum of $k = 1$. This can be obtained in four different ways. The reasoning is similar to that in the case $k = 1$ for $n = 3$. We have four positions to place the 1. Once the 1 is placed, the remaining Y_i have to be automatically equal to 0. Again, the order of placing the 0 values is irrelevant, which eliminates the redundant options through division of the total number by $3 \times 2 \times 1 = 6$. Technically, we have

$$\frac{4 \times 3 \times 2 \times 1}{3 \times 2 \times 1} = \frac{4!}{3!} = 4$$

For a sum X equal to $k = 2$, we have four different positions to place the first 1. Then, we have three positions left to place the second 1. This yields $4 \times 3 = 12$ different options. However, we do not care which one of the 1 values is placed first since, again, their order is irrelevant. So, we divide the total number by 2 to

indicate that the order of the two 1 values is unimportant. Next, we place the first 0, which offers us two possible positions for the remaining Y_i that are not equal to 1 already. For this, we have two options. In total, we then have

$$\frac{4 \times 3 \times 2 \times 1}{2 \times 1} = \frac{4!}{2!} = 12$$

possibilities. Then, the second 0 is placed on the remaining Y_i . So, there is only one choice for this 0. Because we do not care about the order of placement of the 2 values, we divide by 2. The resulting number of different ways to yield a sum X of $k = 2$ is

$$\frac{4 \times 3 \times 2 \times 1}{2 \times 1 \times 2 \times 1} = \frac{4!}{2! \times 2!} = 6$$

which is illustrated in Figures B.2 through B.7.

A sum of X equal to $k = 3$ is achieved by three 1 values and one 0 value. So, since the order of the 1 values is irrelevant due to the previous reasoning, we only care about where to place the 0 value. We have four possibilities, that is,

$$\frac{4 \times 3 \times 2 \times 1}{3 \times 2 \times 1} = \frac{4!}{3!} = 4$$

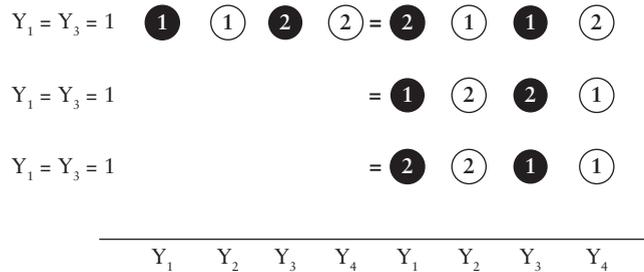


Figure B.3 Four Different Ways to Obtain $Y_1 = Y_3 = 1$

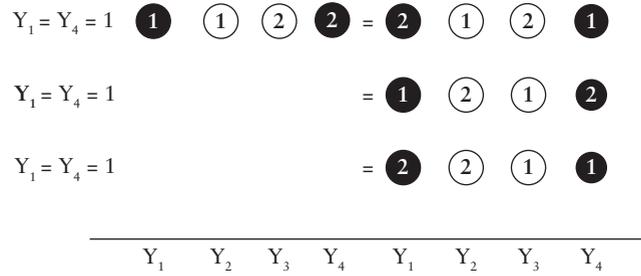


Figure B.4 Four Different Ways to Obtain $Y_1 = Y_4 = 1$

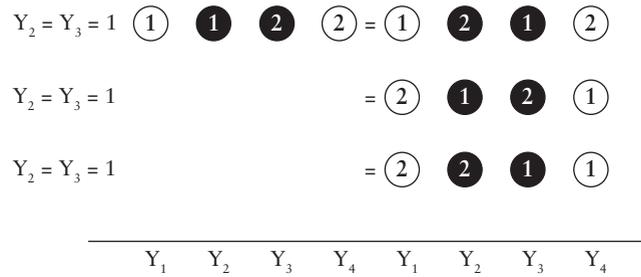


Figure B.5 Four Different Ways to Obtain $Y_2 = Y_3 = 1$

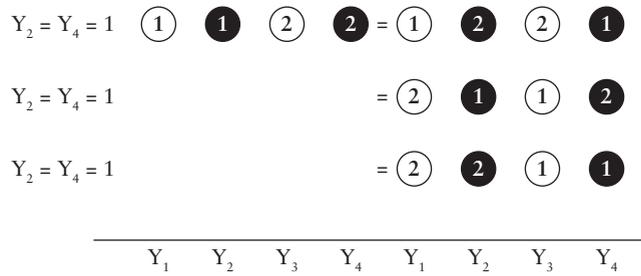


Figure B.6 Four Different Ways to Obtain $Y_2 = Y_4 = 1$

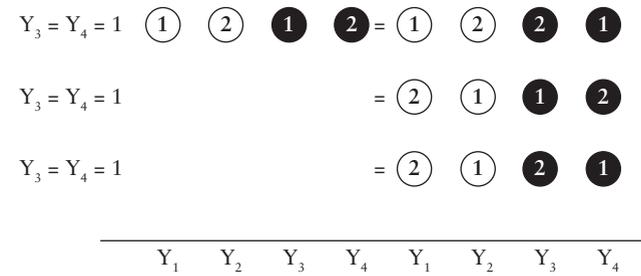


Figure B.7 Four Different Ways to Obtain $Y_3 = Y_4 = 1$

Table B.2 Different Choices to Obtain $X = k$ when $n = 4$

$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$
$1 = \frac{4!}{0! \times 4!} = \binom{4}{0}$	$4 = \frac{4!}{1! \times 3!} = \binom{4}{1}$	$6 = \frac{4!}{2! \times 2!} = \binom{4}{2}$	$4 = \frac{4!}{3! \times 1!} = \binom{4}{3}$	$1 = \frac{4!}{4! \times 0!} = \binom{4}{4}$

Finally, to obtain $k = 4$, we only have one possible way to do so, as in the case where $k = 0$. Mathematically, this is

$$\frac{4 \times 3 \times 2 \times 1}{4 \times 3 \times 2 \times 1} = \frac{4!}{4!} = 1$$

We summarize the results in Table B.2.

General Case

Now we generalize for any $n \in N$ (i.e., some nonnegative integer number). The binomial random variable X is hence the $B(n, p)$ distributed sum of n independent and identically distributed random variables Y_i

From the two special cases (i.e., $n = 3$ and $n = 4$), it seems that to obtain the number of choices for some $0 \leq k \leq n$, we have $n!$ in the numerator to account for all the possibilities to assign the individual n values to the Y_i , no matter how many 1 values and 0 values we have. In the denominator, we correct for the fact that the order of the 1 values and 0 values is irrelevant. That is, we divide by the number of different orders to place the 1 values on the Y_i that are equal to 1, and also by the number of different orders to assign the 0 values to the Y_i being equal to 0. Therefore, we have $n!$ in the

numerator and $k! \times (n - k)!$ in the denominator. The result is illustrated in Table B.3.

MULTINOMIAL COEFFICIENT

The *multinomial coefficient* is defined as

$$\binom{n}{n_1 \ n_2 \ \dots \ n_k} = \frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_k!}$$

for $n_1 + n_2 + \dots + n_k = n$. Sometimes, the multinomial coefficient is referred to as the *polynomial coefficient*.

Assume we have some population of balls with k different colors. Suppose n times we draw some ball and return it to the population such that for each trial (i.e., drawing), we have the identical conditions. Hence, the individual trials are independent of each other. Let Y_i denote the color obtained in the i -th trial for $i = 1, 2, \dots, n$.

How many different possible samples of length n are there? Let us think of the drawings in a different way. That is, we draw one ball after another disregarding color and assign the drawn ball to the trials Y_1 through Y_n in an arbitrary fashion.

Table B.3 Different Choices to Obtain $X = k$ for General n

$k = 0$	$k = 1$	$k = 2$
$1 = \frac{n!}{0! \times n!} = \binom{n}{0}$	$n = \frac{n!}{1! \times (n - 1)!} = \binom{n}{1}$	$\frac{n \times (n - 1)}{2} = \frac{n!}{2! \times (n - 2)!} = \binom{n}{2}$
...	$k = n - 1$	$k = n$
...	$n = \frac{n!}{(n - 1)! \times 1!} = \binom{n}{n - 1}$	$n = \frac{n!}{n! \times 0!} = \binom{n}{n}$

First, we draw a ball with any of the k colors and assign it to one of the n trials, Y_i . Next, we draw the second ball and assign it to one of the remaining $n - 1$ possible trials i as outcome of Y_i . This yields

$$n \times (n - 1)$$

different possibilities. The third ball drawn is assigned to the $n - 2$ trials left so that we have

$$n \times (n - 1) \times (n - 2)$$

possibilities, in total. This is continued until we draw the n th (i.e., the last), color, which can only be placed in the last remaining trial Y_i . In total this yields

$$n \times (n - 1) \times (n - 2) \times \dots \times 2 \times 1 = n!$$

different possibilities of drawing n balls.

The second question is how many different possibilities are there to obtain a sample with the number of occurrences n_1, n_2, \dots , and n_k of the respective colors. Let red be one of these colors and suppose we have a sample with a total of $n_r = 3$ red balls from trials 2, 4, and 7 so that $Y_2 = Y_4 = Y_7 = \text{red}$. The assignment of red to these three trials yields

$$3! = 3 \times 2 \times 1 = 6$$

different orders of assignment. Now, we are indifferent with respect to which of the Y_2, Y_4 , and Y_7 was assigned red first, second, and third. Thus, we divide the total number $n!$ of different samples by $n_r! = 3!$ to obtain only nonredundant results with respect to a red ball. We proceed in the same fashion for the remaining colors and, finally, obtain for the total number of nonredundant samples containing n_1 of color 1, n_2 of color 2, \dots , and n_k of color k

$$\binom{n!}{n_1 n_2 \dots n_k} = \frac{n!}{n_1! \times n_2! \times \dots \times n_k!}$$

which is exactly the multinomial coefficient equation given above.

KEY POINTS

- A discrete law or probability distribution is related to some discrete random variable, that is, a random variable that can assume values from a countable set of values. Typical examples include counts (i.e., the number of items meeting certain requirements) and number of hits.
- The most important discrete random variables used in finance and their probability distribution are the Bernoulli, binomial, hypergeometric, multinomial, Poisson, and discrete uniform.
- The Bernoulli distribution might be the most famous discrete law. It is applied when a random variable can only assume one of two values—0 or 1. A simple example would be the toss of a coin. In financial models, it is applied if it is of interest whether a certain event has occurred (1) or not (0).
- The binomial distribution is the extension of the Bernoulli distribution in the sense that it represents repeated trials where the respective outcomes are either 0 or 1, so that in total we can obtain any integer between 0 and n , where n is the number of Bernoulli trials. A typical example in finance would be given by the binomial stock price model where it is the objective to count the number of improvements of some stock over a given number of periods.
- Drawing with replacement refers to the experiment of repeated trials where each individual trial is conducted under identical conditions as the others and without influencing each other. A prerequisite of the binomial distribution is drawing with replacement.
- The Poisson distribution is related to a discrete random variable that can assume any nonnegative integer value. A typical application is in risk theory when the number of defaults or occurrences of some undesirable event has to be modeled.

NOTE

1. Note that the successive prices S_1, \dots, S_T depend on their respective predecessors. They are said to be *path-dependent*. Only the changes, or factors Y_{t+1} , for each period are independent. In this case, the price S_{t+1} depends only on S_t , however, and not the en-

tire past. This is referred to as the *Markov* property.

REFERENCE

Humpage, O. F. (1998). The Federal Reserve as an informed foreign exchange trader. *Federal Reserve Bank of Cleveland*, Working Paper 9815, September.

Continuous Probability Distributions

MARKUS HÖCHSTÖTTER, PhD

Assistant Professor, University of Karlsruhe

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair-Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Continuous probability distributions are needed when the random variable of interest can assume any value inside of one or more intervals of real numbers such as, for example, any number greater than zero. Asset returns, for example, whether measured monthly, weekly, daily, or at an even higher frequency are commonly modeled as continuous random variables. In contrast to discrete probability distributions that assign positive probability to certain discrete values, continuous probability distributions assign zero probability to any single real number. Instead, only entire intervals of real numbers can have positive probability such as, for example, the event that some asset return is not negative. For each continuous probability distribution, this necessitates the so-called probability density, a function that determines how the entire probability mass of one is distributed. The density often serves as the proxy for the respective probability distribution.

In this entry, we introduce the concept of continuous probability distributions. We present the continuous distribution function with its corresponding density function, a function unique to continuous probability laws. In this entry, parameters of location and scale such as the mean and higher moments—variance and skewness—are defined. For a more technical discussion of continuous distributions, see Evans, Hastings, and Peacock (2000) or Johnson, Kotz, and Balakrishnan (1995).

CONTINUOUS PROBABILITY DISTRIBUTION DESCRIBED

Suppose we are interested in outcomes that are no longer countable. Examples of such outcomes in finance are daily logarithmic stock returns, bond yields, and exchange rates. Technically, without limitations caused by rounding to a certain number of digits, we could imagine that any real number could provide a feasible outcome for the daily logarithmic return of some stock. That is, the set of feasible values

that the outcomes are drawn from (i.e., the space Ω) is uncountable. The events are described by continuous intervals such as, for example, $(-0.05, 0.05]$, which, referring to our example with daily logarithmic returns, would represent the event that the return at a given observation is more than -5% and at most 5% .

In the context of continuous probability distributions, we have the real numbers \mathbb{R} as the uncountable space Ω . The set of events is given by the Borel σ -algebra \mathbb{B} , which is based on the half-open intervals of the form $(-\infty, a]$, for any real a . The space \mathbb{R} and the σ -algebra \mathbb{B} form the measurable space (\mathbb{R}, \mathbb{B}) , which we are to deal with throughout this entry.

DISTRIBUTION FUNCTION

To be able to assign a probability to an event in a unique way, in the context of continuous distributions we introduce as a device the *continuous distribution function* $F(a)$, which expresses the probability that some event of the sort $(-\infty, a]$ occurs (i.e., that a number is realized that is at most a). (Formally, an outcome $\omega \in \Omega$ is realized that lies inside of the interval $(-\infty, a]$.) As with discrete random variables, this function is also referred to as the *cumulative distribution function* (*cdf*) since it aggregates the probability up to a certain value.

To relate to our previous example of daily logarithmic returns, the distribution function evaluated at say 0.05 , that is, $F(0.05)$, states the probability of some return of at most 5% . (The distribution function F is also referred to as the *cumulative probability distribution function* (often abbreviated *cdf*) expressing that the probability is given for the accumulation of all outcomes less than or equal to a certain value.)

For values x approaching $-\infty$, F tends to zero, while for values x approaching ∞ , F goes to 1. In between, F is monotonically increasing and right-continuous. More concisely, we list these

properties below:

$$\text{Property 1. } F(x) \xrightarrow{x \rightarrow -\infty} 0$$

$$\text{Property 2. } F(x) \xrightarrow{x \rightarrow \infty} 1$$

$$\text{Property 3. } F(b) - F(a) \geq 0 \text{ for } b \geq a$$

$$\text{Property 4. } \lim_{x \downarrow a} F(x) = F(a)$$

The behavior in the extremes—that is when x goes to either $-\infty$ or ∞ —is provided by properties 1 and 2, respectively. Property 3 states that F should be monotonically increasing (i.e., never become less for increasing values). Finally, property 4 guarantees that F is right-continuous.

Let us consider in detail the case when $F(x)$ is a continuous distribution, that is, the distribution has no jumps. The continuous probability distribution function F is associated with the probability measure P through the relationship

$$F(a) = P((-\infty, a])$$

that is, that values up to a occur, and

$$F(b) - F(a) = P((a, b]) \quad (1)$$

Therefore, from equation (1) we can see that the probability of some event related to an interval is given by the difference between the value of F at the upper bound b of the interval minus the value of F at the lower bound a . That is, the entire probability that an outcome of at most a occurs is subtracted from the greater event that an outcome of at most b occurs. Using set operations, we can express this as

$$(a, b] = (-\infty, b] \setminus (-\infty, a]$$

For example as we have seen, the event of a daily return of more than -5% and, at most, 5% is given by $(-0.05, 0.05]$. So, the probability associated with this event is given by $P((-0.05, 0.05]) = F(0.05) - F(-0.05)$.

In contrast to a discrete probability distribution, a continuous probability distribution always assigns zero probability to countable events such as individual outcomes a_i or unions

thereof such as

$$\bigcup_{i=1}^{\infty} a_i$$

That is,

$$\begin{aligned} P(\{a_i\}) &= 0, \text{ for all } a_i \\ P\left(\bigcup_{i=1}^{\infty} a_i\right) &= 0 \end{aligned} \quad (2)$$

From equation (2), we can apply the left-hand side of equation (1) also to events of the form (a, b) to obtain

$$P((a, b)) = F(b) - F(a) \quad (3)$$

Thus, it is irrelevant whether we state the probability of the daily logarithmic return being more than -5% and at most 5% , or the probability of the logarithmic return being more than -5% and less than 5% . They are the same because the probability of achieving a return of exactly 5% is zero. With a space Ω consisting of uncountably many possible values such as the set of real numbers, for example, each individual outcome is unlikely to occur. So, from a probabilistic point of view, one should never bet on an exact return or, associated with it, one particular stock price.

Since countable sets produce zero probability from a continuous probability measure, they belong to the so-called *P-null sets*. All events associated with *P-null sets* are unlikely events.

So, how do we assign probabilities to events in a continuous environment? The answer is given by equation (3). That, however, presumes knowledge of the distribution function F . The next task is to define the continuous distribution function F more specifically as explained next.

DENSITY FUNCTION

The continuous distribution function F of a probability measure P on (\mathbb{R}, \mathbb{B}) is defined as follows

$$F(x) = \int_{-\infty}^x f(t) dt \quad (4)$$

where $f(t)$ is the *density function* of the probability measure P .

We interpret equation (4) as follows. Since, at any real value x the distribution function uniquely equals the probability that an outcome of at most x is realized, that is, $F(x) = P((-\infty, x])$, equation (4) states that this probability is obtained by integrating some function f over the interval from $-\infty$ up to the value x .

What is the interpretation of this function f ? The function f is the marginal rate of growth of the distribution function F at some point x . We know that with continuous distribution functions, the probability of exactly a value of x occurring is zero. However, the probability of observing a value inside of the interval between x and some very small step to the right Δx (i.e., $[x, x + \Delta x)$) is not necessarily zero. Between x and $x + \Delta x$, the distribution function F increases by exactly this probability; that is, the increment is

$$F(x + \Delta x) - F(x) = P(X \in [x, x + \Delta x)) \quad (5)$$

Now, if we divide $F(x + \Delta x) - F(x)$ from equation (5) by the width of the interval Δx , we obtain the average probability or average increment of F per unit step on this interval. If we reduce the step size Δx to an infinitesimally small step ∂x , this average approaches the *marginal rate of growth* of F at x , which we denote f ; that is,¹

$$\lim_{\Delta x \rightarrow 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = \frac{\partial F(x)}{\partial x} \equiv f(x) \quad (6)$$

At this point, let us recall the histogram with relative frequency density for class data. Over each class, the height of the histogram is given by the density of the class divided by the width of the corresponding class. Equation (6) is somewhat similar if we think of it this way. We divide the probability that some realization should be inside of the small interval. And, by letting the interval shrink to width zero, we obtain the marginal rate of growth or, equivalently, the

derivative of F . (We assume that F is continuous and that the derivative of F exists.) Hence, we call the *probability density function* or simply the *density function*. Commonly, it is abbreviated as *pdf*.

Now, when we refocus on equation (4), we see that the probability of some occurrence of at most x is given by integration of the density function f over the interval $(-\infty, x]$. Again, there is an analogy to the histogram. The relative frequency of some class is given by the density multiplied by the corresponding class width. With continuous probability distributions, at each value t , we multiply the corresponding density $f(t)$ by the infinitesimally small interval width dt . Finally, we integrate all values of f (weighted by dt) up to x to obtain the probability for $(-\infty, x]$. This, again, is similar to histograms: In order to obtain the cumulative relative frequency at some value x , we compute the area covered by the histogram up to value x .

In Figure 1, we compare the histogram and the probability density function. The histogram

with density h is indicated by the dotted lines while the density function f is given by the solid line. We can now see how the probability $P((-\infty, x^*])$ is derived through integrating the marginal rate f over the interval $(-\infty, x^*]$ with respect to the values t . The resulting total probability is then given by the area A_1 of the example in Figure 1. This is analogous to class data where we would tally the areas of the rectangles whose upper bounds are less than x^* and the part of the area of the rectangle containing x^* up to the dash-dotted vertical line.

Requirements on the Density Function

Given the uncountable space \mathbb{R} (i.e., the real numbers) and the corresponding set of events given by the Borel σ -algebra \mathbb{B} , we can give a more rigorous formal definition of the density function. The *density function* f of probability measure P on the measurable space (\mathbb{R}, \mathbb{B}) with distribution function F is a Borel-measurable

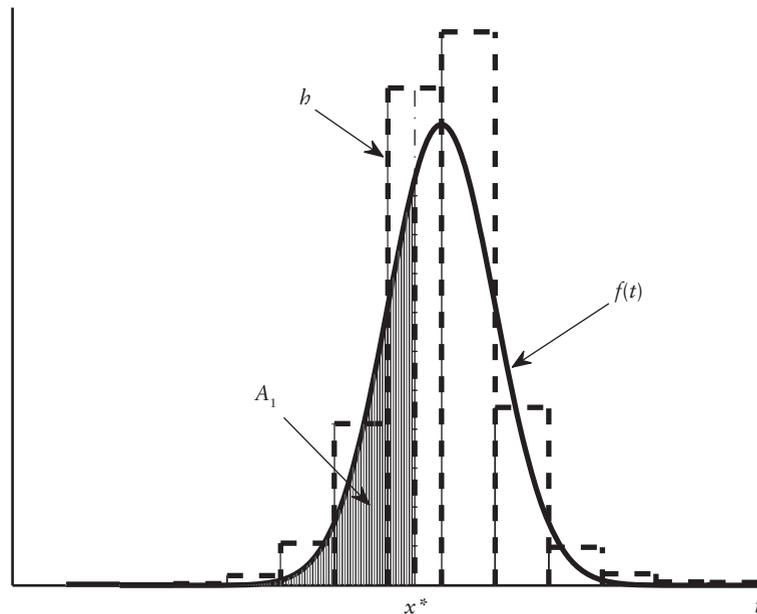


Figure 1 Comparison of Histogram and Density Function

Note: Area A_1 represents probability $P((-\infty, x^*])$ derived through integration of $f(t)$ with respect to t between $-\infty$ and x^*

function f satisfying.

$$P((-\infty, x]) = F(x) = \int_{-\infty}^x f(t)dt \quad (7)$$

with $f(t) \geq 0$, for all $t \in \mathbb{R}$ and

$$\int_{-\infty}^{\infty} f(t)dt = 1$$

By the requirement of Borel-measurability, we simply assume that the real-valued images generated by f have their origins in the Borel σ -algebra \mathbb{B} . Informally, for any value $y = f(t)$, we can trace the corresponding origin(s) t in \mathbb{B} that is (are) mapped to y through the function f . Otherwise, we might incur problems computing the *integral* in equation (7) for reasons that are beyond the scope of this entry.

From definition of the density function given by equation (7), we see that it is reasonable that f be a function that exclusively assumes nonnegative values. Although we have not mentioned this so far, it is immediately intuitive since f is the marginal rate of growth of the continuous distribution function F . At each $t, f(t) \cdot dt$ represents the limit probability that a value inside of the interval $(t, t + dt]$ should occur, which can never be negative. Moreover, we require the in-

tegration of f over the entire domain from $-\infty$ to ∞ to yield 1, which is intuitively reasonable since this integral gives the probability that any real value occurs.

The requirement

$$\int_{-\infty}^{\infty} f(t)dt = 1$$

implies the graphical interpretation that the area enclosed between the graph of f over the entire interval $(-\infty, \infty)$ and the horizontal axis equals one. This is displayed in Figure 2 by the shaded area A . For example, to visualize graphically what is meant by

$$\int_{-\infty}^x f(t)dt$$

in equation (7), we can use Figure 1. Suppose the value x were located at the intersection of the vertical dash-dotted line and the horizontal axis (i.e., x^*). Then, the shaded area A_1 represents the value of the integral and, therefore, the probability of occurrence of a value of at most x . To interpret

$$\int_a^b f(t)dt$$

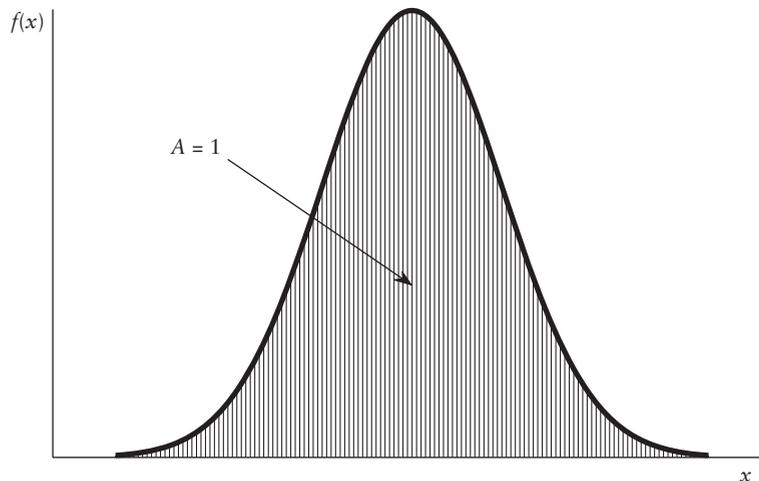


Figure 2 Graphical Interpretation of the Equality $A = \int_{-\infty}^{\infty} f(x)dx = 1$

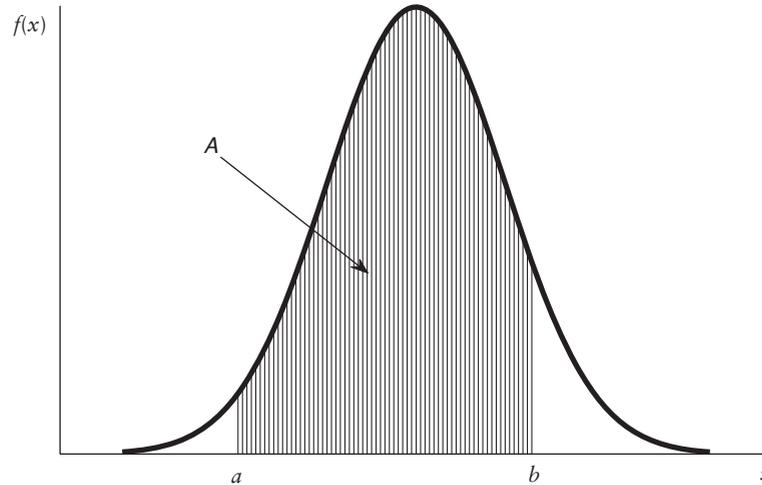


Figure 3 Graphical Interpretation of $A = \int_a^b f(x)dx$

graphically, look at Figure 3. The area representing the value of the interval is indicated by A . So, the probability of some occurrence of at least a and at most b is given by A . Here again, the resemblance to the histogram becomes obvious in that we divide one area above some class, for example, by the total area, and this ratio equates the according relative frequency.

For the sake of completeness, it should be mentioned without indulging in the reasoning behind it that there are probability measures P on (\mathbb{R}, \mathbb{B}) even with continuous distribution functions that do not have density functions as defined in equation (7). But, in our context, we will only regard probability measures with continuous distribution functions with associated density functions so that the equalities of equation (7) are fulfilled.

Sometimes, alternative representations equivalent to equation (7) are used. Typically, the following expressions are used

$$F(x) = \int_{\mathbb{R}} f(t) \cdot 1_{(-\infty, x]} dt \quad (8a)$$

$$F(x) = \int_{-\infty}^{\infty} f(t) \cdot 1_{(-\infty, x]} dt \quad (8b)$$

$$F(x) = \int_{-\infty}^{\infty} P(dt) \quad (8c)$$

$$F(x) = \int_{-\infty}^{\infty} dP(t) \quad (8d)$$

Note that in the first two equalities, (8a) and (8b), the indicator function $1_{(a,b]}$ is used. The last two equalities, (8c) and (8d), can be used even if there is no density function and, therefore, are of a more general form. We will, however, predominantly apply the representation given by equation (7) and occasionally resort to the last two forms above.

We introduce the term *support* at this point to refer to the part of the real line where the density is truly positive, that is, all those x where $f(x) > 0$.

CONTINUOUS RANDOM VARIABLE

So far, we have only considered continuous probability distributions and densities. We yet have to introduce the quantity of greatest interest to us in this entry, the *continuous random variable*. For example, stock returns, bond yields,

and exchange rates are usually modeled as continuous random variables.

Informally stated, a continuous random variable assumes certain values governed by a probability law uniquely linked to a continuous distribution function F . Consequently, it has a density function associated with its distribution. Often, the random variable is merely described by its density function rather than the probability law or the distribution function.

By convention, let us indicate the random variables by capital letters. Recall that any random variable, and in particular a continuous random variable X , is a *measurable function*. Let us assume that X is a function from the probability space $\Omega = \mathbb{R}$ into the *state space* $\Omega' = \mathbb{R}$. That is, origin and image space coincide. The corresponding σ -algebrae containing events of the elementary outcomes ω and the events in the image space $X(\omega)$, respectively, are both given by the Borel σ -algebra \mathbb{B} . Now, we can be more specific by requiring the continuous random variable X to be a $\mathbb{B} - \mathbb{B}$ -measurable real-valued function. That implies, for example, that any event $X \in (a, b]$, which is in \mathbb{B} , has its origin $X^{-1}((a, b])$ in \mathbb{B} , as well. Measurability is important when we want to derive the probability of events in the state space such as $X \in (a, b]$ from original events in the probability space such as $X^{-1}((a, b])$. At this point, one should not be concerned that the theory is somewhat overwhelming. It will become easier to understand once we move to the examples.

COMPUTING PROBABILITIES FROM THE DENSITY FUNCTION

The relationship between the continuous random variable X and its density is given by the following.² Suppose X has density f , then the probability of some event $X \leq x$ or $X \in (a, b]$ is

computed as

$$\begin{aligned} P(X \leq x) &= \int_{-\infty}^x f(t)dt \\ P(X \in (a, b]) &= \int_a^b f(t)dt \end{aligned} \quad (9)$$

which is equivalent to $F(x)$ and $F(b) - F(a)$ respectively, because of the one-to-one relationship between the density f and the distribution function F of X .

As explained earlier, using indicator functions, equation (9) could be alternatively written as

$$\begin{aligned} P(X \leq x) &= \int_{-\infty}^{\infty} 1_{(-\infty, x]}(t) f(t)dt \\ P(X \in (a, b]) &= \int_{-\infty}^{\infty} 1_{(a, b]}(t) f(t)dt \end{aligned}$$

In the following, we will introduce parameters of location and spread such as the mean and the variance, for example. In contrast to the data-dependent statistics, parameters of random variables never change. Some probability distributions can be sufficiently described by their parameters. They are referred to as *parametric distributions*. For example, for the normal distribution we introduce shortly, it is sufficient to know the parameters mean and variance to completely determine the corresponding distribution function. That is, the shape of parametric distributions is governed only by the respective parameters.

LOCATION PARAMETERS

The most important location parameter is the *mean* that is also referred to as the *first moment*. It is the only location parameter presented in this entry.

The mean can be thought of as an average value. It is the number that one would have to

expect for some random variable X with given density function f . The mean is defined as follows: Let X be a real-valued random variable on the space $\Omega = \mathbb{R}$ with Borel σ -algebra \mathbb{B} . The *mean* is given by

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx \quad (10)$$

in case the integral on the right-hand side of equation (10) exists (i.e., is finite). Typically, the mean parameter is denoted as μ .

In equation (10) that defines the mean, we weight each possible value x that the random variable X might assume by the product of the density at this value, $f(x)$, and step size dx . Recall that the product $f(x) \cdot dx$ can be thought of as the limiting probability of attaining the value x . Finally, the mean is given as the integral over these weighted values. Thus, equation (10) is similarly understood as the definition of the mean of a discrete random variable where, instead of integrated, the probability-weighted values are summed.

DISPERSION PARAMETERS

We turn our focus toward measures of spread or, in other words, dispersion measures. Again, as with the previously introduced measures of location, in probability theory the dispersion measures are universally given parameters. Here, we introduce the moments of higher order, variance, standard deviation, and the skewness parameters.

Moments of Higher Order

It might sometimes be necessary to compute *moments of higher order*. As we already know from descriptive statistics, the mean is the moment of order one. (Alternatively, we often say the *first* moment. For the higher orders k , we consequently might refer to the k -th moment.) However, one might not be interested in the expected value of some quantity itself but of its square. If we treat this quantity as a continu-

ous random variable, we compute what is the *second moment*.

Let X be a real-valued random variable on the space $\Omega = \mathbb{R}$ with Borel σ -algebra \mathbb{B} . The *moment of order k* is given by the expression

$$E(X^k) = \int_{-\infty}^{\infty} x^k \cdot f(x) dx \quad (11)$$

in case the integral on the right-hand side of equation (11) exists (i.e., is finite).

From equation (11), we learn that higher-order moments are equivalent to simply computing the mean of X taken to the k -th power.

Variance

The variance involves computing the expected squared deviation from the mean $E(X) = \mu$ of some random variable X . For a continuous random variable X , the variance is defined as follows: Let X be a real-valued random variable on the space $\Omega = \mathbb{R}$ with Borel σ -algebra \mathbb{B} , then the *variance* is

$$\begin{aligned} \text{Var}(X) &= \int_{-\infty}^{\infty} (x - E(X))^2 \cdot f(x) dx \\ &= \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x) dx \quad (12) \end{aligned}$$

in case the integral on the right-hand side of equation (12) exists (i.e., is finite). Often, the variance in equation (12) is denoted by the symbol σ^2 .

In equation (12), at each value x , we square the deviation from the mean and weight it by the density at x times the step size dx . The latter product, again, can be viewed as the limiting probability of the random variable X assuming the value x . The square inflates large deviations even more compared to smaller ones. For some random variable to have a small variance, it is essential to have a quickly vanishing density in the parts where the deviations $(x - \mu)$ become large.

All distributions that we discuss in this entry are parametric distributions. For some of

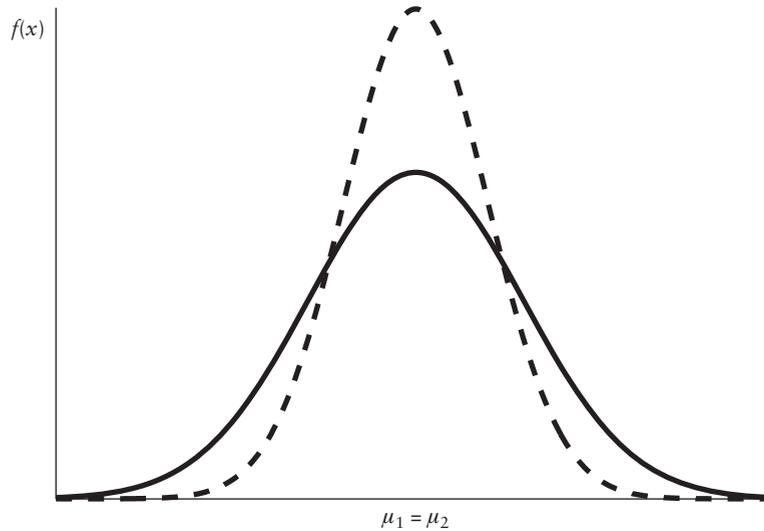


Figure 4 Two Density Functions Yielding Common Means, $\mu_1 = \mu_2$, but Different Variances, $\sigma_1^2 < \sigma_2^2$
 Note: Dashed graph: $\sigma_1^2 = 1$. Solid graph: $\sigma_2^2 = 1.5$.

them, it is enough to know the mean μ and variance σ^2 and consequently, we will resort to these two parameters often. Historically, the variance has often been given the role of risk measure in context of portfolio theory. Suppose we have two random variables R_1 and R_2 representing the returns of two stocks, S_1 and S_2 , with equal means μ_{R_1} and μ_{R_2} , respectively, so that $\mu_{R_1} = \mu_{R_2}$. Moreover, let R_1 and R_2 have variances $\sigma_{R_1}^2$ and $\sigma_{R_2}^2$, respectively, with $\sigma_{R_1}^2 < \sigma_{R_2}^2$. Then, omitting further theory, at this moment, we prefer S_1 to S_2 because of the S_1 's smaller variance. We demonstrate this in Figure 4. The dashed line represents the graph of the first density function while the second one is depicted by the solid line. Both density functions yield the same mean (i.e., $\mu_1 = \mu_2$). However, the variance from the first density function, given by the dashed graph, is smaller than that of the solid graph (i.e., $\sigma_1^2 < \sigma_2^2$). Thus, using variance as the risk measure and resorting to density functions that can be sufficiently described by the mean and variance, we can state that density function for S_1 (dashed graph) is preferable. We can interpret the figure as follows.

Since the variance of the distribution with the dashed density graph is smaller, the probability mass is less dispersed over all x values. Hence, the density is more condensed about the center and more quickly vanishing in the extreme left and right ends, the so-called *tails*. On the other hand, the second distribution with the solid density graph has a larger variance, which can be verified by the overall flatter and more expanded density function. About the center, it is lower and less compressed than the dashed density graph, implying that the second distribution assigns less probability to events immediately near the center. However, the density function of the second distribution decays more slowly in the tails than the first, which means that under the governance of the latter, extreme events are less likely than under the second probability law.

Standard Deviation

The parameter related to the variance is the *standard deviation*. As we know from descriptive statistics described earlier in this book, the standard deviation is the positive square root

of the variance. That is, let X be a real-valued random variable on the space $\Omega = \mathbb{R}$ with Borel σ -algebra \mathbb{B} . Furthermore, let its mean and variance be given by μ and σ^2 , respectively. The standard deviation is defined as

$$\sigma = \sqrt{\text{Var}(X)}$$

For example, in the context of stock returns, one often expresses using the standard deviation the return's fluctuation around its mean. The standard deviation is often more appealing than the variance since the latter uses squares, which are a different scale from the original values of X . Even though mathematically not quite correct, the standard deviation, denoted by σ , is commonly interpreted as the average deviation from the mean.

Skewness

Consider the density function portrayed in Figure 5. The figure is obviously symmetric about some location parameter μ in the sense that $f(-x - \mu) = f(x - \mu)$. Suppose instead that we encounter a density function f of some random variable X that is depicted in Figure 6. This figure is not symmetric about any location parameter. Consequently, some quantity stating the

extent to which the density function is deviating from symmetry is needed. This is accomplished by a parameter referred to as *skewness*. This parameter measures the degree to which the density function leans to either side, if at all.

Let X be a real-valued random variable on the space $\Omega = \mathbb{R}$ with Borel σ -algebra \mathbb{B} , variance σ^2 , and mean $\mu = E(X)$. The skewness parameter, denoted by γ , is given by

$$\gamma = \frac{E\left((x - E(X))^3\right)}{\sigma^{3/2}}$$

The skewness measure given above is referred to as the *Pearson skewness* measure. Negative values indicate skewness to the left (i.e., *left skewed*) while skewness to the right is given by positive values (i.e., *right skewed*).

The design of the skewness parameter follows the following reasoning. In the numerator, we measure the distance from every value x to the mean $E(X)$ of random variable X . To overweight larger deviations, we take them to a higher power than one. In contrast to the variance where we use squares, in the case of skewness we take the third power since three is an odd number and thereby preserves both the signs and directions of the deviations. Due

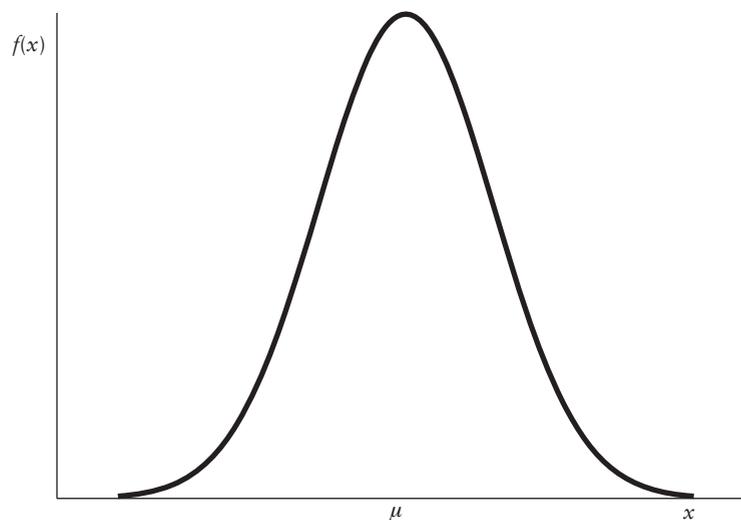


Figure 5 Example of Some Symmetric Density Function $f(x)$

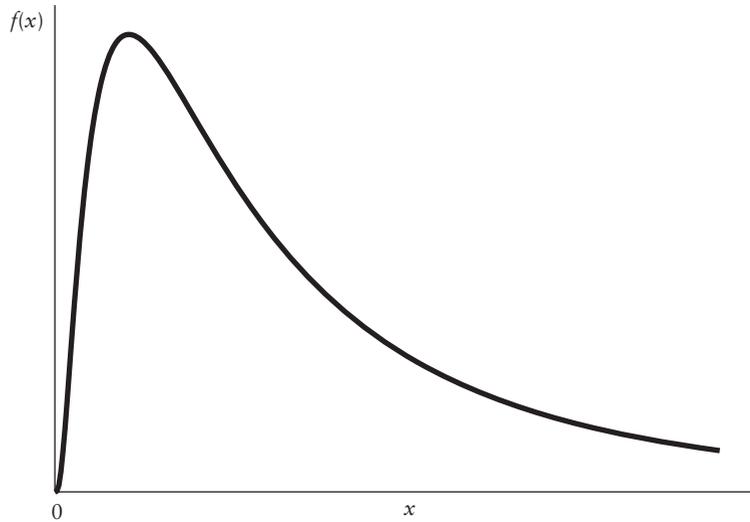


Figure 6 Example of Some Asymmetric Density Function $f(x)$

to this sign preservation, symmetric density functions yield zero skewness since all deviations to the left of the mean cancel their counterparts to the right. To standardize the deviations, we scale them by dividing by the standard deviation, also taken to the third power. So, the skewness parameter is not influenced by the standard deviation of the distributions. If we did not scale the skewness parameter in this way, distribution functions with density functions having large variances would always produce larger skewness even though the density is not really tilted more pronouncedly than some similar density with smaller variance.

We graphically illustrate the skewness parameter γ in Figure 6 for some density function $f(x)$. A density function f that assumes positive values $f(x)$ only for positive real values (i.e., $x > 0$) but zero for $x \leq 0$ is shown in the figure. The random variable X with density function f has mean $\mu = 1.65$. Its standard deviation is computed as $\sigma = 0.957$. The value of the skewness parameter is $\gamma = 0.7224$, indicating a positive skewness. The sign of the skewness parameter can be easily verified by analyzing the density graph. The density peaks just a little to the

right of the leftmost value $x = 0$. Toward the left tail, the density decays abruptly and vanishes at zero. Toward the right tail, things look very different in that f decays very slowly, approaching a level of $f = 0$ as x goes to positive infinity. (The graph is depicted for $x \in [0, 3.3]$.)

KEY POINTS

- A continuous random variable is a random variable that does not only assume values from a set of discrete values but may assume any real value from within one or more intervals. Often, asset returns are modeled as continuous random variables.
- The continuous distribution function is the probability distribution associated with a continuous random variable. It distinguishes itself from the discrete probability distribution in that it gives positive probability only to entire intervals rather than some discrete values only.
- To appreciate continuous random variables, it is necessary to understand the concept of the derivative of some function, which is the marginal rate of growth of some function at a certain point. It can be conceived as the slope

of the function at some point considering only very small increments in the argument of the function away from the point.

- The density function determines how the probability mass of one is distributed across the real line. Hence, it would be counterintuitive if that function were ever negative such that we require it to be nonnegative. Technically, it is the marginal rate of growth of the distribution function at any position or, in other words, its derivative.
- As support of some probability distribution, we define the subset of the real numbers that represents 100% of the probability. For the continuous probability distributions, it is the collection of intervals where the associated probability density is positive.

NOTES

1. The expression $\partial F(x)$ is equivalent to the increment $F(x + \Delta x) - F(x)$ as Δx goes to zero.
2. Sometimes the density of X is explicitly indexed f_x . We will not do so here, however, except where we believe not doing so will lead to confusion. The same holds for its distribution function F .

REFERENCES

- Evans, M., Hastings, N., and Peacock, B. (2000). *Statistical Distributions*, 3rd ed. New York: Wiley.
- Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, vol. 2, 2nd ed. New York: Wiley.

Continuous Probability Distributions with Appealing Statistical Properties

MARKUS HÖCHSTÖTTER, PhD

Assistant Professor, University of Karlsruhe

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair-Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: To model the behavior of certain financial assets in a stochastic environment, we can usually resort to a variety of theoretical distributions. Most commonly, probability distributions are selected that are analytically well known. For example, the normal distribution is often the distribution of choice when asset returns are modeled, or the exponential distribution is applied to characterize the randomness of the time between two successive defaults of firms in a bond portfolio. Many other distributions are related to them or built on them in a well-known manner. These distributions often display pleasant features such as stability under summation—meaning that the return of a portfolio of assets whose returns follow a certain distribution again follows the same distribution. However, one has to be careful using these distributions since their advantage of mathematical tractability is often outweighed by the fact that the stochastic behavior of the true asset returns is not well captured by these distributions.

In this entry, we discuss the more commonly used distributions with appealing statistical properties that are used in finance. The distributions discussed are the normal distribution, the chi-square distribution, the Student's t -distribution, the Fisher's F -distribution, the exponential distribution, the gamma distribution (including the special Erlang distribution), the beta distribution, and the log-normal

distribution. Many of the distributions enjoy widespread attention in finance, or statistical applications in general, due to their well-known characteristics or mathematical simplicity. However, as we emphasize, the use of some of them might be ill-suited to replicate the real-world behavior of financial returns. For a more technical discussion of continuous distributions, see Evans, Hastings,

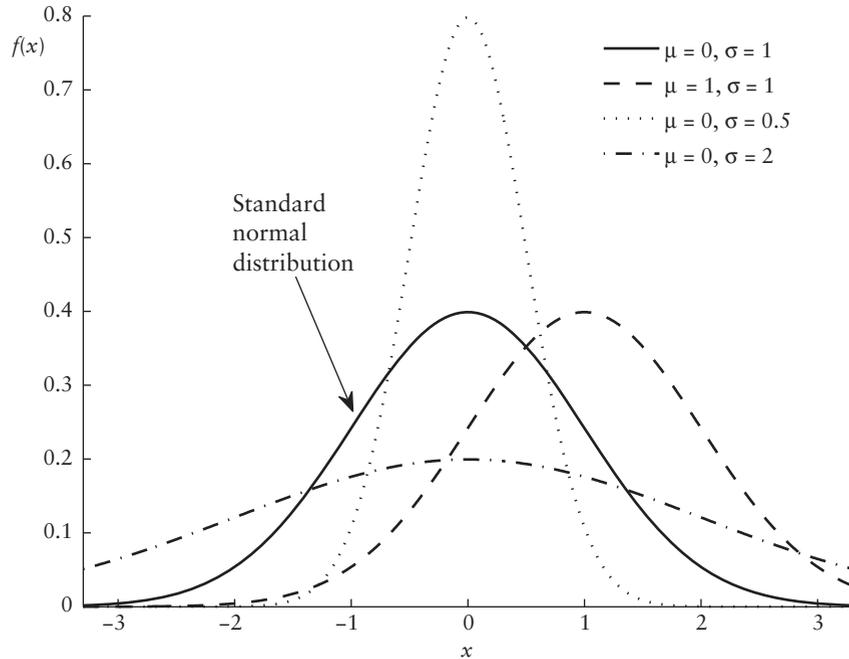


Figure 1 Normal Density Function for Various Parameter Values

and Peacock (2000) or Johnson, Kotz, and Balakrishnan (1995).

NORMAL DISTRIBUTION

The first distribution we discuss is the *normal distribution*. It is the distribution most commonly used in finance despite its many limitations. This distribution, also referred to as the *Gaussian distribution* (named after the mathematician and physicist C. F. Gauss), is characterized by the two parameters: mean (μ) and standard deviation (σ). The distribution is denoted by $N(\mu, \sigma^2)$. When $\mu = 0$ and $\sigma^2 = 1$, then we obtain the *standard normal distribution*.

For $x \in R$, the density function for the normal distribution is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

The density in equation (1) is always positive. Hence, we have support (i.e., positive density) on the entire real line. Furthermore, the density function is symmetric about μ . A plot of the density function for several parameter values is given in Figure 1. As can be seen, the value of μ results in a horizontal shift from 0 while σ inflates or deflates the graph. A characteristic of the normal distribution is that the densities are bell shaped.

A problem is that the distribution function cannot be solved for analytically and therefore has to be approximated numerically. In the particular case of the standard normal distribution, the values are tabulated. Standard statistical software provides the values for the standard normal distribution as well as most of the distributions presented in this entry. The standard normal distribution is commonly denoted by the Greek letter Φ such that we have $\Phi(x) = F(x) = P(X \leq x)$, for some standard normal random variable X . In Figure 2, graphs

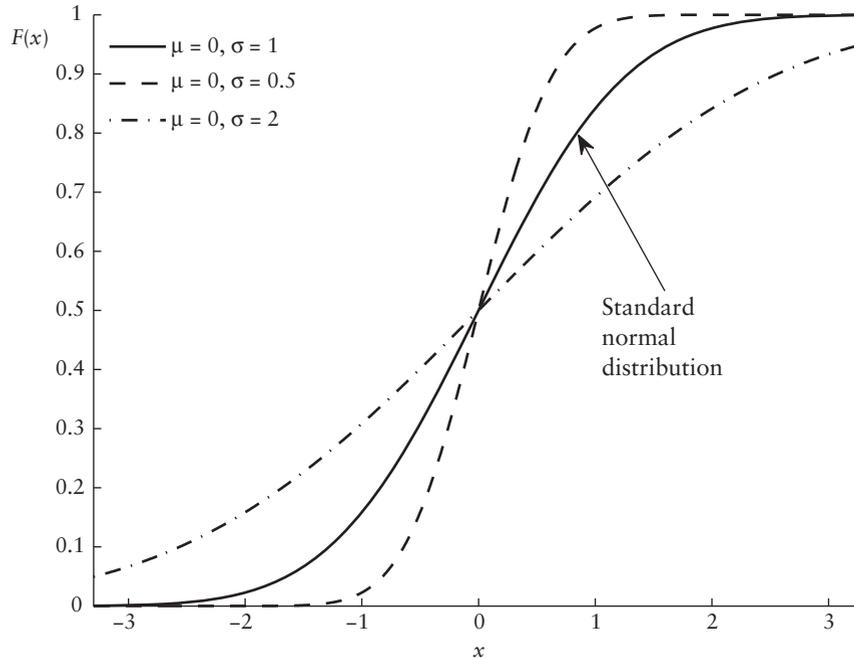


Figure 2 Normal Distribution Function for Various Parameter Values

of the distribution function are given for three different sets of parameters.

Properties of the Normal Distribution

The normal distribution provides one of the most important classes of probability distributions due to two appealing properties that generally are not shared by all distributions:

Property 1. The distribution is *location-scale invariant*.

Property 2. The distribution is *stable under summation*.

Property 1, the location-scale invariance property, guarantees that we may multiply X by b and add a where a and b are any real numbers. Then, the resulting $a + b \cdot X$ is, again, normally distributed, more precisely, $N(a + \mu, b\sigma)$. Consequently, a normal random variable will still be normally distributed if we change the units of measurement. The change into $a + b \cdot X$ can be interpreted as observing the same X , however,

measured in a different scale. In particular, if a and b are such that the mean and variance of the resulting $a + b \cdot X$ are 0 and 1, respectively, then $a + b \cdot X$ is called the *standardization* of X .

Property 2, stability under summation, ensures that the sum of an arbitrary number n of normal random variables, X_1, X_2, \dots, X_n is, again, normally distributed provided that the random variables behave independently of each other. This is important for aggregating quantities.

These properties are illustrated later in the entry.

Furthermore, the normal distribution is often mentioned in the context of the *central limit theorem*. It states that a sum of random variables with identical distributions and being independent of each other results in a normal random variable.¹ We restate this formally as follows:

Let X_1, X_2, \dots, X_n be identically distributed random variables with mean $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ and do not influence the outcome of each other (i.e., are independent).

Then, we have

$$\frac{\sum_{i=1}^n X_i - n \cdot \mu}{\sigma \sqrt{n}} \xrightarrow{D} N(0, 1) \quad (2)$$

as the number n approaches infinity. The D above the convergence arrow in equation (2) indicates that the distribution function of the left expression converges to the standard normal distribution.

Generally, for $n = 30$ in equation (2), we consider equality of the distributions; that is, the left-hand side is $N(0,1)$ distributed. In certain cases, depending on the distribution of the X_i and the corresponding parameter values, $n < 30$ justifies the use of the standard normal distribution for the left-hand side of equation (2). If the X_i are Bernoulli random variables, that is, $X_i \sim B(p)$, with parameter p such that $n \cdot p \geq 5$, then we also assume equality in the distributions in equation (2). Depending on p , this can mean that n is much smaller than 30.

These properties make the normal distribution the most popular distribution in finance. But this popularity is somewhat contentious, however, for reasons that will be given as we progress in this entry.

The last property we will discuss of the normal distribution that is shared with some other distributions is the bell shape of the density function. This particular shape helps in roughly assessing the dispersion of the distribution due to a rule of thumb commonly referred to as the *empirical rule*. Due to this rule, we have

$$\begin{aligned} P(X \in [\mu \pm \sigma]) &= F(\mu + \sigma) - F(\mu - \sigma) \approx 68\% \\ P(X \in [\mu \pm 2\sigma]) &= F(\mu + 2\sigma) - F(\mu - 2\sigma) \approx 95\% \\ P(X \in [\mu \pm 3\sigma]) &= F(\mu + 3\sigma) - F(\mu - 3\sigma) \approx 100\% \end{aligned}$$

The above states that approximately 68% of the probability is given to values that lie in an interval one standard deviation σ about the mean μ . About 95% probability is given to values within 2σ to the mean, while nearly all probability is assigned to values within 3σ from the mean.

By comparison, the so-called *Chebyshev inequalities* valid for any type of distribution—so not necessarily bell-shaped—yield

$$\begin{aligned} P(X \in [\mu \pm \sigma]) &\approx 0\% \\ P(X \in [\mu \pm 2\sigma]) &\approx 75\% \\ P(X \in [\mu \pm 3\sigma]) &\approx 89\% \end{aligned}$$

which provides a much coarser assessment than the empirical rule as we can see, for example, by the assessed 0% of data contained inside of one standard deviation about the mean.

Applications to Stock Returns

Applying Properties 1 and 2 to Stock Returns

With respect to Property 1, consider an example of normally distributed stock returns r with mean μ . If μ is nonzero, this means that the returns are a combination of a constant μ and random behavior centered about zero. If we were only interested in the latter, we would subtract μ from the returns and thereby obtain a new random variable $\tilde{r} = r - \mu$, which is again normally distributed.

With respect to Property 2, we give two examples. First, let us present the effect of aggregation over time. We consider daily stock returns that, by our assumption, follow a normal law. By adding the returns from each trading day during a particular week, we obtain the week's return as $r_w = r_{Mo} + r_{Tu} + \dots + r_{Fr}$ where r_{Mo} , r_{Tu} , \dots , r_{Fr} are the returns from Monday through Friday. The weekly return r_w is normally distributed as well. The second example applies to portfolio returns. Consider a portfolio consisting of n different stocks, each with normally distributed returns. We denote the corresponding returns by R_1 through R_n . Furthermore, in the portfolio we weight each stock i with w_i , for $i = 1, 2, \dots, n$. The resulting portfolio return $R_p = w_1R_1 + w_2R_2 + \dots + w_nR_n$ is also a normal random variable.

Using the Normal Distribution to Approximate the Binomial Distribution

Consider the binomial stock price model. At time $t = 0$, the stock price was $S_0 = \$20$. At time $t = 1$, the stock price was either up or down by 10% so that the resulting price was either $S_0 = \$18$ or $S_0 = \$22$. Both up- and down-movement occurred with probability $P(\$18) = P(\$22) = 0.5$. Now we extend the model to an arbitrary number of n days. Suppose each day i , $i = 1, 2, \dots, n$, the stock price developed in the same manner as on the first day. That is, the price is either up 10% with 50% probability or down 10% with the same probability. If on day i the price is up, we denote this by $X_i = 1$ and $X_i = 0$ if the price is down. The X_i are, hence, $B(0.5)$ random variables. After, say, 50 days, we have a total of $Y = X_1 + X_2 + \dots + X_{50}$ up-movements. Note that because of the assumed independence of the X_i , that Y is a $B(50, 0.5)$ random variable with mean $n \cdot p = 25$ and variance $n \cdot p \cdot (1 - p) = 12.5$. Let us introduce

$$Z_{50} = \frac{Y - 25}{\sqrt{12.5}}$$

From the comments regarding equation (2), we can assume that Z_{50} is approximately $N(25, 12.5)$ distributed. So, the probability of at most 15 up-movements, for example, is given by $P(Y \leq 15) = \Phi((15 - 25)/\sqrt{12.5}) = 0.23\%$. By comparison, the probability of no more than five up-movements is equal to $P(Y \leq 5) = \Phi((5 - 25)/\sqrt{12.5}) = 0\%$.

Normal Distribution for Logarithmic Returns

As another example, let X be some random variable representing a quantitative daily market dynamic such as new information about the economy. A dynamic can be understood as some driving force governing the development of other variables. We assume that it is normally distributed with mean $E(X) = \mu = 0$ and variance $Var(X) = \sigma^2 = 0.2$. Formally, we would write $X \sim N(0, 0.2)$. So, on average, the value of the daily dynamic will be zero with a standard deviation of $\sqrt{0.2}$. In addition, we introduce a

stock price S as a random variable, which is equal to S_0 at the beginning.

After one day, the stock price is modeled to depend on the dynamic X as follows

$$S_1 = S_0 \cdot e^X$$

where S_1 is the stock price after one day. The exponent X in this presentation is referred to as a *logarithmic return* in contrast to a *multiplicative return* R obtained from the formula $R = S_1/S_0 - 1$. So, for example, if $X = 0.01$, S_1 is equal to $e^{0.01} \cdot S_0$. That is almost equal to $1.01 \cdot S_0$, which corresponds to an increase of 1% relative to S_0 .² The probability of X being, for instance, no greater than 0.01 after one day is given by³

$$\begin{aligned} P(X \leq 0.01) &= \int_{-\infty}^{0.01} f(x) dx \\ &= \int_{-\infty}^{0.01} \frac{1}{\sqrt{2\pi} \sqrt{0.2}} e^{-\frac{x^2}{2 \cdot 0.2}} dx \approx 0.51 \end{aligned}$$

Consequently, after one day, the stock price increases, at most, by 1% with 51% probability, that is, $P(S_1 \leq 1.01 \cdot S_0) \approx 0.51$.

Next, suppose we are interested in a five-day outlook where the daily dynamics X_i , $i = 1, 2, \dots, 5$ of each of the following consecutive five days are distributed identically as X and independent of each other. Since the dynamic is modeled to equal exactly the continuously compounded return—that is logarithmic returns—we refer to X as the return in this entry. For the resulting five-day returns, we introduce the random variable $Y = X_1 + X_2 + \dots + X_5$ as the linear combination of the five individual daily returns. We know that Y is normally distributed from Property 2. More precisely, $Y \sim N(0, 1)$. So, on average, the return tends in neither direction, but the volatility measured by the standard deviation is now $\sqrt{5} \approx 2.24$ times that of the daily return X . Consequently, the probability of Y not exceeding a value of 0.01 is

now,

$$P(Y \leq 0.01) = \int_{-\infty}^{0.01} \frac{1}{\sqrt{2\pi}\sqrt{1}} e^{-\frac{y^2}{2 \cdot 1}} dy \approx 0.50$$

We see that the fivefold variance results in a greater likelihood to exceed the threshold 0.01, that is,

$$\begin{aligned} P(Y > 0.01) &= 1 - P(Y \leq 0.01) \\ &\approx 0.50 > 0.49 \approx P(X > 0.01) \end{aligned}$$

We model the stock price after five days as

$$S_5 = S_0 \cdot e^Y = S_0 \cdot e^{X_1 + X_2 + \dots + X_5}$$

So, after five days, the probability for the stock price to have increased by no more than 1% relative to S_0 is equal to

$$P(S_5 \leq e^{0.01} \cdot S_0) = P(S_5 \leq 1.01 \cdot S_0) \approx 0.50$$

There are two reasons why in finance logarithmic returns are commonly used. First, logarithmic returns are often easier to handle than multiplicative returns. Second, if we consider returns that are attributed to ever shorter periods of time (e.g., from yearly to monthly to weekly to daily and so on), the resulting compounded return after some fixed amount of time can be expressed as a logarithmic return. The theory behind this can be obtained from any introductory book on calculus.

CHI-SQUARE DISTRIBUTION

Our next distribution is the *chi-square distribution*. Let Z be a standard normal random variable, in brief $Z \sim N(0,1)$, and let $X = Z^2$. Then X is distributed chi-square with one degree of freedom. We denote this as $X \sim \chi^2(1)$. The *degrees of freedom* indicate how many independently behaving standard normal random variables the resulting variable is composed of. Here X is just composed of one, namely Z , and therefore has one degree of freedom.

Because Z is squared, the chi-square distributed random variable assumes only non-negative values; that is, the support is on the

nonnegative real numbers. It has mean $E(X) = 1$ and variance $Var(X) = 2$.

In general, the chi-square distribution is characterized by the degrees of freedom n , which assume the values $1, 2, \dots$. Let X_1, X_2, \dots, X_n be n $\chi^2(1)$ distributed random variables that are all independent of each other. Then their sum, S , is

$$S = \sum_{i=1}^n X_i \sim \chi^2(n) \quad (3)$$

In words, the sum is again distributed chi-square but this time with n degrees of freedom. The corresponding mean is $E(X) = n$, and the variance equals $Var(X) = 2 \cdot n$. So, the mean and variance are directly related to the degrees of freedom.

From the relationship in equation (3), we see that the degrees of freedom equal the number of independent $\chi^2(1)$ distributed X_i in the sum. If we have $X_1 \sim \chi^2(n_1)$ and $X_2 \sim \chi^2(n_2)$, it follows that

$$X_1 + X_2 \sim \chi^2(n_1 + n_2) \quad (4)$$

From property (4), we have that chi-square distributions have Property 2; that is, they are stable under summation in the sense that the sum of any two chi-squared distributed random variables is itself chi-square distributed.

The chi-square density function with n degrees of freedom is given by

$$f(x) = \begin{cases} f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} \cdot e^{-x/2} \cdot x^{n/2-1}, & x \geq 0 \\ 0 & x < 0 \end{cases}$$

for $n = 1, 2, \dots$ where $\Gamma(\cdot)$ is the *gamma function*. Figure 3 shows a few examples of the chi-square density function with varying degrees of freedom. As can be observed, the chi-square distribution is skewed to the right.

Application to Modeling Short-Term Interest Rates

As an example of an application of the chi-square distribution, we present a simplified

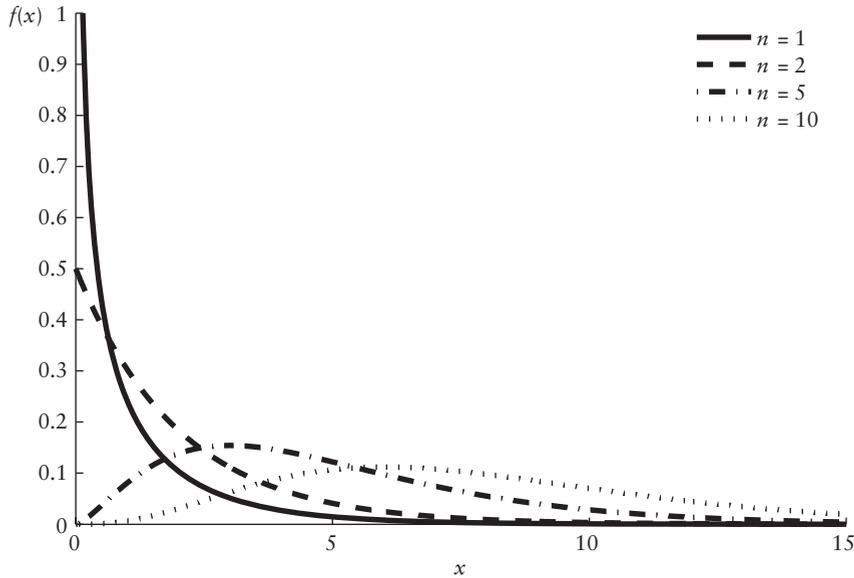


Figure 3 Density Functions of Chi-Square Distributions for Various Degrees of Freedom n

model of short-term interest rates, that is, so-called *short rates*. The short rate given by r_t , at any time t , is assumed to be a nonnegative continuous random variable. Furthermore, we let the short rate be composed of d independent dynamics X_1, X_2, \dots, X_d according to

$$r_t = X_1^2 + X_2^2 + \dots + X_d^2$$

where d is some positive integer number. In addition, each X_i is given as a standard normal random variable independent of all other dynamics. Then, the resulting short rate r_t is chi-square distributed with d degrees of freedom, that is, $r_t \sim \chi^2(d)$.

If we let $d = 2$ (i.e., there are two dynamics governing the short rate), the probability of a short rate between 0 and 1% is 0.5%. That is, we have to expect that on five out of 1,000 days, we will have a short rate assuming some value in the interval $(0, 0.01]$. If, in addition, we had one more dynamic included such that $r_t \sim \chi^2(3)$, then, the same interval would have probability $P(r_t \in (0, 0.01]) \approx 0.03\%$, which is close to being an unlikely event. We see that the more dynam-

ics are involved, the less probable small interest rates such as 1% or less become.

It should be realized, however, that this is merely an approach to model the short rate statistically and not an economic model explaining the factors driving the short rate.

STUDENT'S t -DISTRIBUTION

An important continuous probability distribution when the population variance of a distribution is unknown is the *Students t -distribution* (also referred to as the *t -distribution* and *Student's distribution*).

The t -distribution is a mixture of the normal and chi-square distributions. To derive the distribution, let X be distributed standard normal, that is, $X \sim N(0, 1)$, and S be chi-square distributed with n degrees of freedom, that is, $S \sim \chi^2(n)$. Furthermore, if X and Y are independent of each other (which is to be understood as not influencing the outcome of the other), then

$$Z = \frac{X}{\sqrt{S/n}} \sim t(n) \quad (5)$$

In words, equation (5) states that the resulting random variable Z is Student's t -distributed with n degrees of freedom. The degrees of freedom are inherited from the chi-square distribution of S .

How can we interpret equation (5)? Suppose we have a population of normally distributed values with zero mean. The corresponding normal random variable may be denoted as X . If one also knows the standard deviation of X ,

$$\sigma = \sqrt{\text{Var}(X)}$$

with X/σ , we obtain a standard normal random variable.

However, if σ is not known, we have to use, for example,

$$\sqrt{S/n} = \sqrt{1/n \cdot (X_1^2 + X_2^2 + \dots + X_n^2)}$$

instead where $X_1^2, X_2^2, \dots, X_n^2$ are n random variables identically distributed as X . Moreover, X_1, X_2, \dots, X_n have to assume values independently of each other. Then, the distribution of

$$X/\sqrt{S/n}$$

is the t -distribution with n degrees of freedom, that is,

$$X/\sqrt{S/n} \sim t(n)$$

By dividing by σ or S/n , we generate rescaled random variables that follow a standardized distribution. Quantities similar to $X/\sqrt{S/n}$ play an important role in parameter estimation.

The density function is defined as

$$f(x) = \frac{1}{\sqrt{n \cdot \pi}} \cdot \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \cdot \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad (6)$$

where the gamma function Γ is incorporated again. The density function is symmetric and has support (i.e., is positive) on all \mathbb{R} .

Basically, the Student's t -distribution has a similar shape to the normal distribution, but thicker tails. For large degrees of freedom n , the Student's t -distribution does not significantly differ from the standard normal distribution. As a matter of fact, for $n \geq 100$, it is practically indistinguishable from $N(0,1)$.

Figure 4 shows the Student's t -density function for various degrees of freedom plotted

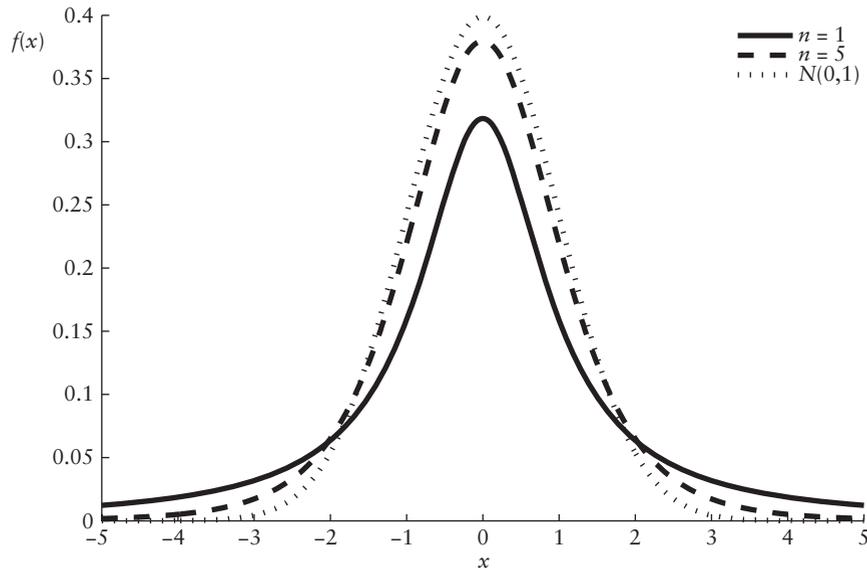


Figure 4 Density Function of the t -Distribution for Various Degrees of Freedom n Compared to the Standard Normal Density Function ($N(0,1)$)

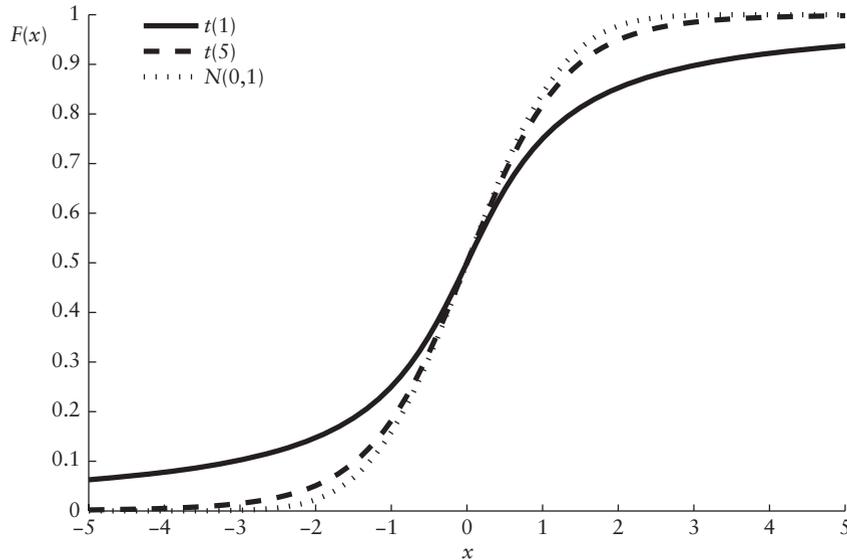


Figure 5 Distribution Function of the t -Distribution for Various Degrees of Freedom n Compared to the Standard Normal Density Function ($N(0,1)$)

against the standard normal density function. The same is done for the distribution function in Figure 5.

In general, the lower the degrees of freedom, the heavier the tails of the distribution, making extreme outcomes much more likely than for greater degrees of freedom or, in the limit, the normal distribution. This can be seen by the distribution function that we depicted in Figure 5 for $n = 1$ and $n = 5$ against the standard normal cumulative distribution function (cdf). For lower degrees of freedom such as $n = 1$, the solid curve starts to rise earlier and approach 1 later than for higher degrees of freedom such as $n = 5$ or the $N(0,1)$ case.

This can be understood as follows. When we rescale X by dividing by $\sqrt{S/n}$ as in equation (5), the resulting $X/\sqrt{S/n}$ obviously inherits randomness from both X and S . Now, when S is composed of few X_i , only, say $n = 3$, such that $X/\sqrt{S/n}$ has three degrees of freedom, there is a lot of dispersion from S relative to the standard normal distribution. By including more independent $N(0,1)$ random variables X_i such that the degrees of freedom increase, S becomes less dispersed. Thus, much uncertainty relative

to the standard normal distribution stemming from the denominator in $X/\sqrt{S/n}$ vanishes. The share of randomness in $X/\sqrt{S/n}$ originating from X alone prevails such that the normal characteristics preponderate. Finally, as n goes to infinity, we have something that is nearly standard normally distributed.

The mean of the Student's t random variable is zero, that is $E(X) = 0$, while the variance is a function of the degrees of freedom n as follows

$$\sigma^2 = \text{Var}(X) = \frac{n}{n-2}$$

For $n = 1$ and 2, there is no finite variance. Distributions with such small degrees of freedom generate extreme movements quite frequently relative to higher degrees of freedom. Precisely for this reason, stock price returns are often found to be modeled quite well using distributions with small degrees of freedom, or alternatively, large variances.

Application to Stock Returns

Let us resume the example at the end of the presentation of the normal distribution. We

consider, once again, the 5-day return Y with standard normal distribution. Suppose that now we do not know the variance. For this reason, at any point in time t , we rescale the observations of Y by

$$\sqrt{\frac{1}{5} \cdot (Y_{-1}^2 + Y_{-2}^2 + \cdots + Y_{-5}^2)}$$

where the $Y_{-1}^2, Y_{-2}^2, \dots, Y_{-5}^2$ are the five independent weekly returns immediately prior to Y . The resulting rescaled weekly returns

$$Z = \frac{Y}{\sqrt{Y_{-1}^2 + Y_{-2}^2 + \cdots + Y_{-5}^2}}$$

then are $t(5)$ distributed. The probability of Y not exceeding a value of 0.01 is now

$$P(Y \leq 0.01) = F(0.01) = 0.5083$$

where F is the cumulative distribution function of the Student's t -distribution with five degrees of freedom. Under the $N(0,1)$, this probability was about the same.

Again, we model the stock price after five days as $S_5 = S_0 \cdot e^Y$ where S_0 is today's price. As we know, when $Y \leq 0.01$, then $S_5 \leq S_0 \cdot e^{0.01} = S_0 \cdot 1.01$. Again, it follows that the stock price increases by at most 1% with probability of about 0.51. So far there is not much difference here between the standard normal and the $t(5)$ distribution.

Let's analyze the stock of American International Group (AIG) in September 2008. During one week, that is, five trading days, the stock lost about 67% of its value. That corresponds to a value of the 5-day return of $Y = -1.0986$ because of $e^Y = e^{-1.0986} = 0.3333 = 1 - 0.6667$. In the $N(0,1)$ model, a decline of this magnitude or even worse would occur with probability

$$P(Y \leq -1.0986) = \Phi(-1.0986) = 13.6\%$$

while under the $t(5)$ assumption, we would obtain

$$P(Y \leq -1.0986) = F(-1.0986) = 16.1\%$$

This is 2.5% more likely in the $t(5)$ model. So, stock price returns exhibiting extreme move-

ments such as that of the AIG stock price should not be modeled using the normal distribution.

F-DISTRIBUTION

Our next distribution is the F -distribution. It is defined as follows. Let $X \sim \chi^2(n_1)$ and $Y \sim \chi^2(n_2)$.

Furthermore, assuming X and Y to be independent, then the ratio

$$F(n_1, n_2) = \frac{X/n_1}{Y/n_2} \quad (7)$$

has an F -distribution with n_1 and n_2 degrees of freedom inherited from the underlying chi-square distributions of X and Y , respectively. We see that the random variable in equation (7) assumes nonnegative values only because neither X nor Y are ever negative. Hence, the support is on the nonnegative real numbers. Also like the chi-square distribution, the F -distribution is skewed to the right.

The F -distribution has a rather complicated looking density function of the form

$$f(x) = \begin{cases} \frac{F\left(\frac{n_1+n_2}{2}\right)}{F\left(\frac{n_1}{2}\right)F\left(\frac{n_2}{2}\right)} \cdot \left(\frac{n_1}{n_2}\right)^{n_1/2} \cdot \frac{x^{n_1/2-1}}{\left[1+x \cdot \frac{n_1}{2}\right]^{\frac{n_1+n_2}{2}}}, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (8)$$

Figure 6 displays the density function (8) for various degrees of freedom. As the degrees of freedom n_1 and n_2 increase, the function graph becomes more peaked and less asymmetric while the tails lose mass.

The mean is given by

$$E(X) = \frac{n_2}{n_2 - 2}, \quad \text{for } n_2 > 2 \quad (9)$$

while the variance equals

$$\sigma^2 = \text{Var}(X) = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)}, \quad \text{for } n_2 > 4 \quad (10)$$

Note that according to equation (9), the mean is not affected by the degrees of freedom n_1 of

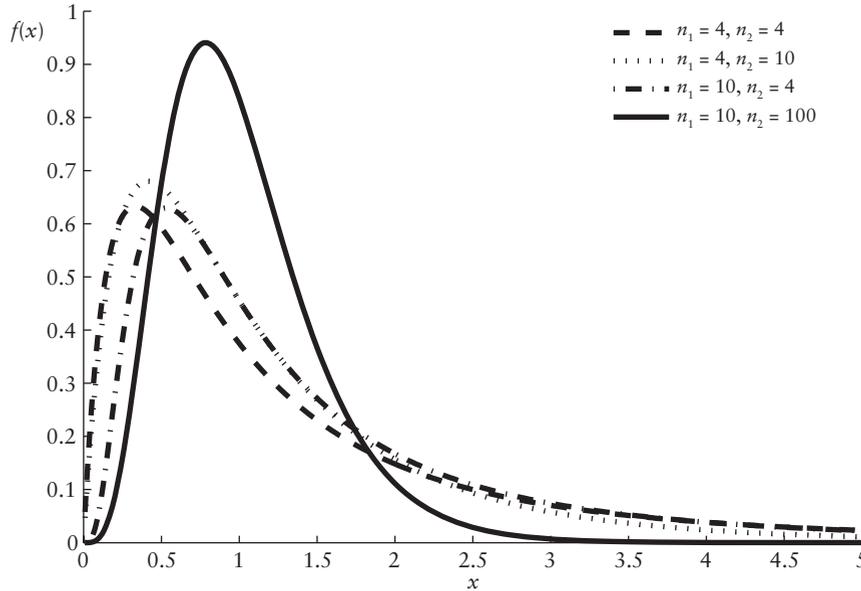


Figure 6 Density Function of the F -Distribution for Various Degrees of Freedom n_1 and n_2

the first chi-square random variable, while the variance in equation (10) is influenced by the degrees of freedom of both random variables.

EXPONENTIAL DISTRIBUTION

The *exponential distribution* is characterized by the positive real-valued parameter λ . In brief, we use the notation $Exp(\lambda)$. An exponential random variable assumes nonnegative values only. The density defined for $\lambda > 0$ by

$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda x}, & x \geq 0 \\ 0 & x < 0 \end{cases}$$

is right skewed. Figure 7 presents the density function for various parameter values λ .

The distribution function is obtained by simple integration as

$$F(x) = 1 - e^{-\lambda x}$$

For identical parameter values as in Figure 7, we have plots of the exponential distribution function shown in Figure 8.

For this distribution, both the mean and variance are relatively simple functions of the pa-

rameter. That is, for the mean

$$E(X) = \frac{1}{\lambda}$$

and for the variance

$$Var(X) = \frac{1}{\lambda^2}$$

There is an inverse relationship between the exponential distribution and the Poisson distribution. Suppose we have a Poisson random variable N with parameter λ , i.e., $N \sim Poi(\lambda)$, counting the occurrences of some event within a time frame of length T . Furthermore, let X_1, X_2, \dots be the $Exp(\lambda)$ distributed interarrival times between the individual occurrences. That is between time zero and the first event, X_1 units of time have passed, between the first event and the second, X_2 units of time have elapsed, and so on. Now, over these T units of time, we expect $T \cdot \lambda = T \cdot E(N)$ events to occur. Alternatively, we have an average of $T/(T \cdot \lambda) = 1/\lambda = E(X)$ units of time to wait between occurrences.

Suppose that by time T we have counted exactly n events. Then the accrued time τ elapsed when the event occurs for the n th time is obtained by the sum of all individual interarrival

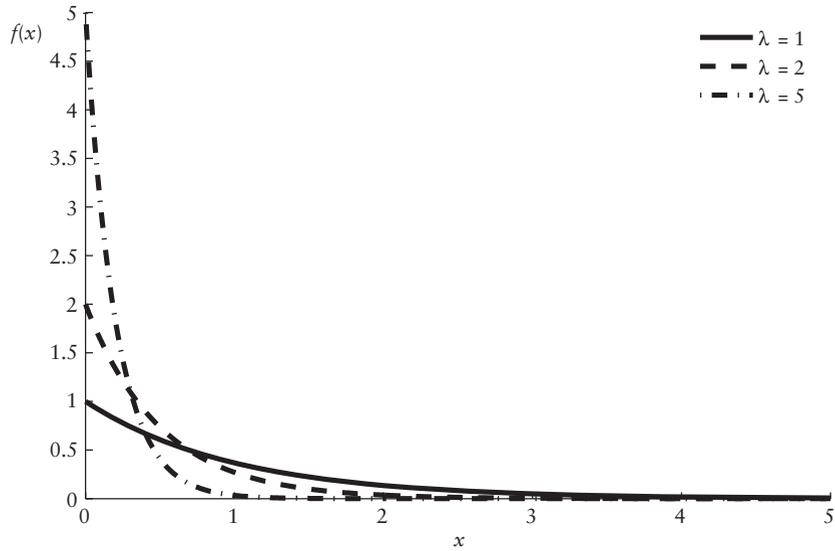


Figure 7 Exponential Density Function for Various Parameter Values λ

times X_1, X_2, \dots, X_n , which cannot be greater than T . Formally

$$\tau = \sum_i^n X_i \leq T \quad (11)$$

A result of this relationship is

$$E(N) = \lambda = \frac{1}{E(X)}$$

The exponential distribution is commonly referred to as a distribution with a “no memory” property in the context of life-span that ends due to some break.

That means that there is no difference in the probability between the following two events. Event one states that the object will live for the first τ units of time after the object’s creation while event two states that the object will

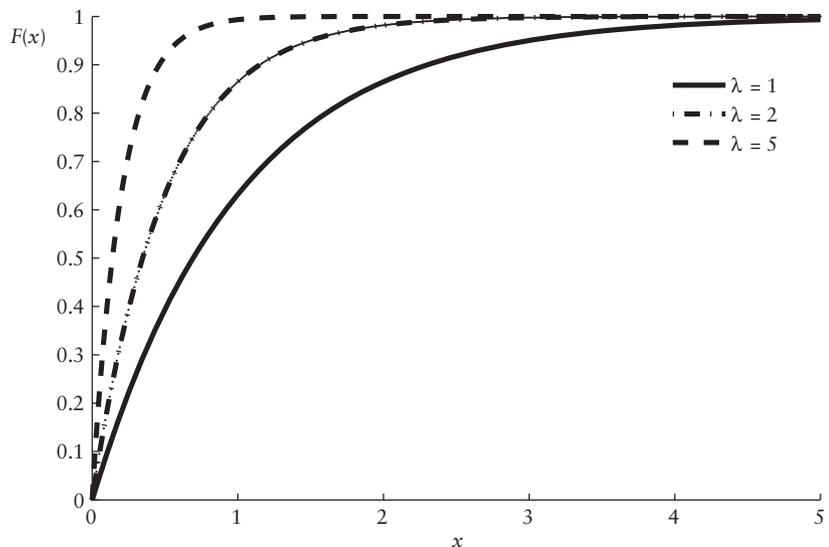


Figure 8 Distribution Function $F(x)$ of the Exponential Distribution for Various Parameter Values λ

continue living for the next τ units of time after it has already survived some t units of time. In other words, if some *interarrival time* or *survival time* (i.e., the time between certain occurrences) is $Exp(\lambda)$ distributed, one starts all over waiting at any given time t provided that the break has not occurred yet. (Technically, these considerations as well as the following equation (12) require the understanding of the notion *conditional distributions*. Here it will suffice to apply pure intuition.) So, for example, let the time until the next default of one of several corporate bonds held in some portfolio be given as an exponential random variable. Then the probability of the first bond defaulting in no more than t units of time given that none have defaulted so far is the same as the probability of the n th bond defaulting after at most t units of time given that $n - 1$ bonds have already defaulted. That is, we only care about the probability distribution of the time of occurrence of the next default regardless of how many bonds have already defaulted.

Finally, an additional property of the exponential distribution is its relationship to the chi-square distribution. Let X be $Exp(\lambda)$. Then X is also chi-square distributed with two degrees of freedom, that is, $X \sim \chi^2(2)$.

Applications in Finance

In applications in finance, the parameter λ often has the meaning of a *default rate*, *default intensity*, or *hazard rate*. This can be understood by observing the ratio

$$\frac{P(X \in (t, t + dt])}{dt \cdot P(X > t)} \quad (12)$$

which expresses the probability of the event of interest such as default of some company occurring between time t and $t + dt$ given that it has not happened by time t , relative to the length of the horizon, dt . Now, let the length of the interval, dt , approach zero, and this ratio in equation (12) will have λ as its limit.

The exponential distribution is often used in credit risk models where the number of defaulting bonds or loans in some portfolio

over some period of time is represented by a Poisson random variable and the random times between successive defaults by exponentially distributed random variables. In general, then, the time until the n th default is given by the sum in equation (11).

Consider, for example, a portfolio of bonds. Moreover, we consider the number of defaults in this portfolio in one year to be some Poisson random variable with parameter $\lambda = 5$, that is, we expect five defaults per year. The same parameter, then, represents the default intensity of the exponentially distributed time between two successive defaults, that is, $\tau \sim Exp(5)$, so that on average, we have to wait $E(\tau) = 1/5$ of a year or 2.4 months. For example, the probability of less than three months (i.e., $1/4$ of a year) between two successive defaults is given by

$$P(\tau \leq 0.25) = 1 - e^{-5 \cdot 0.25} = 0.7135$$

or roughly 71%. Now, the probability of no default in any given year is then

$$P(\tau > 1) = e^{-5 \cdot 1} = 0.0067$$

or 0.67%.

RECTANGULAR DISTRIBUTION

The simplest continuous distribution we are going to introduce is the *rectangular distribution*. Often, it is used to generate simulations of random outcomes of experiments via transformation. If a random variable X is rectangular distributed, we denote this by $X \sim Re(a, b)$ where a and b are the parameters of the distribution.

The support is on the real interval $[a, b]$. The density function is given by

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0 & x \notin [a, b] \end{cases} \quad (13)$$

We see that this density function is always constant, either zero or between the bounds a and b , equal to the inverse of the interval width. Figure 9 displays the density function (13) for some general parameters a and b .

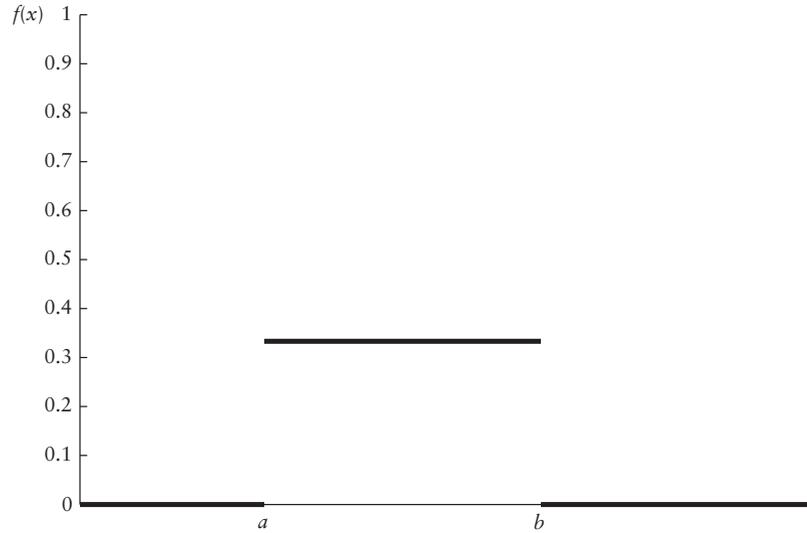


Figure 9 Density Function of a $Re(a, b)$ Distribution

Through integration, the distribution function follows in the form

$$F(x) = \begin{cases} 0 & x < a \\ \frac{1}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases} \quad (14)$$

The mean is equal to

$$E(X) = \frac{a+b}{2}$$

$$Var(X) = \frac{(b-a)^2}{12}$$

In Figure 10, we have the distribution function given by equation (14) with some general parameters a and b . By analyzing the plot, we can see that the distribution function is not differentiable at a or b , since the derivatives of F do not exist for these values. At any other real

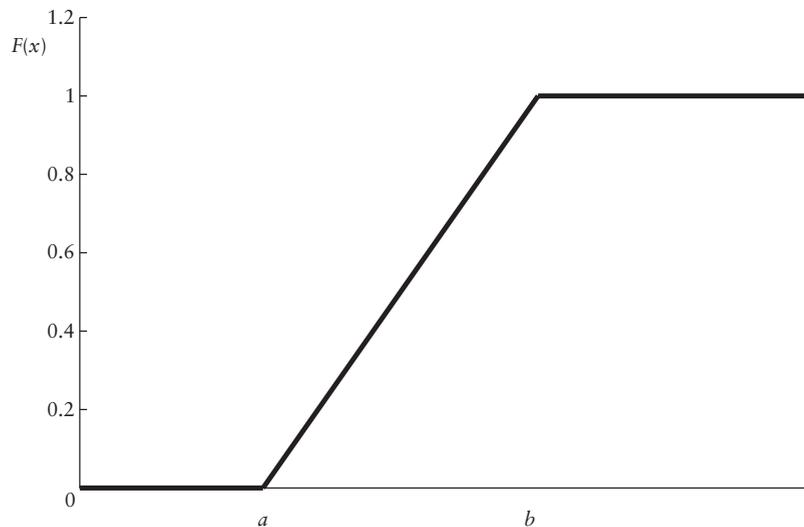


Figure 10 Distribution Function of a $Re(a, b)$ Distribution

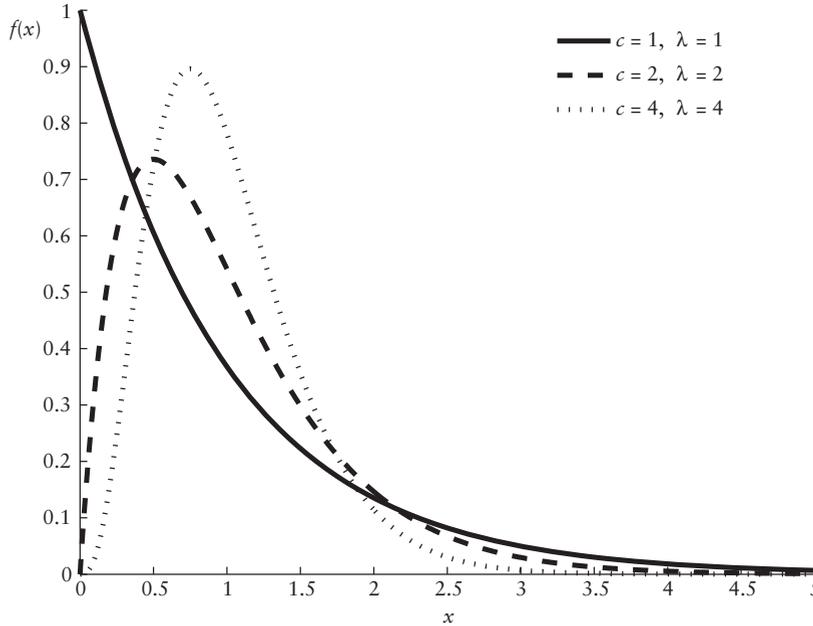


Figure 11 Density Function of a Gamma Distribution $Ga(\lambda, b)$

value x , the derivative exists (being 0) and is continuous. We say in the latter case that f is *smooth* there.

GAMMA DISTRIBUTION

Next we introduce the *gamma distribution* for positive, real-valued random variables. Characterized by two parameters, λ and c , this distribution class embraces several special cases. It is skewed to the right with support on the positive real line. We denote that a random variable X is gamma distributed with parameter λ and c by writing $X \sim Ga(\lambda, c)$ where λ and c are positive real numbers.

The density function is given by

$$f(x) = \begin{cases} \frac{\lambda(\lambda x)^{c-1} \exp\{-\lambda x\}}{\Gamma(c)}, & x > 0 \\ 0 & x \leq 0 \end{cases} \quad (15)$$

with gamma function Γ . A plot of the density function from equation (15) is provided in

Figure 11. The distribution function is

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{\int_0^{\lambda x} u^{c-1} e^{-u} du}{b^c \Gamma(c)}, & x \geq 0 \end{cases}$$

The mean is

$$E(X) = \frac{c}{\lambda}$$

with variance

$$Var(X) = \frac{c}{\lambda^2}$$

Erlang Distribution

A special case is the *Erlang distribution*, which arises for natural number values of the parameter c , that is, $c \in \mathbb{N}$. The intuition behind it is as follows. Suppose we have c exponential random variables with the same parameter λ , that is, $X_1, X_2, \dots, X_c \sim Exp(\lambda)$ all being independent of each other. Then the sum of these

$$S = \sum_{i=1}^c X_i$$

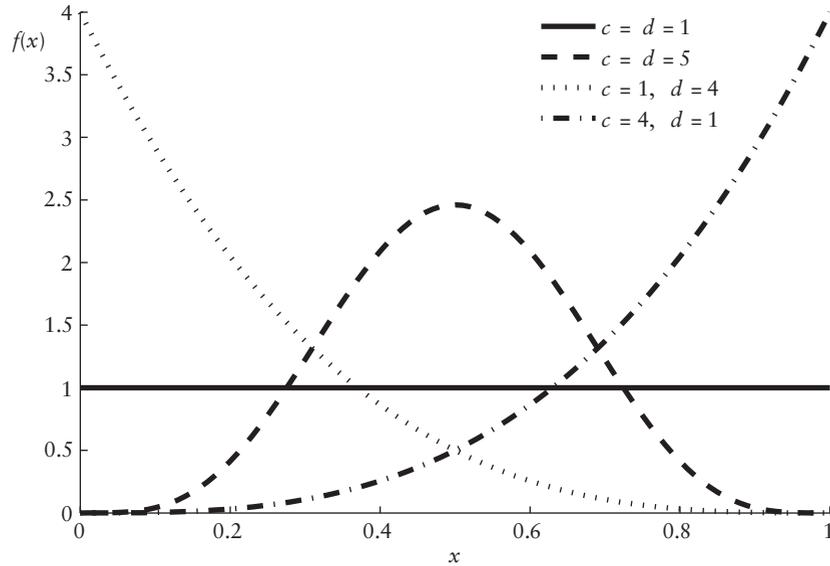


Figure 12 Density Function of a Beta Distribution $Be(c, d)$

is distributed $Ga(\lambda, c)$ such that the resulting distribution function is

$$F(s) = \begin{cases} 0 & s < 0 \\ 1 - e^{-\lambda s} \sum_{i=1}^{c-1} \frac{(\lambda s)^i}{i!}, & s \geq 0 \end{cases}$$

So, when we add the identically $Exp(\lambda)$ distributed interarrival times until the c th default, for example, the resulting combined waiting time is Erlang distributed with parameters c and λ .

BETA DISTRIBUTION

The *beta distribution* is characterized by the two parameters c and d that are any positive real numbers. We abbreviate this distribution by $Be(c, d)$. It has a density function with support on the interval $[0,1]$, that is, only for $x \in [0,1]$ does the density function assume positive values. In the context of credit risk modeling, it commonly serves as an approximation for generating random defaults when the true underlying probabilities of default of certain companies are unknown.

The density function is defined by

$$f(x) = \begin{cases} \frac{1}{B(c, d)} x^{c-1} (1-x)^{d-1}, & 0 \leq x \leq 1 \\ 0 & \text{else} \end{cases}$$

where $B(c, d)$ denotes the *beta function* with parameters c and d . The density function may assume various different shapes depending on c and d . For a few exemplary values, we present the plots in Figure 12. As we can see, for $c = d$, the density function is symmetric about $x = 0.5$.

LOG-NORMAL DISTRIBUTION

Another important distribution in finance is the *log-normal distribution*. It is connected to the normal distribution via the following relationship. Let Y be a normal random variable with mean μ and variance σ^2 . Then the random variable

$$X = e^Y$$

is log-normally distributed with parameters μ and σ^2 . In brief, we denote this distribution by $X \sim Ln(\mu, \sigma^2)$.

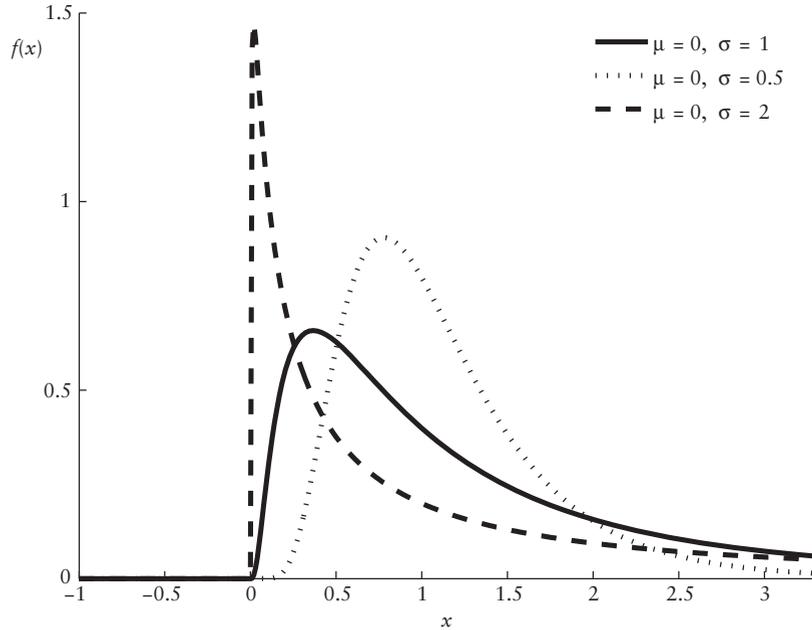


Figure 13 Density Function of the Log-Normal Distribution for Various Values of μ and σ^2

Since the exponential function $e^Y = \exp(Y)$ only yields positive values, the support of the log-normal distribution is on the positive half of the real line only, as will be seen by its density function given by

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, & x > 0 \\ 0 & \text{else} \end{cases} \quad (16)$$

which looks strikingly similar to the normal density function given by (2). Figure 13 depicts the density function for several parameter values.

This density function results in the log-normal distribution function

$$F(x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right)$$

where $\Phi(\cdot)$ is the distribution function of the standard normal distribution. (This is the result of the one-to-one relationship between the values of a log-normal and a standard normal random variable.) A plot of the distribution function for different parameter values can be found in Figure 14.

Mean and variance of a log-normal random variable are

$$E(X) = e^{(\mu + \sigma^2/2)} \quad (17)$$

and

$$\text{Var}(X) = e^{\sigma^2} (e^{\sigma^2} - 1) e^{2\mu} \quad (18)$$

Application to Modeling Asset Returns

The reason for the popularity of the log-normal distribution is that logarithmic asset returns r have been historically modeled as normally distributed such that the related asset prices are modeled by a log-normal distribution. That is, let P_t denote today's asset price and, furthermore, let the daily return r be $N(\mu, \sigma^2)$. Then in a simplified fashion, tomorrow's price is given by $P_{t+1} = P_t \cdot e^r$ while the percentage change between the two prices, e^r , is log-normally distributed, that is, $Ln(\mu, \sigma^2)$.

The log-normal distribution is closed under special operations as well. If we let the n

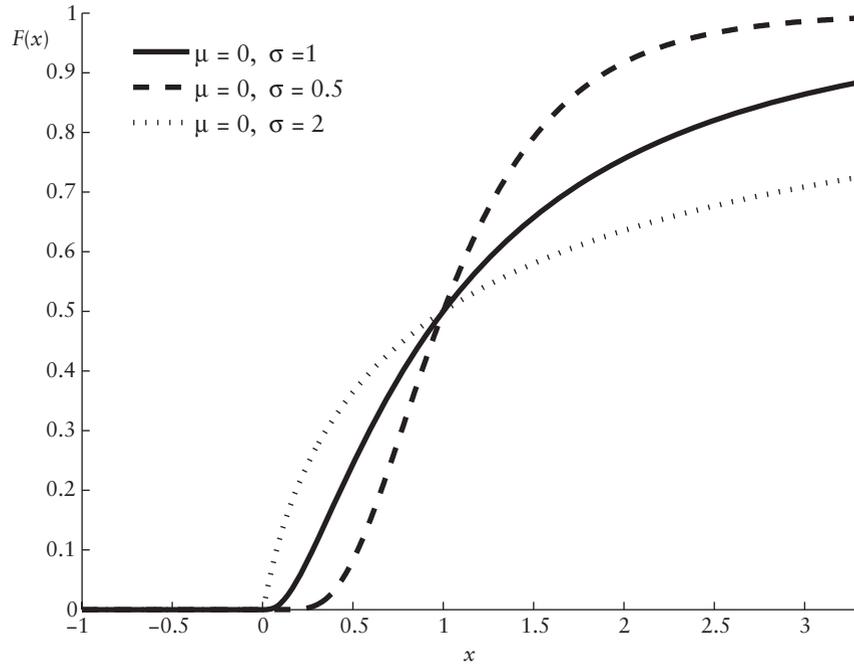


Figure 14 Distribution Function of the Log-Normal Distribution for Various Parameter Values μ and σ^2

random variables X_1, \dots, X_n be log-normally distributed each with parameters μ and σ^2 and uninfluenced by each other, then multiplying all of these and taking the n th root we have that

$$\sqrt[n]{\prod_{i=1}^n X_i} \sim Ln(\mu, \sigma^2)$$

where the product sign is defined as

$$\prod_{i=1}^n X_i \equiv X_1 \times X_2 \times \dots \times X_n$$

As an example, we consider a very simplified stock price model. Let $S = \$100$ be today's stock price of some company. We model tomorrow's price S_1 as driven by the 1-day dynamic X from the previous example of the normal distribution. In particular, the model is

$$S_1 = S_0 \cdot e^X$$

By some slight manipulation of the above equation, we see that the ratio of tomorrow's

price over today's price

$$\frac{S_1}{S_0} = e^X$$

follows a log-normal distribution with parameters μ and σ , that is, $S_1/S_0 \sim LN(\mu, \sigma^2)$. We may now be interested in the probability that tomorrow's price is greater than \$120; that is,

$$\begin{aligned} P(S_1 > 120) &= P(S_0 e^X > 120) \\ &= P(100 \cdot e^X > 120) \end{aligned}$$

This corresponds to

$$\begin{aligned} P\left(\frac{S_1}{S_0} > \frac{120}{S_0}\right) &= P(e^X > 1.20) \\ &= 1 - P(e^X \leq 1.20) \\ &= 1 - F(1.2) \\ &= 1 - 0.8190 = 0.1810 \end{aligned}$$

where in the third equation on the right-hand side, we have applied the log-normal cumulative probability distribution function F . So, in roughly 18% of the scenarios, tomorrow's stock price S_1 will exceed the price of today, $S_0 = \$100$,

by at least 20%. From equation (17), the mean of the ratio is

$$E\left(\frac{S_1}{S_0}\right) = \mu_{S_1/S_0} = e^{0+\frac{0.2}{2}} = 1.1052$$

implying that we have to expect tomorrow's stock price to be roughly 10% greater than today, even though the dynamic X itself has an expected value of 0. Finally, equation (18) yields the variance

$$\text{Var}\left(\frac{S_1}{S_0}\right) = \sigma_{S_1/S_0}^2 = e^{0.2}(e^{0.2} - 1) = 0.2704$$

which is only slightly larger than that of the dynamic X itself.

The statistical concepts learned to this point can be used for pricing certain types of derivative instruments such as the Black-Scholes option pricing model.

KEY POINTS

- The more commonly used distributions with appealing statistical properties that are used in finance are the normal distribution, the chi-square distribution, the Student's t -distribution, the Fisher's F -distribution, the exponential distribution, the gamma, the beta distribution, and the log-normal distribution.
- The normal distribution is probably the most famous probability distribution. Its popularity is credited to the fact that it serves as the distribution of many random sums of random variables. Moreover, it serves as the origin for many other probability distributions with appealing properties.
- The empirical rule is helpful in assessing how the data of most samples are dispersed even if we do not know the underlying distribution. The theoretical counterpart, the Chebychev inequality, provides limits for the dispersion of any probability distribution whose variance we know.
- Logarithmic returns in contrast to percentage returns is the most commonly used method to express changes of asset prices. The reason for the widespread use of returns computed in terms of logarithms lies in the simple mathematical tractability of their form. Moreover, their intuitive appeal results from the fact that they can be understood as the relative price changes obtained from constant trading.
- The default intensity finds extended use in financial models considering stochastic default such as the default of some bond in a bond portfolio. It expresses the probability of defaulting within the next unit of time interval as we let the length of this interval approach zero.
- The interarrival time is the random variable associated with the time between two successive random events. For example, for a bond portfolio manager it is of interest to model the time between some default in the portfolio and the next default. Commonly, the interarrival time is modeled as an exponential random variable.

NOTES

1. There exist generalizations such that the distributions need no longer be identical. However, this is beyond the scope of this entry.
2. For values near 0, the logarithmic return X is virtually equal to the multiplicative return R . Rounding to two decimals, they are both equal to 0.01 here.
3. For some computer software, the probability will be given as 0.5 due to rounding.

REFERENCE

- Evans, M., Hastings, N., and Peacock, B. (2000). *Statistical Distributions*, 3rd ed. New York: Wiley.

Continuous Probability Distributions Dealing with Extreme Events

MARKUS HÖCHSTÖTTER, PhD

Assistant Professor, University of Karlsruhe

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair-Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Continuous probability distributions are commonly the preferred candidates when modeling financial asset returns. The most popular of them is unquestionably the normal distribution because of its appealing properties as well as the fact that it serves as the limit distribution for many sums of random variables such as, for example, aggregated returns. The normal distribution generally renders modeling easy because all moments exist. However, the normal distribution fails to reflect stylized facts commonly encountered in asset returns, namely, the possibility of very extreme movements and skewness. To remedy this shortcoming, probability distributions accounting for such extreme price changes have become increasingly popular. Some of these distributions concentrate exclusively on extreme values and others permit any real number, but in a manner that is capable of reflecting market behavior. Consequently, there is a selection of probability distributions that can realistically reproduce asset price changes. Their common shortcoming is generally that they are mathematically difficult to handle.

In this entry, we present a collection of continuous probability distributions that are used in finance in the context of modeling extreme events. Although there are distributions that are appealing in nature due to their mathematical simplicity, the ones introduced in this entry are sometimes rather complicated, using parameters that are not necessarily intuitive. However, due to the observed behavior of many quanti-

ties in finance, there is a need for more flexible distributions compared to keeping models mathematically simple.

While the Student's t -distribution is able to mimic some behavior inherent in financial data such as so-called *heavy tails* (which means that a lot of the probability mass is attributed to extreme values), it fails to capture other observed behavior such as skewness. Hence,

we decided not to include the Student's t -distribution in this entry.

In this entry, we present the generalized extreme value distribution, the generalized Pareto distribution, the normal inverse Gaussian distribution, and the α -stable distribution together with their parameters of location and spread. The presentation of each distribution is accompanied by some illustration to help render the theory more appealing.

GENERALIZED EXTREME VALUE DISTRIBUTION

Sometimes it is of interest to analyze the probability distribution of extreme values of some random variable rather than the entire distribution. This occurs in risk management (including operational risk, credit risk, and market risk) and risk control in portfolio management. For example, a portfolio manager may be interested in the maximum loss a portfolio might incur with a certain probability. For this purpose, *generalized extreme value (GEV) distributions* are designed. They are characterized by the real-valued parameter ξ . Thus, the abbreviated appellation for this distribution is $GEV(\xi)$.

Technically, one considers a series of identically distributed random variables X_1, X_2, \dots, X_n , which are independent of each other so that each one's value is unaffected by the others' outcomes. Now, the GEV distributions become relevant if we let the length of the series n become ever larger and consider its largest value, that is, the maximum.

The distribution is not applied to the data immediately but, instead, to the so-called *standardized data*. Basically, when standardizing data x , one reduces the data by some constant real parameter a and divides it by some positive parameter b so that one obtains the quantity $(x - a)/b$. (Standardization is a linear transform of the random variable such that its location parameter becomes zero and its scale one.) The parameters are usually chosen such that this standardized quantity has zero mean and unit variance. We have to point out that neither vari-

ance nor mean have to exist for all probability distributions.

Extreme value theory, a branch of statistics that focuses solely on the extremes (tails) of a distribution, distinguishes between three different types of generalized extreme value distributions: Gumbel distribution, Fréchet distribution, and Weibull distribution. In the extreme value theory literature, these distributions are referred to respectively as Type I, Type II, and Type III. (See Embrechts, Klüppelberg, and Mikosch [2003], De Haan and Ferreira [2006], and Kotz and Nadarajah [2002].) The three types are related in that we obtain one type from another by simply varying the value of the parameter ξ . This makes GEV distributions extremely pleasant for handling financial data.

For the *Gumbel distribution*, the general parameter is zero (i.e., $\xi = 0$) and its density function is

$$f(x) = e^{-x} \exp \{-e^{-x}\}$$

A plot of this density is given by the dashed graph in Figure 1 that corresponds to $\xi = 0$. The distribution function of the Gumbel distribution is then

$$F(x) = \exp \{-e^{-x}\}$$

Again, for $\xi = 0$, we have the distribution function displayed by the dashed graph in Figure 2.

The second $GEV(\xi)$ distribution is the *Fréchet distribution*, which is given for $\xi > 0$ and has density

$$f(x) = (1 + \xi x)^{-\xi-1} \cdot \exp\{-x^{-\xi}\}$$

with corresponding distribution function

$$F(x) = \exp \left\{ -(1 + \xi x)^{-1/\xi} \right\}$$

Note that the prerequisite $1 + \xi x > 0$ has to be met. For a parameter value of $\xi = 0.5$, an example of the density and distribution function is given by the dotted graphs in Figures 1 and 3, respectively.

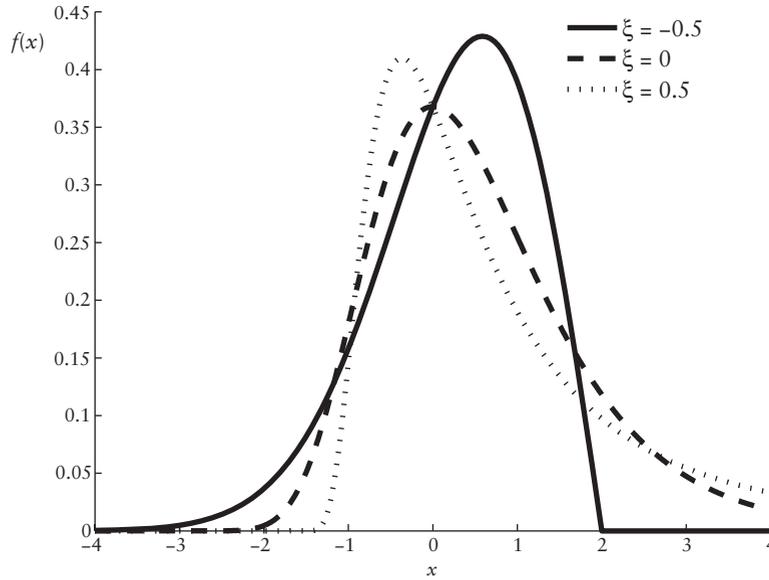


Figure 1 $GEV(\xi)$ Density Function for Various Parameter Values

Finally, the *Weibull distribution* corresponds to $\xi < 0$. It has the density function

$$f(x) = (1 + \xi x)^{-\xi-1} \cdot \exp \{-x^{-\xi}\}$$

A plot of this distribution can be seen in Figure 1, with $\xi = -0.5$ (solid line). Again, $1 + \xi x > 0$ has to be met. It is remarkable that

the density function graph vanishes in a finite right end point, that is, becomes zero. Thus, the support is on $(-\infty, -1/\xi)$. The corresponding distribution function is

$$F(x) = \exp \{-(1 + x)^{-1/\xi}\}$$

a graph of which is depicted in Figure 2 for $\xi = -0.5$ (solid line).

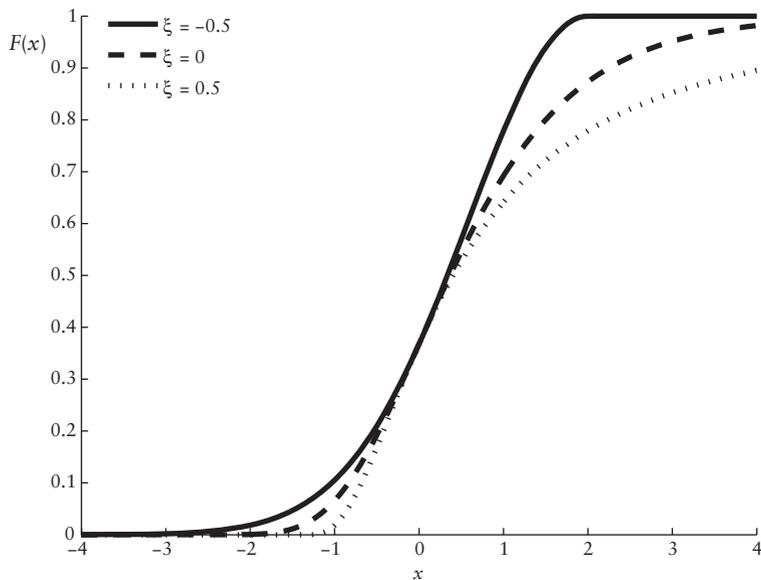


Figure 2 $GEV(\xi)$ Distribution Function for Various Parameter Values

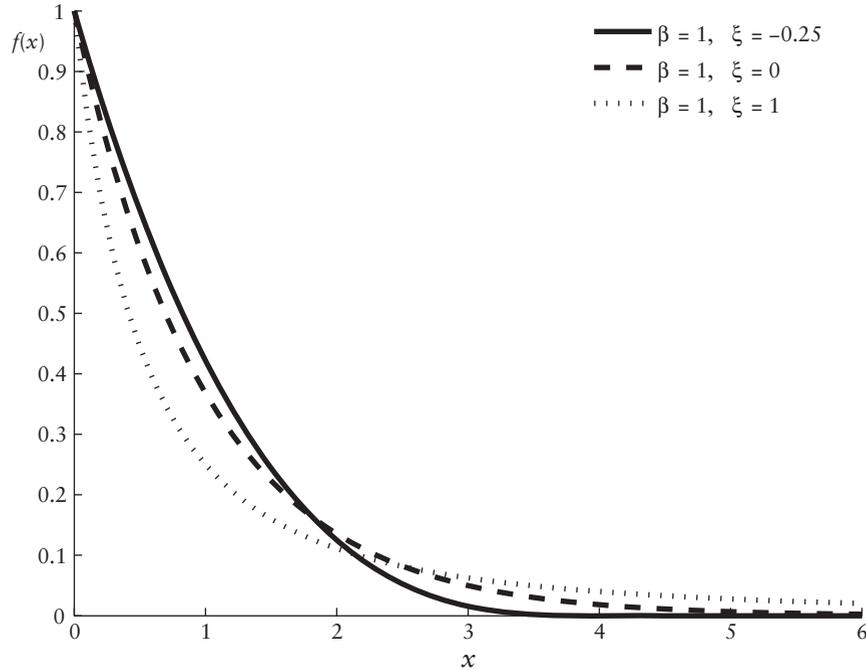


Figure 3 Generalized Pareto Density Function for Various Parameter Values

Notice that the extreme parts of the density function (i.e., the tails) of the Fréchet distribution vanish more slowly than that of the Gumbel distribution. Consequently, a Fréchet type distribution should be applied when dealing with scenarios of large extremes.

GENERALIZED PARETO DISTRIBUTION

A distribution often employed to model large values, such as price changes well beyond the typical change, is the *generalized Pareto distribution* or, as we will often refer to it here, simply *Pareto distribution*. This distribution serves as the distribution of the so called “peaks over thresholds,” which are values exceeding certain benchmarks or loss severity.

For example, consider n random variables X_1, X_2, \dots, X_n that are all identically distributed and independent of each other. Slightly idealized, they might represent the returns of some stock on n different observation days. As the

number of observations n increases, suppose that their maximum observed return follows the distribution law of a GEV distribution with parameter ξ . Furthermore, let u be some sufficiently large threshold return. Suppose that on day i , the return exceeded this threshold. Then, given the exceedance, the amount return X_i surpassed u by, that is, $X_i - u$, is a generalized Pareto distributed random variable.

The following density function characterizes the Pareto distribution

$$f(x) = \begin{cases} \frac{1}{\beta} \left(1 + \xi \frac{x}{\beta}\right)^{-1-1/\xi}, & x \geq 0 \\ 0 & \text{else} \end{cases}$$

with $\beta > 0$ and $1 + (\xi x)/\beta > 0$. Hence, the distribution is right skewed since the support is only on the positive real line. The corresponding distribution function is given by

$$F(x) = \frac{1}{\beta} \left(1 + \xi \frac{x}{\beta}\right)^{-1-1/\xi}, \quad x \geq 0$$

As we can see, the Pareto distribution is characterized by two parameters β and ξ . In brief,

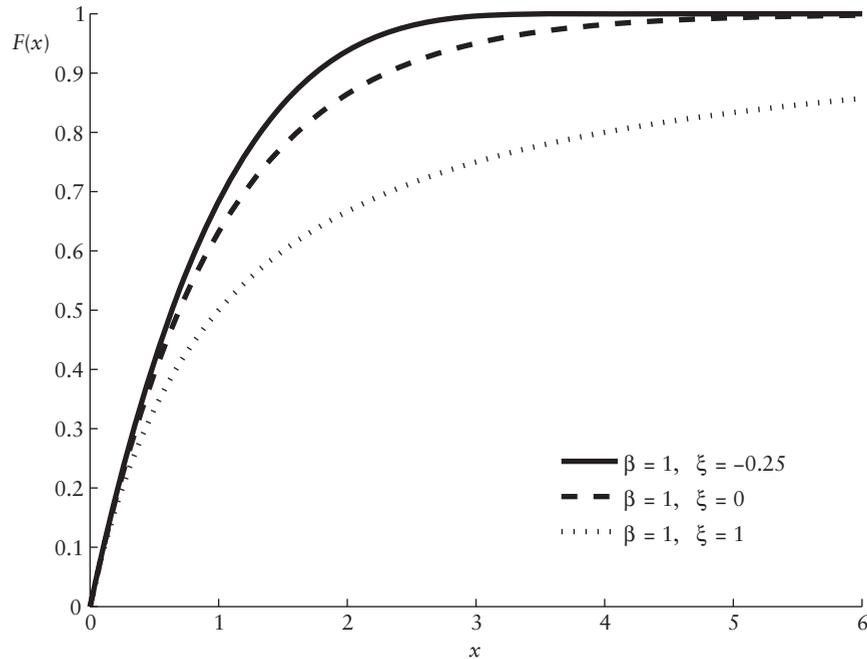


Figure 4 Generalized Pareto Distribution Function for Various Parameter Values

the distribution is denoted by $Pa(\beta, \xi)$. The parameter β serves as a scale parameter while the parameter ξ is responsible for the overall shape as becomes obvious by the density plots in Figure 3. The distribution function is displayed, in Figure 4, for a selection of parameter values.

For $\beta < 1$, the mean is

$$E(X) = \beta/1 - \xi$$

When β becomes very small approaching zero, then the distribution results in the exponential distribution with parameter $\lambda = 1/\beta$.

The Pareto distribution is commonly used to represent the tails of other distributions. For example, while in neighborhoods about the mean, the normal distribution might serve well to model financial returns; for the tails (i.e., the end parts of the density curve), however, one might be better advised to apply the Pareto distribution. The reason is that the normal distribution may not assign sufficient probability to more pronounced price changes measured in log-returns. On the other hand, if one wishes to model behavior that attributes less probability

to extreme values than the normal distribution would suggest, this could be accomplished by the Pareto distribution as well. The reason why the class of the Pareto distributions provides a prime candidate for these tasks is due to the fact that it allows for a great variety of different shapes one can smoothly obtain by altering the parameter values.

NORMAL INVERSE GAUSSIAN DISTRIBUTION

Another candidate for the modeling of financial returns is the *normal inverse Gaussian distribution*. It is considered suitable since it assigns a large amount of probability mass to the tails. This reflects the inherent risks in financial returns that are neglected by the normal distribution since it models asset returns behaving more moderately. But in recent history, we have experienced more extreme shocks than the normal distribution would have suggested with reasonable probability.

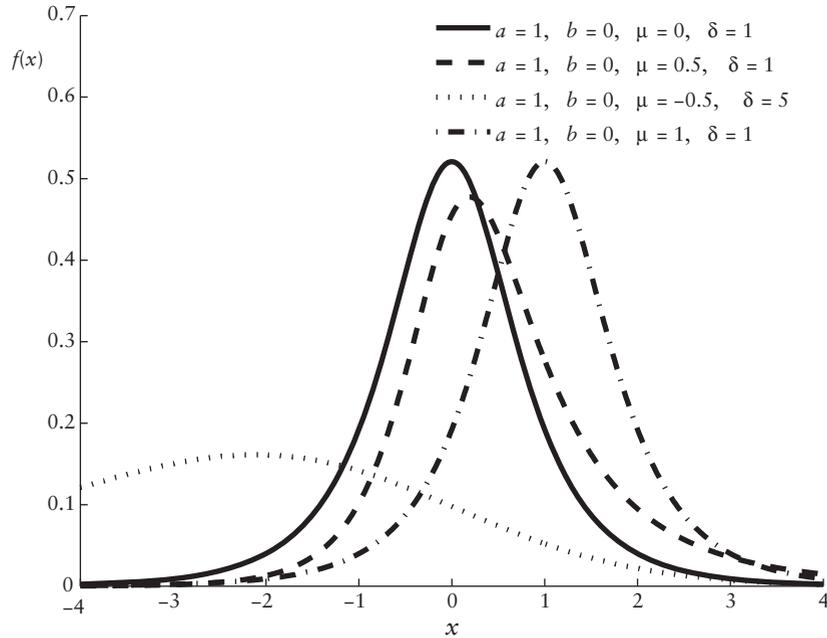


Figure 5 Normal Inverse Gaussian Density Function for Various Parameter Values

The distribution is characterized by four parameters, a , b , μ , and δ . In brief, the distribution is denoted by $NIG(a, b, \mu, \delta)$. For real values x , the density function is given by

$$f(x) = \frac{a \cdot \delta}{\pi} \exp \left\{ \delta \sqrt{a^2 - b^2} + b(x - \mu) \right\} \times \frac{K_1 \left(a \sqrt{\delta^2 + (x - \mu)^2} \right)}{\sqrt{\delta^2 + (x - \mu)^2}}$$

where K_1 is the so-called *Bessel function of the third kind*. In Figure 5, we display the density function for a selection of parameter values.

The distribution function is, as in the normal distribution case, not analytically presentable. It has to be determined with the help of numerical methods. We display the distribution function for a selection of parameter values in Figure 6.

The parameters have the following interpretation. Parameter a determines the overall shape of the density while b controls skewness. The location or position of the density function is governed via parameter μ and δ is responsible for scaling. These parameters have values

according to the following

$$\begin{aligned} a &> 0 \\ 0 &\leq b < a \\ \mu &\in \mathbb{R} \\ \delta &> 0 \end{aligned}$$

The mean of a *NIG* random variable is

$$E(X) = \mu + \frac{\delta \cdot b}{\sqrt{a^2 - b^2}}$$

and the variance is

$$\text{Var}(X) = \delta \frac{a^2}{\left(\sqrt{a^2 - b^2} \right)^3}$$

Normal Distribution versus Normal Inverse Gaussian Distribution

Due to a relationship to the normal distribution that is beyond the scope here, there are some common features between the normal and NIG distributions.

The *scaling property* of the *NIG* distribution guarantees that any *NIG* random variable multiplied by some real constant is again a *NIG* random variable. Formally, for some $k \in \mathbb{R}$ and

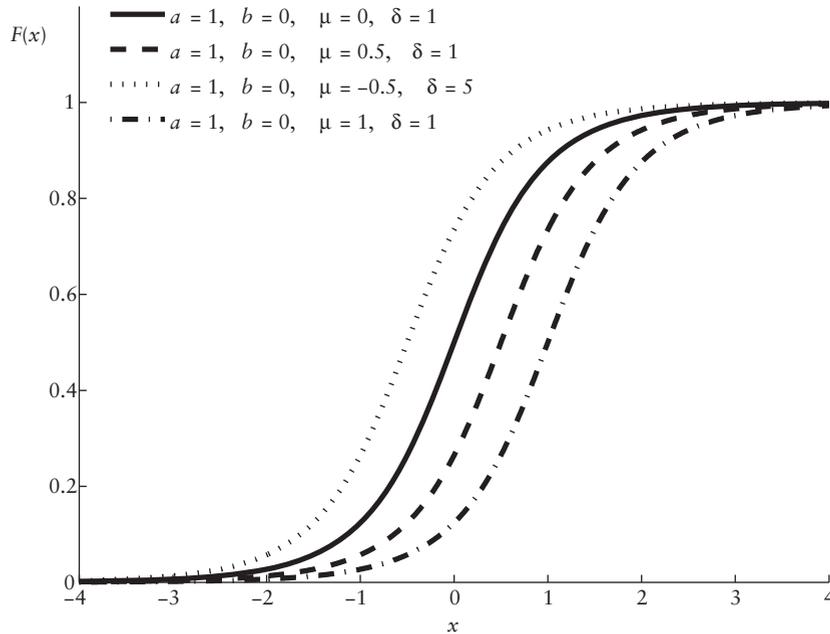


Figure 6 Normal Inverse Gaussian Distribution Function for Various Parameter Values

$X \sim NIG(a, b, \mu, \delta)$, we have that

$$k \cdot X \sim NIG\left(\frac{a}{k}, \frac{b}{k}, k \cdot \mu, k \cdot \delta\right) \quad (1)$$

Among others, the result in equation (1) implies that the factor k shifts the density function by the k -fold of the original position. Moreover, we can reduce skewness in that we inflate X by some factor k .

Also, the NIG distribution is summation stable such that, under certain prerequisites concerning the parameters, independent NIG random variables are again NIG . More precisely, if we have the random variables $X_1 \sim NIG(a, b, \mu_1, \delta_1)$ and $X_2 \sim NIG(a, b, \mu_2, \delta_2)$, the sum is $X_1 + X_2 \sim NIG(a, b, \mu_1 + \mu_2, \delta_1 + \delta_2)$. So, we see that only location and scale are affected by summation.

α -STABLE DISTRIBUTION

The final distribution we introduce is the class of α -stable distributions. (For a further discussion of stable distributions, see Samorodnitsky and Taquq [2000].) Often, these distri-

butions are simply referred to as *stable distributions*. While many models in finance have been modeled historically using the normal distribution based on its pleasant tractability, concerns have been raised that it underestimates the danger of downturns of extreme magnitude inherent in stock markets. The sudden declines of stock prices experienced during several crises since the late 1980s—October 19, 1987 (“Black Monday”), July 1997 (“Asian currency crisis”), 1998–1999 (“Russian ruble crisis”), 2001 (“Dot-com bubble”), and July 2007 and following (“Subprime mortgage crisis”)—are examples that call for distributional alternatives accounting for extreme price shocks more adequately than the normal distribution. This may be even more necessary considering that financial crashes with serious price movements might become even more frequent in time given the major events that transpired throughout the global financial markets in 2008. The immense threat radiating from heavy tails in stock return distributions made industry professionals aware of the urgency to take them seriously and reflect them in their models.

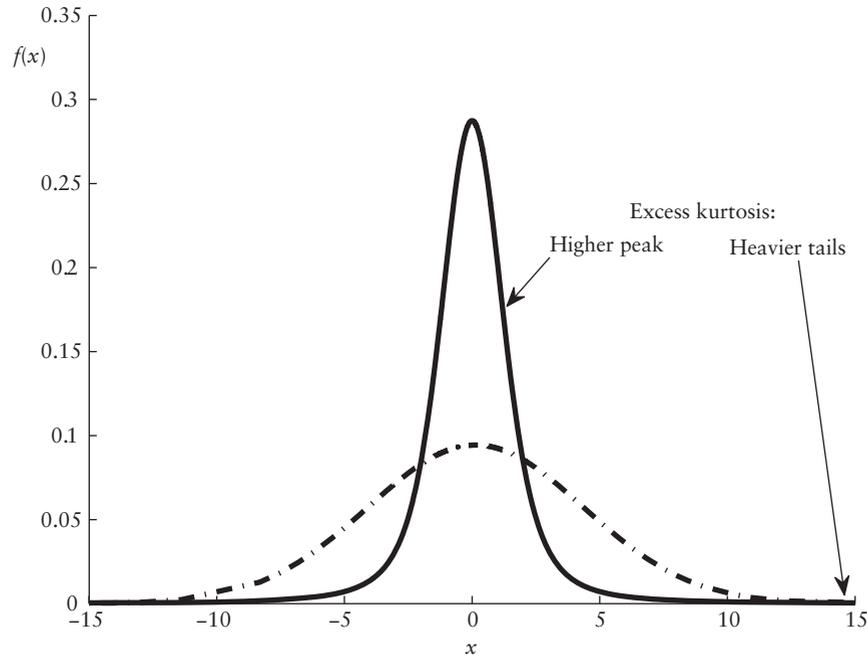


Figure 7 Comparison of the Normal (Dash-Dotted) and α -Stable (Solid) Density Functions

Many distributional alternatives providing more realistic chances to severe price movements are known, such as the Student's t , for example, or the GEV distributions presented earlier in this entry. In the early 1960s, Benoit Mandelbrot suggested as a distribution for commodity price changes the class of stable distributions. The reason is that, through their particular parameterization, they are capable of modeling moderate scenarios as supported by the normal distribution as well as extreme ones beyond the scope of most of the distributions that we have presented in this entry.

The stable distribution is characterized by the four parameters α , β , σ , and μ . In brief, we denote the α -stable distribution by $S(\alpha, \beta, \sigma, \mu)$. Parameter α is the so called *tail index* or *characteristic exponent*. It determines how much probability is assigned around the center and the tails of the distribution. The lower the value α , the more pointed about the center is the density and the heavier are the tails. These two features are referred to as *excess kurtosis* relative to the normal distribution. This can be visualized

graphically as we have done in Figure 7 where we compare the normal density to an α -stable density with a low $\alpha = 1.5$. The parameters for the normal distribution are $\mu = 0.14$ and $\sigma = 4.23$. The parameters for the stable distribution are $\alpha = 1.5$, $\beta = 0$, $\sigma = 1$, and $\mu = 0$. Note that symbols common to both distributions have different meanings.

The density graphs are obtained from fitting the distributions to the same sample data of arbitrarily generated numbers. The parameter α is related to the parameter ξ of the Pareto distribution, resulting in the tails of the density functions of α -stable random variables vanishing at a rate proportional to the Pareto tail.

The tails of the Pareto as well as the α -stable distribution decay at a rate with fixed power α , $x^{-\alpha}$ (i.e., *power law*), which is in contrast to the normal distribution whose tails decay at an exponential rate (i.e., roughly $e^{-x^2/2}$). We illustrate the effect focusing on the probability of exceeding some value x somewhere in the upper tail, say $x = 3$. Moreover, we choose the parameter of stability to be $\alpha = 1.5$. Under the normal law,

the probability of exceedance is roughly $e^{-3^2/2} = 0.011$ while under the power law it is about $3^{-1.5} = 0.1925$. Next, we let the benchmark x become gradually larger. Then the probability of assuming a value at least twice or four times as large (i.e., $2x$ or $4x$) is roughly

$$e^{-\frac{(2 \times 3)^2}{2}} \approx 0$$

or

$$e^{-\frac{(4 \times 3)^2}{2}} \approx 0$$

for the normal distribution. In contrast, under the power law, the same exceedance probabilities would be $(2 \times 3)^{-1.5} = 0.068$ or $(4 \times 3)^{-1.5} \approx 0.024$. This is a much slower rate than under the normal distribution. Note that the value of $x = 3$ plays no role for the power tails while the exceedance probability of the normal distribution decays faster the further out we are in the tails (i.e., the larger is x). The same reasoning applies to the lower tails considering the probability of falling below a benchmark x rather than exceeding it.

The parameter β indicates *skewness* where negative values represent left skewness while positive values indicate right skewness. The *scale* parameter σ has a similar interpretation as the standard deviation. Finally, the parameter μ indicates *location* of the distribution. Its interpretability depends on the parameter α . If the latter is between 1 and 2, then μ is equal to the mean.

Possible values of the parameters are listed below:

$$\begin{aligned} \alpha & (0, 2) \\ \beta & [-1, 1] \\ \sigma & (0, \infty) \\ \mu & \mathbb{R} \end{aligned}$$

Depending on the parameters α and β , the distribution has either support on the entire real line or only the part extending to the right of some location.

In general, the density function is not explicitly presentable. Instead, the distribution of the α -stable random variable is given by its charac-

teristic function. The characteristic function is given by

$$\begin{aligned} \varphi(t) &= \int_{-\infty}^{\infty} e^{itx} f(x) dx \\ &= \begin{cases} \exp \left\{ -\sigma^\alpha |t|^\alpha \left[1 - i\beta \text{sign}(t) \tan \frac{\pi\alpha}{2} \right] + i\mu t \right\}, & \alpha \neq 1 \\ \exp \left\{ -\sigma |t| \left[1 - i\beta \frac{2}{\pi} \text{sign}(t) \ln(t) \right] + i\mu t \right\}, & \alpha = 1 \end{cases} \quad (2) \end{aligned}$$

The density, then, has to be retrieved by an inverse transform to the characteristic function. Numerical procedures are employed for this task to approximate the necessary computations. The characteristic function (2) is presented here more for the sake of completeness rather than necessity. So, one should not be discouraged if it appears overwhelmingly complex.

In Figures 8 and 9, we present the density function for varying parameters β and α , respectively. Note in Figure 9 that for a $\beta = 1$, the density is positive only on a half-line toward the right as α approaches 2.

Only in the case of an α of 0.5, 1, or 2 can the functional form of the density be stated. For our purpose here, only the case $\alpha = 2$ is of interest because for this special case, the stable distribution represents the normal distribution. Then, the parameter β ceases to have any meaning since the normal distribution is not asymmetric.

A feature of the stable distributions is that moments such as the mean, for example, exist only up to the power α . (Recall that a moment exists when the according integral of the absolute values is finite.) So, except for the normal case (where $\alpha = 2$), there exists no finite variance. It becomes even more extreme when α is equal to 1 or less such that not even the mean exists anymore. The nonexistence of the variance is a major drawback when applying stable

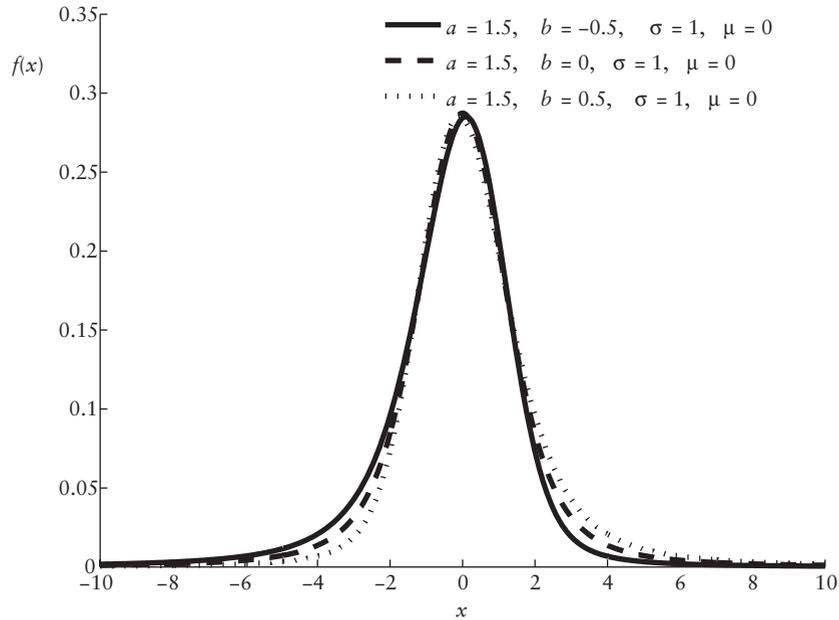


Figure 8 Stable Density Function for Various Values of β

distributions to financial data. This is one reason why the use of this family of distribution in finance is still contended.

This class of distributions owes its name to the *stability property* for the normal distribution

(Property 2): The weighted sum of an arbitrary number of α -stable random variables with the same parameters is, again, α -stable distributed. More formally, let X_1, \dots, X_n be identically distributed and independent of each other. Then,

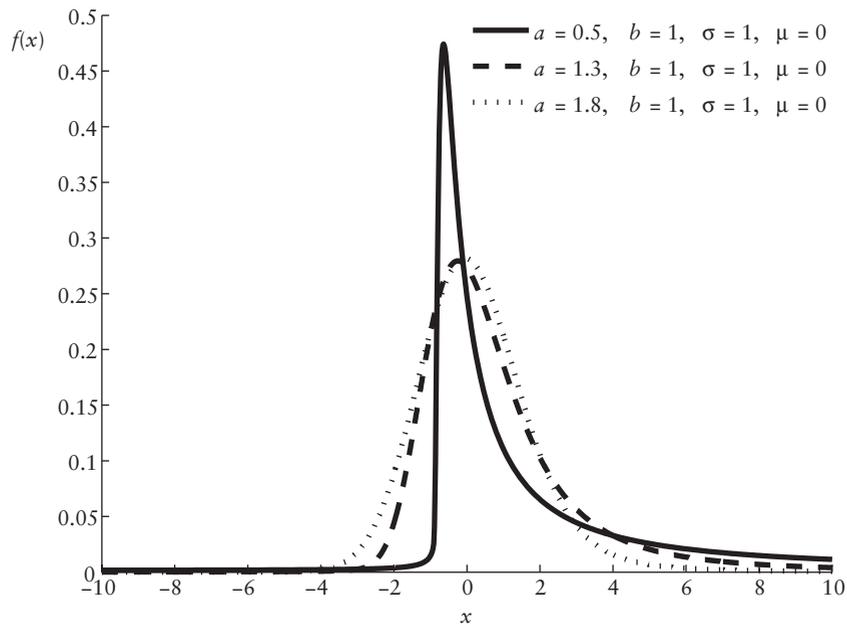


Figure 9 Stable Density Function (totally right-skewed) for Various Values of α

assume that for any $n \in \mathbb{N}$, there exists a positive constant a_n and a real constant b_n such that the normalized sum $Y(n)$

$$Y(n) = a_n(X_1 + X_2 + \cdots + X_n) + b_n \sim S(\alpha, \beta, \sigma, \mu) \quad (3)$$

converges in distribution to a random variable X , then this random variable X must be stable with some parameters α , β , σ , and μ . Again, recall that convergence in distribution means that the distribution function of $Y(n)$ in equation (3) converges to the distribution function on the right-hand side of equation (3).

In the context of financial returns, this means that monthly returns can be treated as the sum of weekly returns and, again, weekly returns themselves can be understood as the sum of daily returns. According to equation (3), they are equally distributed up to rescaling by the parameters a_n and b_n .

From the presentation of the normal distribution, we know that it serves as a limit distribution of a sum of identically distributed random variables that are independent and have finite variance. In particular, the sum converges in distribution to the standard normal distribution once the random variables have been summed and transformed appropriately. The prerequisite, however, was that the variance exists. Now, we can drop the requirement for finite variance and only ask for independence and identical distributions to arrive at the *generalized central limit theorem* expressed by equation (3). The sum of transformed random variables following rather arbitrary laws will have a distribution that follows a stable distribution law as the number n becomes very large. Thus, the class of α -stable distributions provides a greater set of limit distributions than the normal distribution containing the latter as a special case. Theoretically, this justifies the use of α -stable distributions as the choice for modeling asset returns when we consider the returns to be the resulting sum of many independent shocks.

Let us resume the previous example with the random dynamic and the related stock price evolution. Suppose, now, that the 10-day dynamic was S_α distributed. We denote the according random variable by V_{10} . We select a fairly moderate stable parameter of $\alpha = 1.8$. A value in this vicinity is commonly estimated for daily and even weekly stock returns. The skewness and location parameters are both set to zero, that is, $\beta = \mu = 0$. The scale is $\sigma = 1$, so that if the distribution was normal, that is, $\alpha = 2$, the variance would be 2 and, hence, consistent with the previous distributions. Note, however, that for $\alpha = 1.8$, the variance does not exist. Here the probability of the dynamic's exceedance of the lower threshold of 1 is

$$P(V_{10} > 1) = 0.2413 \quad (4)$$

compared to 0.2398 and 0.1870 in the normal and Student's t cases, respectively. Again, the probability in (4) corresponds to the event that in 10 days, the stock price will be greater than \$271. So, it is more likely than in the normal and Student's t model.

For the higher threshold of 3.5, we obtain

$$P(V_{10} > 3.5) = 0.0181$$

compared to 0.0067 and 0.0124 from the normal and Student's t cases, respectively. This event corresponds to a stock price beyond \$3,312, which is an immense increase. Under the normal distribution assumption, this event is virtually unlikely. It would happen in less than 1% of the 10-day periods. However, under the stable as well as the Student's t assumption, this could happen in 1.81% or 1.24% of the scenarios, which is three times or double the probability, respectively. Just for comparison, let us assume $\alpha = 1.6$, which is more common during a rough market climate. The dynamic would now exceed the threshold of 1 with probability

$$P(V_{10} > 1) = 0.2428$$

which fits in with the other distribution. For 3.5, we have

$$P(V_{10} > 3.5) = 0.0315 \quad (5)$$

which is equal to five times the probability under the normal distribution and almost three times the probability under the Student's t distribution assumption. For this threshold, the same probability as in equation (5) could only be achieved with a variance of $\sigma^2 = 4$, which would give the overall distribution a different shape. In the Student's t case, the degree of freedom parameter would have to be less than 3 such that now the variance would not exist any longer.

For the stable parameters chosen, the same results are obtained when the sign of the returns is negative and losses are considered. For example, $P(V_{10} < -3.5) = 0.0315$ corresponds to the probability of obtaining a stock price of \$3 or less. This scenario would only be given 0.67% probability in a normal distribution model. With respect to large portfolios such as those managed by large banks, negative returns deserve much more attention since losses of great magnitude result in widespread damages to industries beyond the financial industry.

As another example, let's look at what happened to the stock price of American International Group (AIG) in September 2008. On one single day, the stock lost 60% of its value. That corresponds to a return of about -0.94 . (Keep in mind that we are analyzing logarithmic returns.) If we choose a normal distribution with $\mu = 0$ and $\sigma^2 = 0.0012$ for the daily returns, a drop in price of this magnitude or less has near zero probability. The distributional parameters were chosen to best mimic the behavior of the AIG returns. By comparison, if we take an α -stable distribution with $\alpha = 1.6$, $\beta = 0$, $\mu = 0$, and $\sigma = 0.001$ where these parameters were selected to fit the AIG returns, we obtain the probability for a decline of at least this size of 0.00003, that is, 0.003%. So even with this distribution, an event of this impact is almost negligible. As a consequence, we have to choose a lower parameter α for the stable distribution. That brings to light the immense risk inherent in the return distributions when they are truly α -stable.

KEY POINTS

- Heavy tails are the general reference term for probability distributions whose probability mass in the tails (i.e., extreme parts of the distribution) is heavier than in the case of a normal distribution. Although there is no unique definition of the feature, there exists a selection of parameters that express whether a distribution is heavy-tailed with respect to the normal distribution. Financial asset returns commonly exhibit heavy tails, which imposes additional risk on asset managers that solely rely on theory based on the normal distribution and other candidates with appealing properties. Hence, it is necessary to account for heavy tails.
- Extreme value theory comprises a collection of distributions dealing with the most extreme values of some set. Either these distributions concentrate on the maxima and minima, respectively, or the most extreme values beyond thresholds. In general, this theory distinguishes among three different kinds of extreme value behavior. Financial risk theory has become intertwined with extreme value theory since it has become common knowledge that it does not suffice to base all analysis on the normal distribution alone.
- Stable distributions form a class of distributions capable of dealing with many stylized facts observed for asset returns. Moreover, the distributions from this class exhibit the property of stability under summation, roughly meaning that sums of random variables following certain probability laws are again distributed as individual random variables. This makes them appealing for the characterization of asset return behavior observed in the real world.
- Skewness is basically a measure of asymmetry of some distribution. While the normal distribution is symmetric about its mean, many other distributions do not share this feature. In fact, when analyzing asset returns, it is often revealed that they are noticeably

skewed to one side; that is, they are asymmetric. Consequently, it is important to consider skewness when dealing with asset returns in order to avoid additional risk arising from its neglect.

- The generalized central limit theorem is the extension of the central limit theorem stating that the appropriately scaled sum of certain random variables is eventually standard normally distributed when their number becomes large. However, the criteria for these random variables for the central limit theorem to hold are sometimes unrealistic. The generalized central limit theorem, in contrast, relaxes some of these criteria to include a larger selection of random variables that would fail to sum up to a standard normally distributed random variable. The lim-

iting distributions of these sums are instead members of the class of α -stable distributions. This theorem provides a justification for the use of stable distributions in mathematical finance.

REFERENCES

- De Haan, L., and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. New York: Springer.
- Embrechts, P., Klüppelberg, C., and Mikosch, T. (2003). *Modelling Extremal Events: For Insurance and Finance*. Berlin: Springer.
- Kotz, S., and Nadarajah, S. (2002). *Extreme Value Distributions: Theory and Applications*. London: Imperial College Press.
- Samorodnitsky, G., and Taqqu, M. (2000). *Stable Non-Gaussian Random Processes: Stochastic Models*. Boca Raton, FL: Chapman & Hall/CRC.

Stable and Tempered Stable Distributions

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair-Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

YOUNG SHIN KIM, PhD

Research Assistant Professor, School of Economics and Business Engineering, University of Karlsruhe and KIT

MICHELE LEONARDO BIANCHI, PhD

Research Analyst, Specialized Intermediaries Supervision Department, Bank of Italy

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: In financial models for asset pricing and asset allocation, asset returns and prices are assumed to follow a normal or Gaussian distribution. However, the properties of the normal distribution are not consistent with the observed behavior found for real-world asset returns. More specifically, the symmetric and rapidly decreasing tail properties of asset return distributions cannot describe the skewed and fat-tailed properties of the empirical distribution of asset returns. The alpha-stable distribution or α -stable distribution has been proposed as an alternative to the normal distribution for modeling asset returns because it allows for skewness and fat tails. Recent research since the turn of the century has introduced alternative distributions such as the tempered stable distributions to better describe asset returns.

In finance, the normal or Gaussian distribution has been the underlying assumption in describing asset returns in major financial theories such as the capital asset pricing theory and option pricing theory. In the early 1960s, Benoit Mandelbrot, a mathematician at

IBM's Thomas J. Watson Research Center, presented empirical evidence regarding returns on commodity prices and interest rate movements that strongly rejected the assumption that asset returns are normally distributed (see Mandelbrot, 1963). The mainstream financial

Dr. Bianchi acknowledges that the views expressed in this entry are his own and do not necessarily reflect those of the Bank of Italy.

models at the time relied on the work of Louis Bachelier, a French mathematician who at the beginning of the 20th century was the first to formulate random walk models for stock prices (see Bachelier, 1900). Bachelier's work assumed that relative price changes followed a normal distribution. Mandelbrot's findings led a leading financial economist, Paul Cootner of MIT, to warn the academic community that Mandelbrot's finding may mean that "past econometric work is meaningless" (see Cootner, 1964).

In Mandelbrot's attack on the normal distribution, he suggested that asset returns are more appropriately described by a non-normal *stable distribution* referred to as a stable Paretian distribution or alpha-stable distribution (α -stable distribution), so named because the tails of this distribution have Pareto power-type decay. The reason for describing this distribution as "non-normal stable" is because the normal distribution is a special case of the stable distribution. Because of the work by Paul Lévy, a French mathematician who introduced and characterized the non-normal stable distribution, this distribution is also referred to as the Lévy stable distribution and the Pareto-Lévy stable distribution.

There are two other facts about asset return distributions that have been supported by empirical evidence. First, distributions have been observed to be skewed or nonsymmetric. That is, unlike in the case of the normal distribution where there is a mirror imaging of the two sides of the probability distribution, typically in a skewed distribution one tail of the distribution is much longer (i.e., has greater probability of extreme values occurring) than the other tail of the probability distribution. Probability distributions with this attribute are referred to as having *fat tails* or *heavy tails*. The second finding is the tendency of large changes in asset prices (either positive or negative) to be followed by large changes, and small changes to be followed by small changes. This attribute of asset return distributions is referred to as *volatility clustering*. In contrast to the normal distribution, the

α -stable distribution allows for skewness and fat tails.

While the α -stable distribution has certain desirable properties that will be discussed in more detail in this entry, it is not suitable in certain modeling applications such as the modeling of option prices. In order to obtain a well-defined model for pricing options, the mean, variance, and exponential moments of the return distribution have to exist. For this reason, the smoothly truncated stable distribution and various types of tempered stable distributions have been proposed for financial modeling. Those distributions are obtained by tempering the tail properties of the α -stable distribution. Because they converge weakly to the α -stable distribution, the α -stable distribution is embedded in the class of the tempered stable distributions.

In this entry, we discuss the α -stable and tempered stable distributions. The more general distribution, named the infinitely divisible distribution, will be discussed as well. The distributions in this entry are defined by their characteristic functions. The density functions are not given by a closed-form formula in general but obtained by a numerical method discussed in Rachev et al. (2011).

α -STABLE DISTRIBUTION

In this section, we discuss a wide class of α -stable distributions. We review the definition and the basic properties of the α -stable distribution. We further present the class of smoothly truncated stable distributions which has been proposed by Menn and Rachev (2009) for dealing with the drawbacks of the α -stable distribution.

Definition of an α -Stable Random Variable

We begin with a definition of an α -stable random variable.¹ Suppose that X_1, X_2, \dots, X_n are independent and identically distributed (IID) random variables, independent copies of X .

Then a random variable X is said to follow an α -stable distribution if there exist a positive constant C_n and a real number D_n such that the following relation holds:

$$X_1 + X_2 + \dots + X_n \stackrel{d}{=} C_n X + D_n$$

The notation $\stackrel{d}{=}$ denotes equality in distribution. The constant $C_n = n^{\frac{1}{\alpha}}$ dictates the stability property, which we will discuss later. When $\alpha = 2$, we have the Gaussian (normal) case. In subsequent discussions of the α -stable distributions in this entry, we restrict ourselves to the non-Gaussian case in which $0 < \alpha < 2$.

For the general case, the density of the α -stable distribution does not have a closed-form solution. The distribution is expressed by its characteristic function:

$$\phi_{\text{stable}}(\mu; \alpha, \beta, \mu) = E[e^{iuX}] \begin{cases} \exp(i\mu u - |\sigma u|^\alpha (1 - i\beta(\text{sign } u) \tan \frac{\pi\alpha}{2})), & \alpha \neq 1 \\ \exp(i\mu u - \sigma |u| (1 - i\beta \frac{2}{\pi}(\text{sign } u) \ln |u|)), & \alpha = 1 \end{cases} \quad (1)$$

where

$$\text{sign } t = \begin{cases} 1, & t > 0 \\ 0, & t = 0 \\ -1, & t < 0 \end{cases}$$

The distribution is characterized by four parameters:

- α : the index of stability or the shape parameter, $\alpha \in (0, 2)$.
- β : the skewness parameter, $\beta \in [-1, +1]$.
- σ : the scale parameter, $\sigma \in (0, +\infty)$.
- μ : the location parameter, $\mu \in (-\infty, +\infty)$.

When a random variable X follows the α -stable distribution characterized by those parameters, then we denote it by $X \sim S_\alpha(\sigma, \beta, \mu)$.

The three special cases where there is a closed-form solution for the densities are (1) the Gaussian case ($\alpha = 2$), (2) the Cauchy case ($\alpha = 1, \beta = 0$), and (3) the Lévy case ($\alpha = 1/2, \beta = \pm 1$) with the following respective densities:

- Gaussian: $f(x) = \frac{1}{2\sigma\sqrt{\pi}} e^{-\frac{(x-\mu)^2}{4\sigma^2}}, -\infty < x < \infty$
- Cauchy: $f(x) = \frac{\sigma}{\pi((x-\mu)^2 + \sigma^2)}, -\infty < x < \infty$
- Lévy: $f(x) = \frac{\sqrt{\sigma}}{\sqrt{2\pi}(x-\mu)^{3/2}} e^{-\frac{\sigma}{2(x-\mu)}}, \mu < x < \infty$

Because of the four parameters, the α -stable distribution is highly flexible and suitable for modeling nonsymmetric, highly kurtotic, and heavy-tailed data. Figures 1 and 2 illustrate the effects of the shape and skewness parameters, respectively, on the shape of the distribution, with other parameters kept constant. As is evident from Figure 1, a lower value for α is attributed to heavier tails and higher kurtosis.

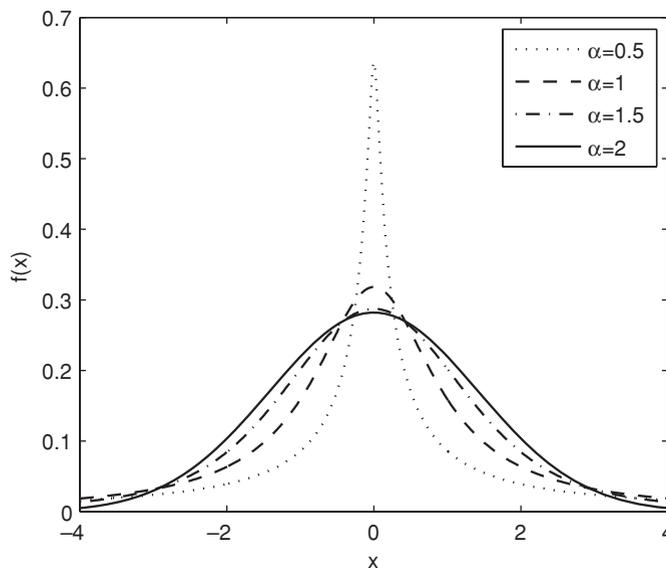


Figure 1 Illustration of α -Stable Densities for Varying α 's, with $\beta = 0, \sigma = 1$, and $\mu = 0$

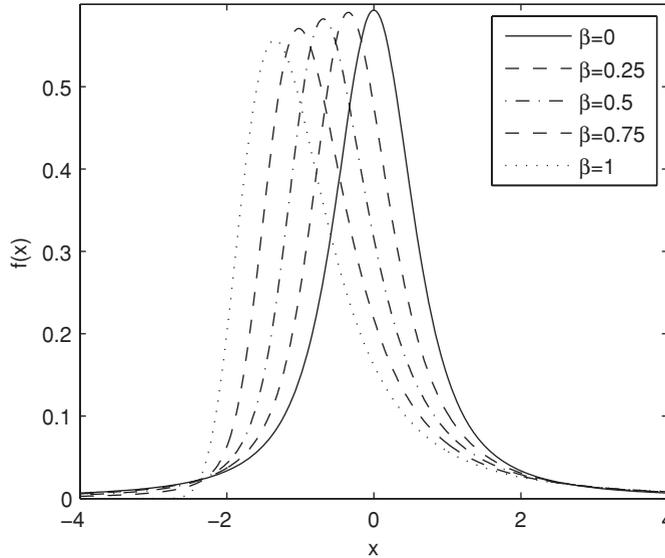


Figure 2 Illustration of α -stable Densities for Varying β 's, with $\alpha = 1.25, \sigma = 0.5,$ and $\mu = 0$

Useful Properties of an α -Stable Random Variable

The four basic properties of the α -stable distribution:

- *Property 1.* The power tail decay property means that the tail of the density function decays like a power function (slower than the exponential decay), which is what allows the distribution to capture extreme events in the tails:

$$P(|X| > x) \propto C \cdot x^{-\alpha}, x \rightarrow \infty$$

for some constant C . More precisely, if $X \sim S_\alpha(\sigma, \beta, \mu)$ with $0 < \alpha < 2$ then

$$\begin{cases} \lim_{\lambda \rightarrow \infty} \lambda^\alpha P(X > \lambda) = C_\alpha \frac{1+\beta}{2} \sigma^\alpha \\ \lim_{\lambda \rightarrow \infty} \lambda^\alpha P(X > -\lambda) = C_\alpha \frac{1-\beta}{2} \sigma^\alpha \end{cases}$$

where

$$C_\alpha = \begin{cases} \frac{1-\alpha}{\Gamma(2-\alpha)\cos(\pi\alpha/2)} & \text{if } \alpha \neq 1 \\ \frac{2}{\pi} & \text{if } \alpha = 1 \end{cases}$$

- *Property 2.* Raw moments satisfy the property:

$$\begin{aligned} E|X|^p &< \infty \text{ for any } 0 < p < \alpha \\ E|X|^p &= \infty \text{ for any } p \geq \alpha \end{aligned}$$

- *Property 3.* Because of Property 2, the mean is finite only for $\alpha > 1$:

$$\begin{aligned} E[X] &= \mu \quad \text{for } \alpha > 1 \\ E[X] &= \infty \quad \text{for } 0 < \alpha \leq 1 \end{aligned}$$

The second and higher moments are infinite, leading to infinite variance together with the skewness and kurtosis coefficients.

- *Property 4.* The stability property is a useful and convenient property and dictates that the distributional form of the variable is preserved under linear transformations. The stability property is governed by the stability parameter α in the constant C_n (which appeared earlier in the definition of an α -stable random variable): $C_n = n^{1/\alpha}$. As was stated earlier, smaller values of α refer to a heavier-tailed distribution. The standard central limit theorem does not apply to the non-Gaussian case: An appropriately standardized large sum of IID random variables converges to an α -stable random variable instead of a normal random variable.

The following examples illustrate the stability property. Suppose that X_1, X_2, \dots, X_n are IID

random variables with $X_i \sim S_\alpha(\sigma_i, \beta_i, \mu_i)$, $i = 1, 2, \dots, n$ and a fixed α . Then:

- The distribution of $Y = \sum_i^n X_i$ is α -stable with the index of stability α and parameters:

$$\beta = \frac{\sum_i^n \beta_i \sigma_i^\alpha}{\sum_i^n \sigma_i^\alpha}, \sigma = \left(\sum_i^n \sigma_i^\alpha \right)^{1/\alpha}, \mu = \sum_i^n \mu_i$$

- The distribution of $Y = X_1 + a$ for some real constant a is α -stable with the index of stability α and parameters:

$$\beta = \beta_1, \sigma = \sigma_1, \mu = \mu_1 + a$$

- The distribution of $Y = aX_1$ for some real constant $a(a \neq 0)$ is α -stable with the index of stability α and parameters:

$$\begin{aligned} \beta &= (\text{sign } a)\beta_1 \\ \sigma &= |a| \sigma_1 \\ \mu &= \begin{cases} a\mu_1 & \text{for } \alpha \neq 1 \\ a\mu_1 - \frac{2}{\pi}a(\ln a)\sigma_1\beta_1 & \text{for } \alpha = 1 \end{cases} \end{aligned}$$

- The distribution of $Y = -X_1$ is α -stable with the index of stability α and parameters:

$$\beta = -\beta_1, \sigma = \sigma_1, \mu = \mu_1$$

Smoothly Truncated Stable Distribution

In some special cases of financial modeling it might occur that the infinite variance of stable distributions make their application impossible. In many cases, the infinite variance of the return might lead to an infinite price for derivative instruments such as options, clearly contradicting reality and intuition. The modeler is confronted with a dilemma. On the one hand, the skewed and heavy-tailed return distribution disqualifies the normal distribution as a suitable candidate; on the other hand, theoretical restrictions in option pricing do not allow the application of the stable distribution due to its infinite moments of order higher than α . For this reason, Menn and Rachev (2009) have sug-

gested the use of appropriately truncated stable distributions.

The exact definition of truncated stable distributions is not that important at this point; that is why we restrict ourselves to a brief description of the idea. The density function of a smoothly truncated stable distribution (STS distribution) is obtained by replacing the heavy tails of the density function g of some stable distribution with parameters $(\alpha, \beta, \sigma, \mu)$ by the thin tails of two appropriately chosen normal distributions h_1 and h_2 :

$$f(x) = \begin{cases} h_1(x), & x < a \\ g(x), & a \leq x \leq b \\ h_2(x), & x > b \end{cases}$$

The parameters of the normal distributions are chosen such that the resulting function is the continuous density function of a probability measure on the real line. If it is possible to choose the cutting points a and b in a way that the resulting distribution possesses zero mean and unit variance, then we have found an easy way to characterize standardized STS distributions. In Figure 3, the influence of the stable parameters on the appropriate cutting points is examined. As α approaches 2 (i.e., when the stable distribution approaches the normal distribution), we observe that the cutting points move to infinity. For small values of α , in contrast, the interval $[a, b]$ shrinks, reflecting the increasing heaviness of the tails of the stable distribution in the center.

Due to the thin tails of the normal density functions, the STS distributions admit finite moments of arbitrary order but nevertheless are able to explain extreme observations. Table 1 provides a comparison of tail probabilities for an arbitrarily chosen STS distribution with zero mean and unit variance and the standard normal distribution. As can be seen from the table, the probability of extreme events is much higher under the assumption of an STS distribution. STS distributions allow for skewness in the returns. Moreover, the tails behave like fat

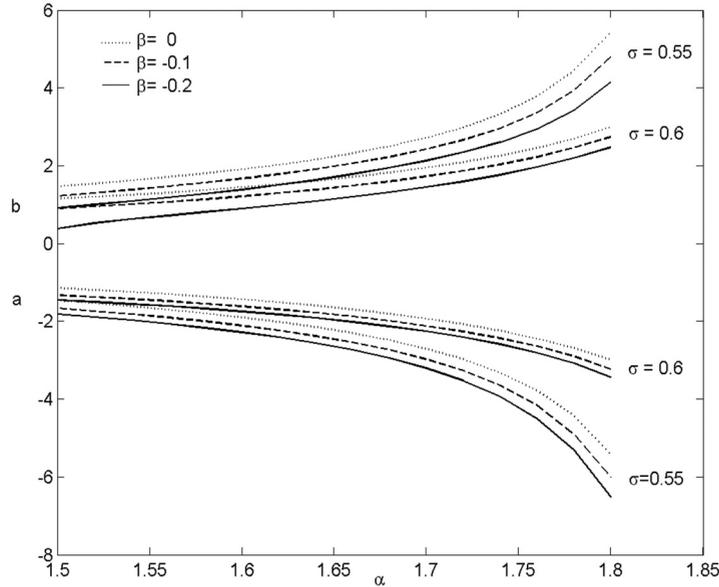


Figure 3 Influence of the Stable Parameters on the Cutting Points a and b

Table 1 Comparison of Tail Probabilities for a Standard Normal and a Standardized STS Distribution

x	$\mathbb{P}(X_1 \leq x)$ with $X_1 \sim N(0,1)$	$\mathbb{P}(X_2 \leq x)$ with $X_2 \sim STS$
-1	15.866%	11.794%
-2	2.275%	2.014%
-3	0.135%	0.670%
-4	0.003%	0.356%
-5	$\approx 10^{-5}\%$	0.210%
-6	$\approx 10^{-8}\%$	0.120%
-7	$\approx 10^{-10}\%$	0.067%
-8	$\approx 10^{-14}\%$	0.036%
-9	$\approx 10^{-17}\%$	0.019%
-10	$\approx 10^{-22}\%$	0.010%

tails but are light tails in the mathematical sense. Hence, all moments of arbitrary order exist and are finite. For this reason, advocates of the class of STS distribution argue that it is an appropriate class for modeling the return distribution of various financial assets.

TEMPERED STABLE DISTRIBUTIONS

In this section, we discuss six types of *tempered stable distributions*.

Classical Tempered Stable Distribution

Let $\alpha \in (0,1) \cup (1, 2)$, $C, \lambda_+, \lambda_- > 0$, and $m \in \mathbb{R}$. X is said to follow the *classical tempered stable (CTS) distribution* if the characteristic function of X is given by

$$\begin{aligned} \phi_X(u) &= \phi_{CTS}(u; \alpha, C, \lambda_+, \lambda_-, m) \\ &= \exp(ium - iuC\Gamma(1 - \alpha)(\lambda_+^{\alpha-1} - \lambda_-^{\alpha-1}) \\ &\quad + C\Gamma(-\alpha)((\lambda_+ - iu)^\alpha - \lambda_+^\alpha \\ &\quad + (\lambda_- + iu)^\alpha - \lambda_-^\alpha)) \end{aligned} \tag{2}$$

and we denote it by $X \sim CTS(\alpha, C, \lambda_+, \lambda_-, m)$.

Using the n th derivative of $\psi(u) = \log\phi_X(u)$ evaluated around zero, the cumulants $c_n(X) = \frac{1}{i^n} \frac{\partial^n \psi}{\partial u^n}(0)$ of X are obtained by

$$\begin{aligned} c_1(X) &= m \\ c_n(X) &= C\Gamma(n - \alpha)(\lambda_+^{\alpha-n} \\ &\quad + (-1)^n \lambda_-^{\alpha-n}), \text{ for } n = 2, 3, \dots \end{aligned}$$

The role of the parameters is as follows:

- The parameter m determines the location of the distribution.

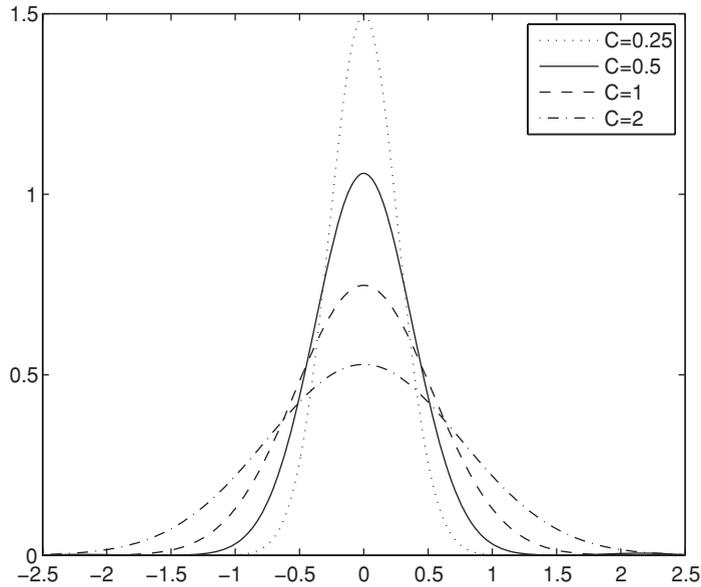


Figure 4 Probability Density of the CTS Distributions' Dependence on C
 Note: $C \in \{0.25, 0.5, 1, 2\}$, $\alpha = 1.4$, $\lambda_+ = 50$, $\lambda_- = 50$, $m = 0$.

- The parameter C is the scale parameter. Figure 4 shows the density function of the CTS distributions' dependence on C .
- The parameters λ_+ and λ_- control the rate of decay on the positive and negative tails,

respectively. If $\lambda_+ > \lambda_-$ ($\lambda_+ < \lambda_-$), then the distribution is skewed to the left (right), and if $\lambda_+ = \lambda_-$, then it is symmetric. Figure 5 illustrates left and right skewed density functions of the CTS distribution, as well as the symmetric case.

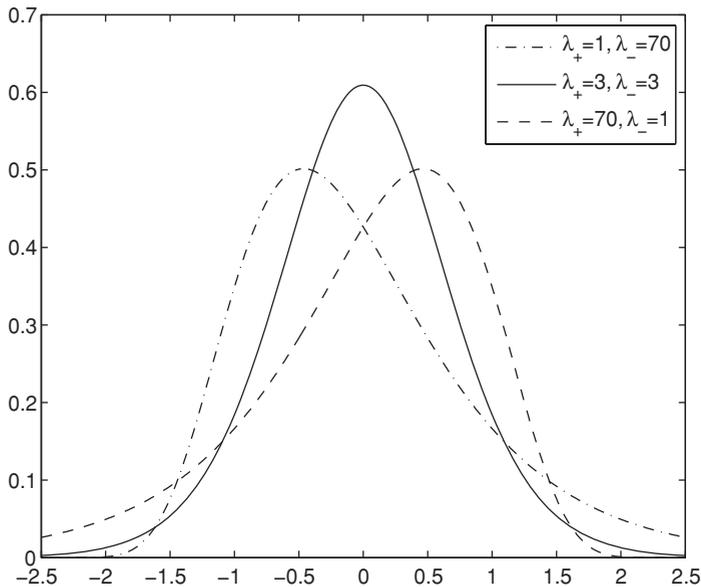


Figure 5 Probability Density of the CTS Distributions: Dependence on λ_+ and λ_-
 Note: $(\lambda_+, \lambda_-) \in \{(1, 70), (3, 3), (70, 1)\}$, $\alpha = 0.8$, $C = 1$, $m = 0$.

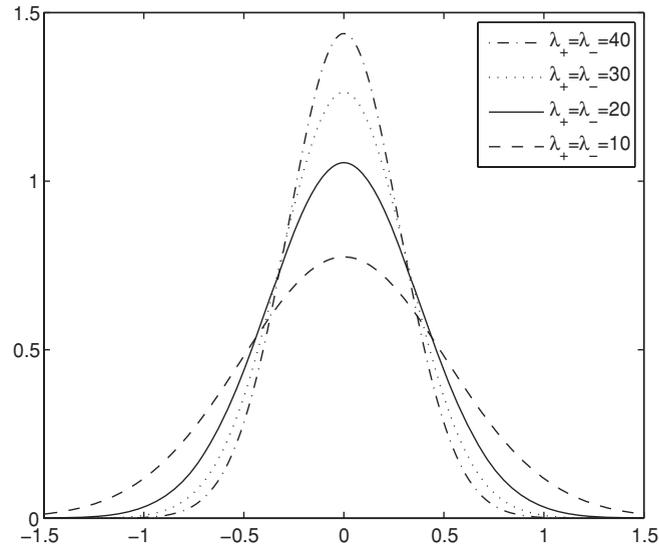


Figure 6 Probability Density of the Symmetric CTS Distributions' Dependence on Parameters λ_+, λ_-
Note: $\lambda_+ = \lambda_- \in \{10, 20, 30, 40\}$, $\alpha = 1.1$, $C = 1$, $m = 0$.

- The parameters λ_+, λ_- , and α are related to tail weights. Figures 6 and 7 illustrate this fact. We will discuss another role of α later.
- If α approaches to 0, the CTS distribution converges to the variance-gamma distribution (discussed later in this entry) in distribution sense.

If we take a special parameter C defined by

$$C = (\Gamma(2 - \alpha)(\lambda_+^{\alpha-2} + \lambda_-^{\alpha-2}))^{-1} \quad (3)$$

then $X \sim \text{CTS}(\alpha, C, \lambda_+, \lambda_-, 0)$ has zero mean and unit variance. In this case, X is called the standard CTS distribution with parameters $(\alpha, \lambda_+, \lambda_-)$ and denoted by $X \sim \text{stdCTS}(\alpha, \lambda_+, \lambda_-)$.

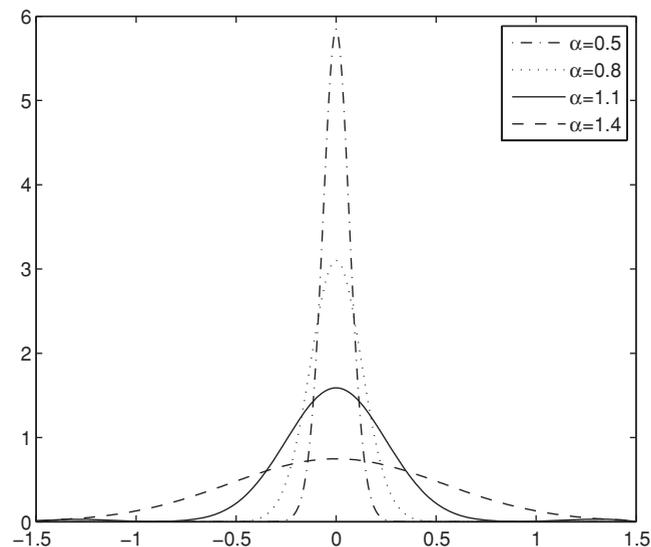


Figure 7 Probability Density of the CTS Distributions: Dependence on α
Note: $\alpha \in \{0.5, 0.8, 1.1, 1.4\}$, $C = 1$, $\lambda_+ = 50$, $\lambda_- = 50$, $m = 0$.

λ_-). Let m be a real number, σ be a positive real number, and $X \sim \text{stdCTS}(\alpha, \lambda_+, \lambda_-)$. Then

$$Y = \sigma X + m \approx \text{CTS} \left(\alpha, \frac{\sigma^\alpha}{\Gamma(2-\alpha)(\lambda_+^{\alpha-2} + \lambda_-^{\alpha-2})}, \frac{\lambda_+}{\sigma}, \frac{\lambda_-}{\sigma}, m \right)$$

The random variable Y is the CTS distributed, and its mean and variance are m and σ^2 , respectively.

Generalized Classical Tempered Stable Distribution

A more general form of the characteristic function for the CTS distribution is

$$\begin{aligned} \phi_X(u) = & \exp(ium - iu\Gamma(1-\alpha)(C + \lambda_+^{\alpha+1} - C_- \lambda_-^{\alpha-1}) \\ & + C_+ \Gamma(-\alpha_+)((\lambda_+ - iu)^{\alpha_+} - \lambda_+^{\alpha_+}) \\ & + C_- \Gamma(-\alpha_-)((\lambda_- + iu)^{\alpha_-} - \lambda_-^{\alpha_-})) \end{aligned} \quad (4)$$

where $\alpha_+, \alpha_- \in (0, 1) \cup (1, 2)$, $C_+, C_-, \lambda_+, \lambda_- > 0$, and $m \in \mathbb{R}$. This distribution has been referred to as the *generalized classical tempered stable (GTS) distribution* and we denote it by $X \sim \text{GTS}(\alpha_+, \alpha_-, C_+, C_-, \lambda_+, \lambda_-, m)$.²

The cumulants of X are $c_1(X) = m$ and

$$\begin{aligned} c_n(X) = & C_+ \Gamma(n - \alpha_+) \lambda_+^{\alpha_+ - n} \\ & + (-1)^n C_- \Gamma(n - \alpha_-) \lambda_-^{\alpha_- - n} \end{aligned}$$

for $n = 2, 3, \dots$. If we substitute

$$C_+ = \frac{p\lambda_+^{2-\alpha_+}}{\Gamma(2-\alpha_+)}, \quad C_- = \frac{(1-p)\lambda_-^{2-\alpha_-}}{\Gamma(2-\alpha_-)} \quad (5)$$

where $p \in (0, 1)$, then $X \sim \text{GTS}(\alpha_+, \alpha_-, C_+, C_-, \lambda_+, \lambda_-, 0)$ has zero mean and unit variance. In this case, X is called the standard GTS distribution with parameters $(\alpha_+, \alpha_-, \lambda_+, \lambda_-, p)$ and denoted by $X \sim \text{stdGTS}(\alpha_+, \alpha_-, \lambda_+, \lambda_-, p)$.

Modified Tempered Stable Distribution

Let $\alpha \in (0, 1) \cup (1, 2)$, $C, \lambda_+, \lambda_- > 0$, and $m \in \mathbb{R}$. X is said to follow the *modified tempered stable*

(MTS) *distribution* (see Kim et al., 2009) if the characteristic function of X is given by

$$\begin{aligned} \phi_X(u) = & \phi_{\text{MTS}}(u; \alpha, C, \lambda_+, \lambda_-, m) \\ = & \exp(ium + C(G_R(u; \alpha, \lambda_+) + G_R(u; \alpha, \lambda_-)) \\ & + iuC(G_I(u; \alpha, \lambda_+) - G_I(u; \alpha, \lambda_-))) \end{aligned} \quad (6)$$

where for $u \in \mathbb{R}$,

$$G_R(x; \alpha, \lambda) = 2^{-\frac{\alpha+3}{2}} \sqrt{\pi} \Gamma\left(-\frac{\alpha}{2}\right) ((\lambda^2 + x^2)^{\frac{\alpha}{2}} - \lambda^\alpha)$$

and

$$\begin{aligned} G_I(x; \alpha, \lambda) = & 2^{-\frac{\alpha+1}{2}} \Gamma\left(\frac{1-\alpha}{2}\right) \lambda^{\alpha-1} \\ & \times \left[{}_2F_1\left(1, \frac{1-\alpha}{2}; \frac{3}{2}; -\frac{x^2}{\lambda^2}\right) - 1 \right] \end{aligned}$$

where ${}_2F_1$ is the hypergeometric function. We denote an MTS distributed random variable X by $X \sim \text{MTS}(\alpha, C, \lambda_+, \lambda_-, m)$.

The role of the parameters of the MTS distribution is same as in the case of the CTS distribution. For example, the parameters λ_+ and λ_- control the rate of decay on the positive and negative tails, respectively, and if $\lambda_+ = \lambda_-$, then it is symmetric. The characteristic function of the symmetric MTS distribution is defined not only for the case $\alpha \in (0, 1) \cup (1, 2)$ but also for the case $\alpha = 1$. The form of the characteristic function for the symmetric case is given by

$$\begin{aligned} \phi_X(u) = & \phi_{\text{MTS}}(u; \alpha, C, \lambda, \lambda, m) \\ = & \exp\left(ium + C2^{-\frac{\alpha+1}{2}} \sqrt{\pi} \Gamma\left(-\frac{\alpha}{2}\right) \right. \\ & \left. \times ((\lambda^2 + x^2)^{\frac{\alpha}{2}} - \lambda^\alpha)\right) \end{aligned}$$

The mean of X is m , and the cumulants of X are equal to

$$\begin{aligned} c_n(X) = & 2^{n-\frac{\alpha+3}{2}} C \Gamma\left(\frac{n+1}{2}\right) \Gamma\left(\frac{n-\alpha}{2}\right) \\ & \times (\lambda_+^{\alpha-n} + (-1)^n \lambda_-^{\alpha-n}) \end{aligned}$$

for $n = 2, 3, \dots$.

If we substitute

$$C = 2^{\frac{\alpha+1}{2}} \left(\sqrt{\pi} \Gamma\left(1 - \frac{\alpha}{2}\right) (\lambda_+^{\alpha-2} + \lambda_-^{\alpha-2}) \right)^{-1} \quad (7)$$

then $X \sim \text{MTS}(\alpha, C, \lambda_+, \lambda_-, 0)$ has zero mean and unit variance. In this case, the random variable X is called the standard MTS distribution and

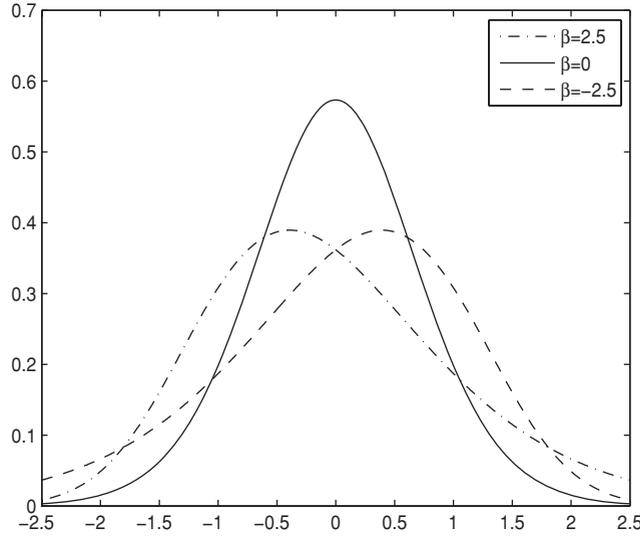


Figure 8 Probability Density of the NTS Distributions' Dependence on β
 Note: $\beta \in \{-2.5, 0, 2.5\}$, $\alpha = 0.8$, $C = 1$, $\lambda = 4$, $m = 0$.

denoted by $X \sim \text{stdMTS}(\alpha, \lambda_+, \lambda_-)$. Let m be a real number, σ be a positive real number, and $X \sim \text{stdMTS}(\alpha, \lambda_+, \lambda_-)$. Then

$$Y = \sigma X + m \sim \text{MTS}(\alpha, \sigma^\alpha C, \lambda_+/\sigma, \lambda_-/\sigma, m)$$

where C is equal to (7). The random variable Y is MTS distributed, and its mean and variance are m and σ^2 , respectively.

Normal Tempered Stable Distribution

Let $\alpha \in (0, 2)$, $C, \lambda > 0$, $|\beta| < \lambda$, and $m \in \mathbb{R}$. X is said to follow the *normal tempered stable (NTS) distribution*.³ If the characteristic function of X is given by

$$\begin{aligned} \phi_X(u) &= \phi_{NTS}(u; \alpha, C, \lambda, \beta, m) \\ &= \exp(ium - iu2^{-\frac{\alpha-1}{2}} \sqrt{\pi} C \Gamma\left(1 - \frac{\alpha}{2}\right) \\ &\quad \times \beta(\lambda^2 - \beta^2)^{\frac{\alpha}{2}-1} + 2^{-\frac{\alpha+1}{2}} C \sqrt{\pi} \Gamma\left(-\frac{\alpha}{2}\right) \\ &\quad \times \left(\left(\lambda^2 - (\beta + iu)^2\right)^{\frac{\alpha}{2}} - (\lambda^2 - \beta^2)^{\frac{\alpha}{2}}\right) \end{aligned} \quad (8)$$

We denote an NTS distributed random variable X by $X \sim \text{NTS}(\alpha, C, \lambda, \beta, m)$.

The mean of X is m . The general expressions for cumulants of X are omitted since they are

rather complicated. Instead of the general form, we present three cumulants

$$\begin{aligned} c_2(X) &= \bar{C}(\lambda^2 - \beta^2)^{\frac{\alpha}{2}-2} \alpha(\alpha\beta^2 - \lambda^2 - \beta^2) \\ c_3(X) &= -\bar{C} \alpha \beta (\lambda^2 - \beta^2)^{\frac{\alpha}{2}-3} (\alpha^2 \beta^2 - 3\alpha\lambda^2 - 3\alpha\beta^2 + 6\lambda^2 + 2\beta^2) \\ c_4(X) &= \bar{C} \alpha (\alpha - 2) (\lambda^2 - \beta^2)^{\frac{\alpha}{2}-4} \\ &\quad \times (\alpha^2 \beta^4 - 6\alpha\lambda^2 \beta^2 - 4\alpha\beta^4 + 3\beta^4 + 18\lambda^2 \beta^2 + 3\lambda^4) \end{aligned}$$

where $\bar{C} = 2^{-\frac{\alpha+1}{2}} C \sqrt{\pi} \Gamma\left(-\frac{\alpha}{2}\right)$

The roles of parameters α, C , and λ are same as in the case of the symmetric MTS distribution. The parameter β is related to the distribution's skewness. If $\beta < 0$ ($\beta > 0$), then the distribution is skewed to the left (right). Moreover, if $\beta = 0$, then it is symmetric. This fact is illustrated in Figure 8.

If we substitute

$$\begin{aligned} C &= 2^{\frac{\alpha+1}{2}} \\ &\quad \times \left(\sqrt{\pi} \Gamma\left(-\frac{\alpha}{2}\right) \alpha (\lambda^2 - \beta^2)^{\frac{\alpha}{2}-2} (\alpha\beta^2 - \lambda^2 - \beta^2)\right)^{-1} \end{aligned} \quad (9)$$

then $X \sim \text{NTS}(\alpha, C, \lambda, \beta, 0)$ has zero mean and unit variance. In this case, X is called the standard NTS distribution and denoted by $X \sim \text{stdNTS}(\alpha, \lambda, \beta)$. Let m be a real number, σ be a

positive real number, and $X \sim \text{stdNTS}(\alpha, \lambda, \beta)$. Then

$$Y = \sigma X + m \sim \text{NTS}(\alpha, \sigma^\alpha C, \lambda/\sigma, \beta/\sigma, m)$$

where C is equal to (9). The random variable Y is NTS distributed, and its mean and variance are m and σ^2 , respectively.

If we substitute $\alpha = 1$ and $C = \frac{c}{\pi}$ into the definition of the NTS distribution, we obtain the normal inverse Gaussian (NIG) distribution.⁴ That is, if random variable $X \sim \text{NTS}(1, c/\pi, \lambda, \beta, m)$, then X becomes an NIG distributed random variable. In this case, we denote $X \sim \text{NIG}(c, \lambda, \beta, m)$.

By substituting $\alpha = 1$ and $C = \frac{c}{\pi}$ into (8), we obtain the characteristic function of the NIG distributed X as

$$\begin{aligned} \phi_X(u) &= \phi_{\text{NIG}}(u; c, \lambda, \beta, m) \\ &= \exp\left(ium - \frac{iuc\beta}{\sqrt{\lambda^2 - \beta^2}} - c \right. \\ &\quad \left. \times \left(\sqrt{\lambda^2 - (\beta + iu)^2} - \sqrt{\lambda^2 - \beta^2}\right)\right) \end{aligned} \quad (10)$$

If we substitute

$$c = \frac{(\lambda^2 - \beta^2)^{\frac{3}{2}}}{\lambda^2} \quad (11)$$

then $X \sim \text{NIG}(c, \lambda, \beta, 0)$ has zero mean and unit variance. In this case, X is called the standard NIG distribution and denoted by $X \sim \text{stdNIG}(\lambda, \beta)$.

Kim-Rachev Tempered Stable Distribution

Let $\alpha \in (0, 1) \cup (1, 2)$, $k_+, k_-, r_+, r_- > 0$, $p_+, p_- \in \{p > -\alpha \mid p \neq -1, p \neq 0\}$, and $m \in \mathbb{R}$. X is said to follow the *Kim-Rachev tempered stable* (KRTS) distribution (see Kim et al., 2008b) if the characteristic function of X is given by

$$\begin{aligned} \phi_X(u) &= \phi_{\text{KRTS}}(u; \alpha, k_+, k_-, r_+, r_-, p_+, p_-, m) \\ &= \exp\left(ium - iu\Gamma(1 - \alpha) \left(\frac{k_+ r_+}{p_+ + 1} - \frac{k_- r_-}{p_- + 1}\right) \right. \\ &\quad \left. + k_+ H(iu; \alpha, r_+, p_+) + k_- H(-iu; \alpha, r_-, p_-)\right) \end{aligned} \quad (12)$$

where

$$H(x; \alpha, r, p) = \frac{\Gamma(-\alpha)}{p} ({}_2F_1(p, -\alpha; 1 + p; rx) - 1)$$

We denote a KRTS distributed random variable X by $X \sim \text{KRTS}(\alpha, k_+, k_-, r_+, r_-, p_+, p_-, m)$.

The KRTS distribution is an extension of the CTS distribution. Indeed, the distribution $\text{KRTS}(\alpha, k_+, k_-, r_+, r_-, p_+, p_-, m)$ converges weakly to the CTS distribution as $p_\pm \rightarrow \infty$ provided that $C_\pm = c(\alpha + p_\pm)r_\pm^{-\alpha}$ where $c > 0$ (see Kim et al., 2008a). Figure 9 shows that the KRTS distribution converges to the CTS distribution when parameter $p = p_+ = p_-$ increases to infinity.

The cumulants of the KRTS distributed random variable X are $c_1(X) = m$ and

$$\begin{aligned} c_n(X) &= \Gamma(n - \alpha) \left(\frac{k_+ r_+^n}{p_+ + n} + (-1)^n \frac{k_- r_-^n}{p_- + n} \right), \\ &\quad \text{for } n = 2, 3, \dots \end{aligned}$$

If we substitute

$$\begin{aligned} k_+ &= C \frac{\alpha + p_+}{r_+^\alpha} \\ k_- &= C \frac{\alpha + p_-}{r_-^\alpha} \end{aligned}$$

where

$$C = \frac{1}{\Gamma(2 - \alpha)} \left(\frac{\alpha + p_+}{2 + p_+} r_+^{2-\alpha} + \frac{\alpha + p_-}{2 + p_-} r_-^{2-\alpha} \right)^{-1} \quad (13)$$

then $X \sim \text{KRTS}(\alpha, k_+, k_-, r_+, r_-, p_+, p_-, 0)$ has zero mean and unit variance. In this case, X is said to be standard KRTS distributed and denoted by $X \sim \text{stdKRTS}(\alpha, r_+, r_-, p_+, p_-)$. Let m be a real number, σ be a positive real number, and $X \sim \text{stdKRTS}(\alpha, r_+, r_-, p_+, p_-)$. Then

$$\begin{aligned} Y &= \sigma X + m \\ &\sim \text{KRTS}(\alpha, C(\alpha + p_+)(\sigma r_+)^{-\alpha}, C(\alpha + p_-) \\ &\quad (\sigma r_-)^{-\alpha}, \sigma r_+, \sigma r_-, p_+, p_-, m) \end{aligned}$$

where C is equal to (13). The random variable Y is KRTS distributed, and its mean and variance are m and σ^2 , respectively.

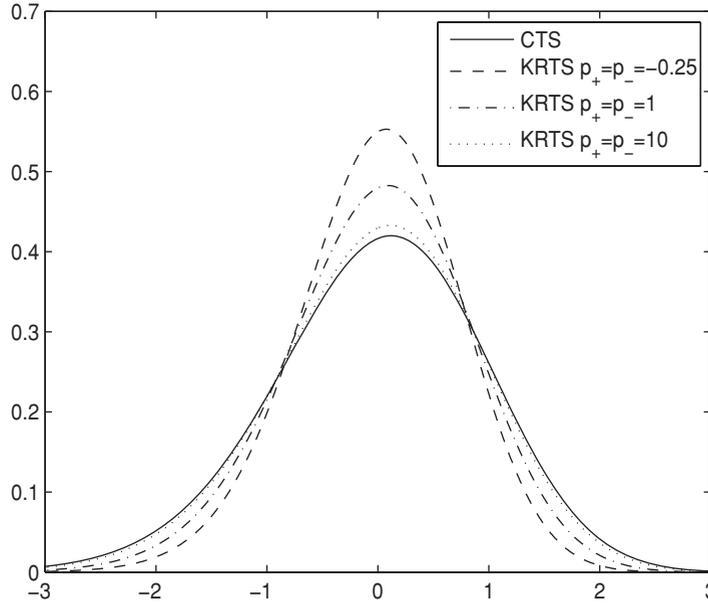


Figure 9 Probability Density of the CTS Distribution with Parameters $C = 1, \lambda_+ = 10, \lambda_- = 2, \alpha = 1.25$, and the KRTS Distributions with $k_{\pm} = C(\alpha + p)_{\pm}^{\alpha}, r_+ = 1/\lambda_+, r_- = 1/\lambda_-$, where $p = p_+ = p_- \in \{-0.25, 1, 10\}$

Rapidly Decreasing Tempered Stable Distribution

Let $\alpha \in (0, 1) \cup (1, 2), C, \lambda_+, \lambda_- > 0$, and $m \in \mathbb{R}$. A random variable X is said to follow the *rapidly decreasing tempered stable* (RDTS) distribution (see Bianchi et al., 2010 and Kim et al., 2010) if the characteristic function of X is given by

$$\phi_X(u) = \phi_{RDTS}(u; \alpha, C, \lambda_+, \lambda_-, m) \exp(ium + C(G(iu; \alpha, \lambda_+) + G(-iu; \alpha, \lambda_-))) \tag{14}$$

where

$$G(x; \alpha, \lambda) = 2^{-\frac{\alpha}{2}-1} \lambda^{\alpha} \Gamma\left(-\frac{\alpha}{2}\right) \left(M\left(-\frac{\alpha}{2}, \frac{1}{2}; \frac{x^2}{2\lambda^2}\right) - 1\right) + 2^{-\frac{\alpha}{2}-\frac{1}{2}} \lambda^{\alpha-1} x \Gamma\left(\frac{1-\alpha}{2}\right) \times \left(M\left(\frac{1-\alpha}{2}, \frac{3}{2}; \frac{x^2}{2\lambda^2}\right) - 1\right)$$

and M is the confluent hypergeometric function. Further details of the confluent hypergeometric function are presented at the end of this entry. In this case, we denote $X \sim RDTS(\alpha, C, \lambda_+, \lambda_-, m)$. The role of the parameters are the same as for the case of the CTS distribution.

The mean of X is m , and the cumulants of X are

$$c_n(X) = 2^{\frac{n-\alpha-2}{2}} C \Gamma\left(\frac{n-\alpha}{2}\right) \times (\lambda_+^{\alpha-n} + (-1)^n \lambda_-^{\alpha-n}), \text{ for } n = 2, 3, \dots$$

If we substitute

$$C = 2^{\frac{\alpha}{2}} \left(\Gamma\left(1 - \frac{\alpha}{2}\right) (\lambda_+^{\alpha-2} + \lambda_-^{\alpha-2})\right)^{-1} \tag{15}$$

then $X \sim RDTS(\alpha, C, \lambda_+, \lambda_-, 0)$ has zero mean and unit variance, and X is called the standard RDTS distribution and denoted by $X \sim \text{stdRDTS}(\alpha, \lambda_+, \lambda_-)$. Let m be a real number, σ be a positive real number, and $X \sim \text{stdCTS}(\alpha, \lambda_+, \lambda_-)$. Then

$$\sigma X + m \sim RDTS(\alpha, \sigma^{\alpha} C, \lambda_+/\sigma, \lambda_-/\sigma, m)$$

where C is equal to (15). The random variable Y is RDTS distributed, and its mean and variance are m and σ^2 , respectively.

INFINITELY DIVISIBLE DISTRIBUTIONS

A random variable Y is referred to as *infinitely divisible* if for each positive integer n , there are IID random variables Y_1, Y_2, \dots, Y_n such that $Y \stackrel{d}{=} \sum_{k=1}^n Y_k$ that is, the distribution of Y is the same as the distribution of $\sum_{k=1}^n Y_k$

For example, the normal distribution is infinitely divisible. Using the characteristic function for the normal distribution, we can easily check the property. Suppose $Y \sim N(\mu, \sigma^2)$. For any positive integer n , consider a sequence of IID random variables Y_1, Y_2, \dots, Y_n such that $Y_k \sim N(\mu/n, \sigma^2/n)$. Since Y_k 's are independent we have

$$E \left[\exp \left(iu \sum_{k=1}^n Y_k \right) \right] = \prod_{k=1}^n E [iuY_k]$$

The characteristic function of Y_k is given by

$$E [iuY_k] = \exp \left(\frac{iu\mu}{n} - \frac{\sigma^2 u^2}{2n} \right)$$

Hence, the characteristic function of $\sum_{k=1}^n Y_k$ is

$$E \left[\exp \left(iu \sum_{k=1}^n Y_k \right) \right] = \exp \left(iu\mu - \frac{\sigma^2 u^2}{2} \right)$$

which is the same as the characteristic function of Y . Therefore, $Y \stackrel{d}{=} \sum_{k=1}^n Y_k$.

Using similar arguments, we can show that the Poisson, gamma, variance-gamma (VG), inverse Gaussian (IG), α -stable, CTS, GTS, MTS, NTS(NIG), RDTS, and KRTS distributions are infinitely divisible. The relations of Y and $Y_k, k = 1, \dots, n$ for those distributions are presented in Table 2. We can show that the sum of infinitely divisible random variables is again infinitely divisible.

In the literature, the characteristic function of the one-dimensional infinitely divisible distribution is generalized by the Lévy-Khinchine formula:

$$\exp \left(i\gamma u - \frac{1}{2}\alpha^2 u^2 + \int_{-\infty}^{\infty} (e^{iux} - 1 - iux1_{|x|\leq 1})v(dx) \right) \tag{16}$$

Table 2 Infinitely Divisible Distributions

	$Y \stackrel{d}{=} \sum_{k=1}^n Y_k$	Y_k
Poisson	$Poiss(\lambda)$	$Poiss(\frac{\lambda}{n})$
Gamma	$Gamma(c, \lambda)$	$Gamma(\frac{c}{n}, \lambda)$
Variance gamma	$VG(C, \lambda_+, \lambda_-)$	$VG(\frac{C}{n}, \lambda_+, \lambda_-)$
Inverse Gaussian	$IG(c, \lambda)$	$IG(\frac{c}{n}, \lambda)$
Normal	$N(\mu, \sigma^2)$	$N(\frac{\mu}{n}, \frac{\sigma^2}{n})$
α -stable	$S_\alpha(\sigma, \beta, \mu)$	$S_\alpha(\frac{\sigma}{n}, \beta, \frac{\mu}{n})$
CTS	$CTS(\alpha, C, \lambda_+, \lambda_-, m)$	$CTS(\alpha, \frac{C}{n}, \lambda_+, \lambda_-, \frac{m}{n})$
GTS	$GTS(\alpha_+, \alpha_-, C_+, C_-, \lambda_+, \lambda_-, m)$	$GTS(\alpha_+ \alpha_-, \frac{C_+}{N}, \frac{C_-}{N}, \lambda_+, \lambda_-, \frac{m}{n})$
MTS	$MTS(\alpha, C, \lambda_+, \lambda_-, m)$	$MTS(\alpha, \frac{C}{n}, \lambda_+, \lambda_-, \frac{m}{n})$
NTS	$NTS(\alpha, C, \lambda, \beta, m)$	$NTS(\alpha, \frac{C}{n}, \lambda, \beta, \frac{m}{n})$
KRTS	$KRTS(\alpha_+, k_+, k_-, r_+, r_-, p_+, p_-, m)$	$KRTS(\alpha, \frac{k_+}{n}, \frac{k_-}{n}, r_+, r_-, p_+, p_-, \frac{m}{n})$
RDTS	$RDTS(\alpha, C, \lambda_+, \lambda_-, m)$	$RDTS(\alpha, \frac{C}{n}, \lambda_+, \lambda_-, \frac{m}{n})$

Table 3 Lévy Measures

Distributions	Lévy Measure
Poisson	$\nu_{\text{Poisson}}(dx) = \lambda \delta_x(dx)^5$
Gamma	$\nu_{\text{gamma}}(dx) = \frac{ce^{-\lambda x}}{x} 1_{x>0} dx$
Variance gamma	$\nu_{\text{VG}}(dx) = \left(\frac{C_+ e^{-\lambda_+ x}}{x} 1_{x>0} + \frac{C_- e^{-\lambda_- x }}{x} 1_{x<0} \right) dx$
Inverse Gaussian	$\nu_{\text{IG}}(dx) = \frac{ce^{-\frac{\lambda^2}{2} x}}{\sqrt{2\pi x^3}} 1_{x>0} dx$

In the formula, the measure ν is referred to as the Lévy measure. The measure is a Borel measure satisfying the conditions that $\nu(0) = 0$ and $\int_{\mathbb{R}} (1 \wedge |x^2|) \nu(dx) < \infty$. The parameters γ and σ are real numbers. The variable γ is referred to as the center or drift and determines the location. This triplet (σ^2, ν, γ) is uniquely defined for each infinitely divisible distribution and called a Lévy triplet.

If $\nu(dx) = 0$, then the characteristic function equals the characteristic function of the normal distribution. That is, the infinitely divisible distribution with $\nu(dx) = 0$ becomes the normal distribution with mean γ and variance σ^2 .

If $\sigma = 0$, then the distribution is referred to as a purely non-Gaussian distribution. The characteristic functions of purely non-Gaussian distributions are computed by

$$\exp \left(i\gamma u + \int_{-\infty}^{\infty} (e^{iux} - 1 - iux 1_{|x| \leq 1}) \nu(dx) \right)$$

Hence, except for the location determined by γ , all the properties of the distribution are characterized by the Lévy measure $\nu(dx)$. The Poisson, gamma, VG, IG, α -stable, CTS, GTS, MTS, NTS, RDTS, and KRTS distributions are purely non-Gaussian distributions. The Lévy measure of the Poisson, gamma, VG, and IG distributions are given in Table 3.

The Lévy measure of the α -stable distribution is given by

$$\nu_{\text{stable}}(dx) = \left(\frac{C_+}{x^{1+\alpha}} 1_{x>0} + \frac{C_-}{|x|^{1+\alpha}} 1_{x<0} \right) dx \quad (17)$$

Using the Lévy Khinchine formula we can obtain the characteristic function in (1).⁵

The Lévy measure of the CTS, MTS, NTS, KRTS, and RDTS distributions can be obtained by multiplying the tempering function by the Lévy measure of α -stable distribution. For example, if we take $q(x) = e^{-\lambda_+ x} 1_{x>0} + e^{-\lambda_- |x|} 1_{x<0}$ as the tempering function, then we obtain the Lévy measure of the CTS distribution as

$$\begin{aligned} \nu(dx) &= q(x) \nu_{\text{stable}}(dx) \\ &= \left(\frac{C + e^{-\lambda_+ x}}{x^{1+\alpha}} 1_{x>0} + \frac{C_- e^{-\lambda_- |x|}}{|x|^{1+\alpha}} 1_{x<0} \right) dx \end{aligned}$$

Tempering functions of the other distributions are presented in Table 4. For this reason, they are referred to as the *tempered stable distributions*. The GTS distribution is also a purely non-Gaussian distribution, but not a tempered stable distribution in this sense. Indeed, its Lévy measure is given by

$$\nu(dx) = \left(\frac{C + e^{-\lambda_+ x}}{x^{1+\alpha_+}} 1_{x>0} + \frac{C_- e^{-\lambda_- |x|}}{|x|^{1+\alpha_-}} 1_{x<0} \right) dx$$

However, we will refer to the GTS distribution as a tempered stable distribution for convenience. Using the Lévy measures and the Lévy-Khinchine formula, we can obtain the characteristic functions (1), (2), (4), (6), (8), (12), and (14).

Generalizations of the tempering function and the tempered stable distribution have been studied in the literature.⁶

Table 4 Tempering Functions

	Tempering Function $q(x)$
CTS	$e^{-\lambda_+x}1_{x>0} + e^{-\lambda_- x }1_{x<0}$
MTS	$(\lambda_+x)^{\frac{\alpha+1}{2}}K_{\frac{\alpha+1}{2}}(\lambda_+x)1_{x>0} + (\lambda_- x)^{\frac{\alpha+1}{2}}K_{\frac{\alpha+1}{2}}(\lambda_- x)1_{x<0}$
NTS	$e^{\beta x}(\lambda x)^{\frac{\alpha+1}{2}}K_{\frac{\alpha+1}{2}}(\lambda x)$
KRTS	$r_+^{-p_+} \int_0^{r_+} e^{-x/s} s^{\alpha+p_+-1} ds 1_{x>0} + r_-^{-p_-} \int_0^{r_-} e^{- x /s} s^{\alpha+p_- -1} ds 1_{x<0}$
RDTs	$e^{-\frac{\lambda_+x^2}{2}}1_{x>0} + e^{-\frac{\lambda_- x ^2}{2}}1_{x<0}$

Exponential Moments

The exponential moment of a random variable X is defined by $E[e^{uX}]$ for some real number u . Existence of the exponential moment is important for modeling an asset price process in option pricing theory.

The exponential moment of the normal distribution is given by

$$E[e^{uX}] = \exp\left(\mu u + \frac{\sigma^2 u^2}{2}\right)$$

where $X \sim N(\mu, \sigma)$.

Using the Lévy measure we can check the existence of the exponential moment for an infinitely divisible random variable. The following theorem (see Sato, 1999) provides a useful tool to verify the existence of an exponential moment of an infinitely divisible distribution.

Theorem Let X be an infinitely divisible random variable with the Lévy triplet (σ^2, ν, γ) and let $u \in \mathbb{R}$. Then $E[e^{uX}] < \infty$ if and only if

$$\int_{|x|>1} e^{ux} \nu(dx) < \infty \tag{18}$$

In this case,

$$E[e^{uX}] = \phi_X(-iu)$$

where ϕ is the characteristic function of X and $i = \sqrt{-1}$.

The existence of exponential moments in the tempered stable distributions is as following:

- For the α -stable random variable X , the exponential moment of X generally does not exist.

However, if $X \sim S_\alpha(\sigma, 1, 0)$, then $E[e^{uX}] < \infty$ for $u < 0$. In this case,

$$E[e^{uX}] = \begin{cases} \exp\left(-\frac{\sigma^\alpha}{\cos\frac{\pi\alpha}{2}}\mu\alpha\right), & \alpha \neq 1 \\ \exp\left(\frac{2\sigma}{\pi}u \ln u\right), & \alpha = 1 \end{cases}$$

- For the CTS, GTS, and MTS distributions, the condition (18) is satisfied if and only if $-\lambda_- \leq u \leq \lambda_+$. Hence, $E[e^{uX}] < \infty$ for $u \in [-\lambda_-, \lambda_+]$.
- For the KRTS distribution, $E[e^{uX}] < \infty$ for $u \in [-1/r_-, 1/r_+]$.
- For the NTS and the NIG distributions, $E[e^{uX}] < \infty$ for $u \in [-\lambda - \beta, \lambda - \beta]$.
- For the RDTs distribution, (18) is satisfied for the entire real number u . Hence, $E[e^{uX}] < \infty$ for all $u \in \mathbb{R}$.

If $E[e^{uX}] < \infty$, then we can define the log-Laplace transform for the random variable X . The log-Laplace transform is given by

$$L(u) = \log E[e^{uX}] = \log \phi(-iu)$$

if (18) is satisfied.

For example, let $X \sim \text{stdCTS}(\alpha, \lambda_+, \lambda_-)$. The log-Laplace transform L_{CTS} of X is defined on $u \in [-\lambda_-, \lambda_+]$, and is given by

$$\begin{aligned} L_{\text{CTS}}(u; a, \lambda_+, \lambda_-) &= \log \phi_{\text{CTS}}(-iu; \alpha, C, \lambda_+, \lambda_-, 0) \\ &= \frac{(\lambda_+ - u)^\alpha - \lambda_+^\alpha + (\lambda_- + u)^\alpha - \lambda_-^\alpha}{\alpha(\alpha - 1)(\lambda_+^{\alpha-2} + \lambda_-^{\alpha-2})} \\ &\quad - \frac{u(\lambda_+^{\alpha-1} - \lambda_-^{\alpha-1})}{(1 - \alpha)(\lambda_+^{\alpha-2} + \lambda_-^{\alpha-2})} \end{aligned}$$

where C is satisfied (3). Using the same method, we can obtain the log-Laplace transform of the other standard tempered stable distributions as follows:

- Standard GTS distribution:

$$L_{GTS}(u; \alpha_+, \alpha_-, \lambda_+, \lambda_-) = \log \phi_{GTS}(-iu; \alpha_+, \alpha_-, C_+, C_-, \lambda_+, \lambda_-, 0)$$

on $u \in [-\lambda_-, \lambda_+]$ where C_+ and C_- satisfy (5).

- Standard MTS distribution:

$$L_{MTS}(u; \alpha, \lambda_+, \lambda_-) = \log \phi_{MTS}(-iu; \alpha, C, \lambda_+, \lambda_-, 0)$$

on $u \in [-\lambda_-, \lambda_+]$ where C satisfies (7).

- Standard NTS distribution:

$$L_{NTS}(u; \alpha, \lambda, \beta) = \log \phi_{NTS}(-iu; \alpha, C, \lambda, \beta, 0)$$

on $u \in [-\lambda - \beta, \lambda - \beta]$ where C satisfies (9).

- Standard NIG distribution:

$$L_{NIG}(u; \lambda, \beta) = \log \phi_{NIG}(-iu; C, \lambda, \beta, 0)$$

on $u \in [-\lambda - \beta, \lambda - \beta]$ where C satisfies (11).

- Standard KRTS distribution:

$$L_{KRTS}(u; \alpha, r_+, r_-, p_+, p_-) = \log \phi_{KRTS}(-iu; \alpha, k_+, k_-, r_+, r_-, p_+, p_-, 0)$$

on $u \in [-\lambda_-, \lambda_+]$ where k_+ and k_- satisfy (13).

- Standard RDTS distribution:

$$L_{RDTS}(u; \alpha, \lambda_+, \lambda_-) = \log \phi_{RDTS}(-iu; \alpha, C, \lambda_+, \lambda_-, 0)$$

on $u \in \mathbb{R}$ where C satisfies (15).

HYPERGEOMETRIC FUNCTION AND CONFLUENT HYPERGEOMETRIC FUNCTION

In this entry, we referred to the hypergeometric function and the confluent hypergeometric function. Here we describe these two special functions. (For more details, see Andrews,

1998.) We begin by introducing the following notation

$$(a)_0 = 1, \quad (a)_n = a(a+1)\cdots(a+n-1) \\ n = 1, 2, 3, \dots, a \in \mathbb{R} \tag{19}$$

and we refer to the notation as the Pochhammer symbol. By properties of the gamma function, the Pochhammer symbol can also be defined by

$$(a)_n = \frac{\Gamma(a+n)}{\Gamma(a)}, \quad n = 0, 1, 2, 3, \dots$$

From (19), we obtain

$$(2n+1)! = 2^{2n}n! \left(\frac{3}{2}\right)_n \tag{20}$$

The Hypergeometric Function

The function

$${}_2F_1(a, b; c; x) = \sum_{n=0}^{\infty} \frac{(a)_n(b)_n}{(c)_n} \frac{x^n}{n!}, \quad |x| < 1 \tag{21}$$

is called the hypergeometric function. If $c \neq 0, -1, -2, \dots$, the function $F(a, b; c; x)$ is a solution to the linear second-order differential equation

$$x(1-x)y'' + (c - (a+b+1)x)y' - aby = 0 \tag{22}$$

referred to as the hypergeometric equation. Moreover, if $c \neq 0, \pm 1, \pm 2, \dots$,

$$y = C_1 {}_2F_1(a, b; c; x) + C_2 x^{1-c} {}_2F_1(1+a-c, 1+b-c; 2-c; x)$$

for any constants C_1 and C_2 , is a general solution to equation (22). For $k = 1, 2, 3, \dots$, k th derivatives are obtained from the following equation:

$$\frac{d^k}{dx^k} {}_2F_1(a, b; c; x) = \frac{(a)_k(b)_k}{(c)_k} {}_2F_1(a+k, b+k; c+k; x) \tag{23}$$

The Confluent Hypergeometric Function

The function

$$M(a; c; x) = \sum_{n=0}^{\infty} \frac{(a)_n x^n}{(c)_n n!}, \quad -\infty < x < \infty \quad (24)$$

is called the confluent hypergeometric function and is obtained by the limit of the hypergeometric function as follows:

$$M(a; c; x) = \lim_{b \rightarrow \infty} F(a, b; c; x/b)$$

The function $M(a; c; x)$ is a solution of the linear second-order differential equation

$$xy'' + (c - x)y' - ay = 0 \quad (25)$$

referred to as the confluent hypergeometric equation. Moreover, if $c \neq 0, \pm 1, \pm 2, \dots$,

$$y = C_1 M(a; c; x) + C_2 x^{1-c} F(1 + a - c; 2 - c; x)$$

for any constants C_1 and C_2 , is a general solution of equation (25). For $k = 1, 2, 3 \dots$, k th derivatives are obtained by the following equation:

$$\frac{d^k}{dx^k} M(a; c; x) = \frac{(a)_k}{(c)_k} M(a + k; c + k; x) \quad (26)$$

KEY POINTS

- The distribution assumed in financial models for asset returns is the normal or Gaussian distribution. Real-world asset returns, however, have been observed to be skewed and non-symmetric, two features that are inconsistent with the normal distribution.
- Although the non-Gaussian alpha-stable distribution is superior to the normal distribution because it allows for skewness and fat tails, it is not suitable in certain modeling applications such as in modeling option prices. This is because the mean, variance, and exponential moments of the return distribution have to exist. The smoothly truncated sta-

ble distribution, obtained by tempering the tail properties of the alpha-stable distribution, have been proposed for modeling in such instances.

- There are six tempered stable distributions: classical tempered stable distribution, generalized classical tempered stable distribution, modified tempered stable distribution, normal tempered stable distribution, Kim-Rachev tempered stable distribution, and rapidly decreasing tempered stable distribution. All six tempered stable distributions and the alpha-stable distribution are defined by their characteristic functions.
- The infinitely divisible distribution is characterized by the Lévy-Khinchine formula and contains the alpha-stable and the tempered stable distributions as special cases.

NOTES

1. Extensive analysis of α -stable distributions and their properties can be found in Samorodnitsky and Taqqu (1994), Rachev and Mittnik (2000), and Stoyanov and Racheva-Iotova (2004a, 2004b).
2. The KoBoL distribution (see Boyarchenko and Levendorskii, 2000) is obtained by substituting $\alpha = \alpha_+ = \alpha_-$, the truncated Lévy flight is obtained by substituting $\lambda = \lambda_+ = \lambda_-$ and $\alpha = \alpha_+ = \alpha_-$, while the CGMY distribution (see Carr et al., 2002) is obtained by substituting $C = C_+ = C_-$, $G = \lambda_-$, $M = X_+$ and $Y = \alpha_+ = \alpha_-$.
3. The NTS distribution was originally obtained using a time-changed Brownian motion with a tempered stable subordinator by Barndorff-Nielsen and Levendorskii (2001). Later, Kim, Rachev, Chung, and Bianchi (2008c) define the NTS distribution by the exponential tilting for the symmetric MTS distribution.
4. The NIG distribution has been used for financial modeling by Barndorff-Nielsen (1998, 1997) and Rydberg (1997).

5. More details about the calculation can be found in Samorodnitsky and Taqqu (1994) and Sato (1999).
6. The tempered stable distribution has been generalized by Rosiński (2007) and Bianchi et al. (2010). Rosiński (2007) defined the tempering function as the completely monotone function. The complete monotonicity of the tempering function $q(x)$ means that $(-1)^n \frac{d^n}{dx^n} q(x) > 0$ for all $n = 0, 1, 2, \dots$ and $x \in \mathbb{R}$ with $x \neq 0$. The CTS and the KRTS distributions are included in Rosiński's generalization. In Bianchi et al. (2010), the tempering function is defined by the positive definite radial function. The RDTs and the MTS distributions are subclasses of the class of the TID distributions.

REFERENCES

- Barndorff-Nielsen, O. E. (1997). Normal inverse Gaussian distributions and stochastic volatility modelling. *Scandinavian Journal of Statistics* 24: 1–13.
- Barndorff-Nielsen, O. E. (1998). Processes of normal inverse Gaussian type. *Finance and Stochastics* 41–68.
- Barndorff-Nielsen, O. E., and Levendorskii, S. (2001). Feller processes of normal inverse Gaussian type. *Quantitative Finance* 1: 318–331.
- Bianchi, M. L., Rachev, S. T., Kim, Y. S., and Fabozzi, F. J. (2010). Tempered infinitely divisible distributions and processes. *Theory of Probability and Its Applications (TVP), Society for Industrial and Applied Mathematics (SIAM)* 55, 1: 58–86.
- Boyarchenko, S. I., and Levendorskii, S. Z. (2000). Option pricing for truncated Lévy processes. *International Journal of Theoretical and Applied Finance* 3.
- Carr, P., Geman, H., Madan, D., and Yor, M. (2002). The fine structure of asset returns: An empirical investigation. *Journal of Business* 75, 2: 305–332.
- Cootner, P. (1964). *The Random Character of Stock Market Prices*. Cambridge, MA: The MIT Press.
- Kim, Y. S., Rachev, S. T., Bianchi, M. L., and Fabozzi, F. (2008a). Financial market models with Lévy processes and time-varying volatility. *Journal of Banking and Finance* 32, 7: 1363–1378.
- Kim, Y. S., Rachev, S. T., Bianchi, M. L., and Fabozzi, F. (2008b). A new tempered stable distribution and its application to finance. In G. Bol, S. T. Rachev, and R. Wurth (Eds.), *Risk Assessment: Decisions in Banking and Finance*. Heidelberg: Physica-Verlag.
- Kim, Y. S., Rachev, S. T., Bianchi, M. L., and Fabozzi, F. (2010). Tempered stable and tempered infinitely divisible GARCH models. *Journal of Banking and Finance* 34: 2096–2109.
- Kim, Y. S., Rachev, S. T., Chung, D., and Bianchi, M. (2008c). A modified tempered stable distribution with volatility clustering. In J. O. Soares, J. P. Pina, and M. Catalaõ-Lopes (Eds.), *New Developments in Financial Modelling* 344–365. Cambridge Scholars Publishing.
- Kim, Y. S., Rachev, S. T., Chung, D. M., and Bianchi, M. L. (2009). The modified tempered stable distribution, GARCH-models and option pricing. *Probability and Mathematical Statistics* 29, 1: 91–117.
- Menn, C., and Rachev, S. T. (2009). Smoothly truncated stable distributions, GARCH-models, and option pricing. *Mathematical Methods of Operations Research* 63, 3: 411–438.
- Rachev, S. T., Kim, Y. S., Bianchi, M. L., and Fabozzi, F. J. (2011). *Financial Models with Lévy Processes and Volatility Clustering*. Hoboken, NJ: John Wiley & Sons.
- Rachev, S. T., and Mittnik, S. (2000). *Stable Parettian Models in Finance*. Chichester: John Wiley & Sons.
- Rosiński, J. (2007). Tempering stable processes. *Stochastic Processes and Their Applications* 117, 6: 677–707.
- Rydberg, T. (1997). The normal inverse Gaussian Lévy process: Simulation and approximation. *Communications in Statistics. Stochastic Models* 13: 887–910.
- Samorodnitsky, G., and Taqqu, M. (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. New York: CRC Press.
- Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. New York: Cambridge University Press.
- Stoyanov, S., and Racheva-Iotova, B. (2004a). Univariate stable laws in the field of finance. Approximation of density and distribution functions. *Journal of Concrete and Applicable Mathematics* 2, 1: 37–58.
- Stoyanov, S., and Racheva-Iotova, B. (2004b). Univariate stable laws in the field of finance. parameter estimation. *Journal of Concrete and Applicable Mathematics* 2, 4: 24–49.

Fat Tails, Scaling, and Stable Laws

SERGIO M. FOCARDI, PhD
Partner, The Intertek Group

FRANK J. FABOZZI, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

Abstract: Fat-tailed laws have been found in many economic variables. Fully approximating a finite economic system with fat-tailed laws depends on an accurate statistical analysis of the phenomena, but also on a number of the theoretical implications of subexponentiality and scaling. Modeling financial variables with stable laws implies the assumption of infinite variance, which seems to contradict empirical observations. Nevertheless, scaling laws might still be an appropriate modeling paradigm given the complex interaction of distributional shape and correlations in price processes. They might help in understanding not only the sheer size of economic fluctuations but also the complexity of economic cycles. There are applications where scaling laws play a fundamental role, in particular in risk management and financial optimization. Ignoring the possibility of large deviations would render financial risk management ineffective and dangerous.

Most models of stochastic processes and time series assume that distributions have finite mean and finite variance. In this entry we describe fat-tailed distributions with infinite variance. Fat-tailed distributions have been found in many financial economic variables ranging from forecasting returns on financial assets to modeling recovery distributions in bankruptcies. They have also been found in numerous insurance applications such as catastrophic insurance claims and in value-at-risk measures employed by risk managers.

In this entry, we review the related concepts of *fat-tailed*, *power-law*, and *Levy-stable distributions*, *scaling*, and *self-similarity*, as well as explore the mechanisms that generate these distributions.

We discuss the key intuition relative to the applicability of fat-tailed or scaling processes to finance: In a fat-tailed or scaling world (as opposed to an ergodic world), the past does not offer an exhaustive set of possible configurations. Adopting, as an approximation, a scaling description of financial phenomena implies the belief that only a small space of possible configurations has been explored; vast regions remain unexplored.

We begin with the mathematics of fat-tailed processes, followed by a discussion of classical *extreme value theory* for independent and identically distributed sequences. We then explore the consequences of eliminating the assumption of independence and discuss different concepts of

scaling and self similarity. We will not provide a review of the literature on the evidence of fat tails in financial markets. For a review, see Rachev, Menn, and Fabozzi (2005).

SCALING, STABLE LAWS, AND FAT TAILS

Let's begin with a review of the different but related concepts and properties of fat tails, power laws, and stable laws. These concepts appear frequently in the financial and economic literature, applied to both random variables and stochastic processes.

Fat Tails

Consider a random variable X . By definition, X is a real-valued function from the set Ω of the possible outcomes to the set R of real numbers, such that the set $(X \leq x)$ is an event. If $P(X \leq x)$ is the probability of the event $(X \leq x)$, the function $F(x) = P(X \leq x)$ is a well-defined function for every real number x . The function $F(x)$ is called the cumulative distribution function, or simply the distribution function, of the random variable X . Note that X denotes a function $\Omega \rightarrow R$, x is a real variable, and $F(x)$ is an ordinary real-valued function that assumes values in the interval $[0,1]$. If the function $F(x)$ admits a derivative

$$f(x) = \frac{dF(x)}{dx}$$

The function $f(x)$ is called the probability density of the random variable X . The function $\bar{F}(x) = 1 - F(x)$ is the tail of the distribution $F(x)$. The function $\bar{F}(x)$ is called the survival function.

Fat tails are somewhat arbitrarily defined. Intuitively, a fat-tailed distribution is a distribution that has more weight in the tails than some reference distribution. The exponential decay of the tail is generally assumed as the borderline separating fat-tailed from light-tailed distributions. In the literature, distributions with

a power-law decay of the tails are referred to as *heavy-tailed distributions*. It is sometimes assumed that the reference distribution is Gaussian (i.e., normal), but this is unsatisfactory; it implies, for instance, that exponential distributions are fat-tailed because Gaussian tails decay as the square of an exponential and thus faster than an exponential.

These characterizations of fat-tailedness (or heavy-tailedness) are not convenient from a mathematical and statistical point of view. It would be preferable to define fat-tailedness in terms of a function of some essential property that can be associated to it. Several proposals have been advanced. Widely used definitions focus on the moments of the distribution. Definitions of fat-tailedness based on a single moment focus either on the second moment, the variance, or the kurtosis, defined as the fourth moment divided by the square of the variance. In fact, a distribution is often considered fat-tailed if its variance is infinite or if it is leptokurtic (i.e., its kurtosis is greater than 3). However, as remarked by Bryson (1982), definitions of this type are too crude and should be replaced by more complete descriptions of tail behavior.

Others consider a distribution fat-tailed if all its exponential moments are infinite, $E[e^{sX}] = \infty$ for every $s \geq 0$. This condition implies that the moment-generating function does not exist. Some suggest weakening this condition, defining fat-tailed distributions as those distributions that do not have a finite exponential moment of first order. Exponential moments are particularly important in finance and economics when the logarithm of variables, for instance logprices, are the primary quantity to be modeled.¹

Fat-tailedness has a consequence of practical importance: The probability of *extremal events* (i.e., the probability that the random variable assumes large values) is much higher than in the case of normal distributions. A fat-tailed distribution assigns higher probabilities to extremal events than would a normal distribution.

For instance, a six-sigma event (i.e., a realized value of a random variable whose difference from the mean is six times the size of the standard deviation) has a near zero probability in a Gaussian distribution but might have a non-negligible probability in fat-tailed distributions.

The notion of fat-tailedness can be made quantitative as different distributions have different degrees of fat-tailedness. The degree of fat-tailedness dictates the weight of the tails and thus the probability of extremal events. Extreme value theory attempts to estimate the entire tail region, and therefore the degree of fat-tailedness, from a finite sample. A number of indicators for evaluating the size of extremal events have been proposed; among these is the extremal claim index proposed in Embrechts, Kluppelberg, and Mikosch (1999), which plays an important role in risk management.

The Class \mathcal{L} of Fat-Tailed Distributions

Many important classes of fat-tailed distributions have been defined; each is characterized by special statistical properties that are important in given application domains. We will introduce a number of such classes in order of inclusion, starting from the class with the broadest membership: the class \mathcal{L} , which is defined as follows. Suppose that F is a distribution function defined in the domain $(0, \infty)$ with $F < 1$ in the entire domain (i.e., F is the distribution function of a positive random variable with a tail that never decays to zero). It is said that $F \in \mathcal{L}$ if, for any $y > 0$, the following property holds:

$$\lim_{x \rightarrow \infty} \frac{\bar{F}(x - y)}{\bar{F}(x)} = 1, \quad \forall y > 0$$

We can rewrite the above property in an equivalent (and perhaps more intuitive from the probabilistic point of view) way. Under the same assumptions as above, it is said that, given a positive random variable X , its distribution function $F \in \mathcal{L}$ if the following property holds

for any $y > 0$:

$$\begin{aligned} & \lim_{x \rightarrow \infty} P(X > x + y | X > x) \\ &= \lim_{x \rightarrow \infty} \frac{\bar{F}(x + y)}{\bar{F}(x)} = 1, \quad \forall y > 0 \end{aligned}$$

Intuitively, this second property means that if it is known that a random variable exceeds a given value x , then it will exceed any bigger value with certainty as the value x tends to infinity. Some authors define a distribution as being heavy-tailed if it satisfies this property.²

It can be demonstrated that if a distribution $F(x) \in \mathcal{L}$, then it has the following properties:

- Infinite exponential moments of every order: $E[e^{sX}] = \infty$ for every $s \geq 0$
- $\lim_{x \rightarrow \infty} \bar{F}(x)e^{\lambda x} = \infty, \quad \forall \lambda > 0$

As distributions in class \mathcal{L} have infinite exponential moments of every order, they satisfy one of the previous definitions of fat-tailedness. However, they might have finite or infinite mean and variance.

The class \mathcal{L} is in fact quite broad. It includes, in particular, the two classes of subexponential distributions and distributions with regularly varying tails that are discussed in the following sections.

Subexponential Distributions

A class of fat-tailed distributions, widely used in insurance and telecommunications, is the class S of *subexponential distributions*. Introduced by Chistyakov (1964), subexponential distributions can be characterized by two equivalent properties: (1) the convolution closure property of the tails and (2) the property of the sums.³

The *convolution closure property* of the tails prescribes that the shape of the tail is preserved after the summation of identical and independent copies of a variable. This property asserts that, for $x \rightarrow \infty$, the tail of a sum of independent and identical variables has the same shape as the tail of the variable itself. As the distribution of a sum of n independent variables is the n -convolution of their distributions, the

convolution closure property can be written as

$$\lim_{x \rightarrow \infty} \frac{\bar{F}^{n^*}(x)}{\bar{F}(x)} = n$$

Note that Gaussian distributions do not have this property although the sum of independent Gaussian distributions is again a Gaussian distribution. Subexponential distributions can be characterized by another important (and perhaps more intuitive) property, which is equivalent to the convolution closure property: In a sum of n variables, the largest value will be of the same order of magnitude as the sum itself. For any n , define

$$S_n(x) = \sum_{i=1}^n X_i$$

as a sum of independent and identical copies of a variable X and call M_n their maxima. In the limit of large x , the probability that the tail of the sum exceeds x equals the probability that the largest summand exceeds x :

$$\lim_{x \rightarrow \infty} \frac{P(S_n > x)}{P(M_n > x)} = 1$$

The class S of subexponential distributions is a proper subset of the class \mathfrak{L} . Every subexponential distribution belongs to the class \mathfrak{L} while it can be demonstrated (but this is not trivial) that there are distributions that belong to the class \mathfrak{L} but not to the class S . Distributions that have both properties are called subexponential as it can be demonstrated that, as all distributions in \mathfrak{L} , they satisfy the property:

$$\lim_{x \rightarrow \infty} \bar{F}(x)e^{\lambda x} = \infty, \quad \forall \lambda > 0$$

Note, however, that the class of distributions that satisfies the latter property is broader than the class of subexponential distributions; this is because the former includes, for instance, the class \mathfrak{L} .⁴

Subexponential distributions do not have finite exponential moments of any order, that is, $E[e^{sX}] = \infty$ for every $s \geq 0$. They may or may not have a finite mean and/or a finite variance. Consider, in fact, that the class of subexpo-

ponential distributions includes both Pareto and Weibull distributions. The former have infinite variance but might have finite or infinite mean depending on the index; the latter have finite moments of every order (see below).

The key indicators of subexponentiality are (1) the equivalence in the distribution of the tail between a variable and a sum of independent copies of the same variable and (2) the fact that a sum is dominated by its largest term. The importance of the largest terms in a sum can be made more quantitative using measures such as the large claims index introduced in Embrechts, Kluppelberg, and Mikosch (1999) that quantifies the ratio between the largest p terms in a sum and the entire sum.

The class of subexponential distributions is quite large. It includes not only Pareto and stable distributions but also log-gamma, lognormal, Benkander, Burr, and Weibull distributions. Pareto distributions and stable distributions are a particularly important subclass of subexponential distributions; these will be described in some detail below.

Power-Law Distributions

Power-law distributions are a particularly important subset of subexponential distributions. Their tails follow approximately an inverse power law, decaying as $x^{-\alpha}$. The exponent α is called the tail index of the distribution. To express formally the notion of approximate power-law decay, we need to introduce the class $\mathfrak{R}(\alpha)$, equivalently written as \mathfrak{R}_α of regularly varying functions.

A positive function f is said to be regularly varying with index α or $f \in \mathfrak{R}(\alpha)$ if the following condition holds:

$$\lim_{x \rightarrow \infty} \frac{f(tx)}{f(x)} = t^\alpha$$

A function $f \in \mathfrak{R}(\alpha)$ is called slowly varying. It can be demonstrated that a regularly varying function $f(x)$ of index α admits the representation $f(x) = x^\alpha l(x)$ where $l(x)$ is a slowly varying function.

A distribution F is said to have a regularly varying tail if the following property holds:

$$\bar{F} = x^{-\alpha}l(x)$$

where l is a slowly varying function. An example of a distribution with a regularly varying tail is Pareto's law. The latter can be written in various ways, including the following:

$$\bar{F}(x) = P(X > x) = \frac{c}{c + x^\alpha} \text{ for } x \geq 0$$

Power-law distributions are thus distributions with regularly varying tails. It can be demonstrated that they satisfy the convolution closure property of the tail. The distribution of the sum of n independent variables of tail index α is a power-law distribution of the same index α . Note that this property holds in the limit for $x \rightarrow \infty$. Distributions with regularly varying tails are therefore a proper subset of subexponential distributions.

Being subexponential, power laws have all the general properties of fat-tailed distributions and some additional ones. One particularly important property of distributions with regularly varying tails, valid for every tail index, is the rank-size order property. Suppose that samples from a power law of tail index α are ordered by size, and call S_r the size of the r th sample. One then finds that the law

$$S_r = ar^{-\frac{1}{\alpha}}$$

is approximately verified. The well-known Zipf's law is an example of this rank-size ordering. Zipf's law states that the size of an observation is inversely proportional to its rank. For example, the frequency of words in an English text is inversely proportional to their rank. The same is approximately valid for the size of U.S. cities.

Many properties of power-law distributions are distinctly different in the three following ranges of α : $0 < \alpha \leq 1$, $1 < \alpha \leq 2$, $\alpha > 2$. The threshold $\alpha = 2$ for the tail index is important as it marks the separation between the applicability of the standard central limit theorem (CLT); the threshold $\alpha = 1$ is important as it separates

variables with a finite mean from those with infinite mean. Let's take a closer look at the *law of large numbers* and the CLT.

The Law of Large Numbers and the Central Limit Theorem

There are four basic versions of the law of large numbers (LLN), two weak laws of large numbers (WLLN), and two strong laws of large numbers (SLLN).

The two versions of the WLLN are formulated as follows.

1. Suppose that the variables X_i are IID with finite mean $E[X_i] = E[X] = \mu$. Under this condition it can be demonstrated that the empirical average tends to the mean in probability:

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} \xrightarrow[n \rightarrow \infty]{P} E[X] = \mu$$

2. If the variables are only independently distributed (ID) but have finite means and variances (μ_i, σ_i) , then the following relationship holds:

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} \xrightarrow[n \rightarrow \infty]{P} \frac{\sum_{i=1}^n \bar{X}_i}{n} = \frac{\sum_{i=1}^n \mu_i}{n}$$

In other words, the empirical average of a sequence of finite-mean finite-variance variables tends to the average of the means.

The two versions of the SLLN are formulated as follows.

1. The empirical average of a sequence of IID variables X_i tends almost surely to a constant a if and only if the expected value of the variables is finite. In addition, the constant a is equal to μ . Therefore, if and only if $|E[X_i]| = |E[X]| = |\mu| < \infty$ the following relationship holds:

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} \xrightarrow[n \rightarrow \infty]{A.S.} E[X] = \mu$$

where convergence is in the sense of almost sure convergence.

2. If the variables X_i are only independently distributed (ID) but have finite means and variances (μ_i, σ_i) and

$$\lim_{n \rightarrow \infty} \frac{1}{n^2} \sum_{i=1}^n \sigma_i^2 < \infty$$

then the following relationship holds:

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} \xrightarrow[n \rightarrow \infty]{A.S.} \frac{\sum_{i=1}^n \bar{X}_i}{n} = \frac{\sum_{i=1}^n \mu_i}{n}$$

Suppose the variables are IID. If the scaling factor n is replaced with \sqrt{n} , then the limit relation no longer holds as the normalized sum

$$\frac{\sum_{i=1}^n X_i}{\sqrt{n}}$$

diverges. However, if the variables have finite second-order moments, the classical version of the CLT can be demonstrated. In fact, under the assumption that both first- and second-order moments are finite, it can be shown that

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{D} \Phi$$

$$S_n = \sum_{i=1}^n X_i$$

where μ, σ are respectively the expected value and standard deviation of X , and Φ the standard normal distribution.

If the tail index $\alpha > 1$, variables have finite expected value and the SLNN holds. If the tail index $\alpha > 2$, variables have finite variance and the CLT in the previous form holds. If the tail index $\alpha \leq 2$, then variables have infinite variance: The CLT in the previous form does not hold. In fact, variables with $\alpha \leq 2$ belong to the domain of attraction of a stable law of index α . This means that a sequence of properly normalized and centered sums tends to a stable distribution with infinite variance. In this case, the CLT takes the form

$$\frac{S_n - n\mu}{n^{\frac{1}{\alpha}}} \xrightarrow{D} G_\alpha, \text{ if } 1 < \alpha \leq 2$$

$$\frac{S_n}{n^{\frac{1}{\alpha}}} \xrightarrow{D} G_\alpha, \text{ if } 0 < \alpha \leq 1$$

where G are stable distributions as defined below. Note that the case $\alpha = 2$ is somewhat special: variables with this tail index have infinite variance but fall nevertheless in the domain of attraction of a normal variable, that is, G_2 . Below the threshold 1, distributions have neither finite variance nor finite mean. There is a sharp change in the normalization behavior at this tail-index threshold.

Stable Distributions

Stable distributions are not, in their generality, a subset of fat-tailed distributions as they include the normal distribution. There are different, equivalent ways to define stable distributions. Let's begin with a key property: the equality in distribution between a random variable and the (normalized) independent sum of any number of identical replicas of the same variable. This is a different property than the closure property of the tail insofar as (1) it involves not only the tail but the entire distribution and (2) equality in distribution means that distributions have the same functional form but, possibly, with different parameters. Normal distributions have this property: The sum of two or more normally distributed variables is again a normally distributed variable. But this property holds for a more general class of distributions called stable distributions or Levy-stable distributions.⁵ Normal distributions are thus a special type of stable distributions.

The above can be formalized as follows: Stable distributions can be defined as those distributions for which the following identity in distribution holds for any number $n \geq 2$:

$$\sum_{i=1}^n X_i \stackrel{D}{=} C_n X + D_n$$

where X_i are identical independent copies of X and the C_n, D_n are constants. Alternatively, the same property can be expressed stating that stable distributions are distributions for which the following identity in distribution holds:

$$AX_1 + BX_2 \stackrel{D}{=} CX + D$$

Stable distributions are also characterized by another property that might be used in defining them: a stable distribution has a domain of attraction (i.e., it is the limit in distribution of a normalized and centered sum of identical and independent variables). Stable distributions coincide with all variables that have a domain of attraction.

Except in the special cases of Gaussian ($\alpha = 2$), symmetric Cauchy ($\alpha = 1, \beta = 0$), and stable inverse Gaussian ($\alpha = \frac{1}{2}, \beta = 0$) distributions, stable distributions cannot be written as simple formulas; formulas have been discovered but are not simple. However, stable distributions can be characterized in a simple way through their characteristic function, the Fourier transform of the distribution function. In fact, this function can be written as

$$\Phi_X(t) = \exp\{i\gamma t - c|t|^\alpha[1 - i\beta\text{sign}(t)z(t, \alpha)]\}$$

where $t \in R, \gamma \in R, c > 0, \alpha \in (0, 2), \beta \in [-1, 1]$, and

$$z(t, \alpha) = \tan \frac{\pi\alpha}{2} \text{ if } \alpha \neq 1$$

$$z(t, \alpha) = -2 \log |t| \text{ if } \alpha = 1$$

It can be shown that only distributions with this characteristic function are stable distributions (i.e., they are the only distributions closed under summation). A stable law is characterized by four parameters: α, β, c , and γ . Normal distributions correspond to the parameters: $\alpha = 2, \beta = 0, \gamma = 0$.

Even if stable distributions cannot be written as simple formulas, the asymptotic shape of their tails can be written in a simple way. In fact, with the exception of Gaussian distributions, the tails of stable laws obey an inverse power law with exponent α (between 0 and 2). Normal distributions are stable but are an exception as their tails decay exponentially.

For stable distributions, the CLT holds in the same form as for inverse power-law distributions. In addition, the functions in the domain of attraction of a stable law of index $\alpha < 2$ are characterized by the same tail index. This means that a distribution G belongs to the domain of

attraction of a stable law of parameter $\alpha < 2$ if and only if its tail decays as α . In particular, Pareto's law belongs to the domain of attraction of stable laws of the same tail index.

EXTREME VALUE THEORY FOR IID PROCESSES

In this section we introduce a number of important probabilistic concepts that form the conceptual basis of extreme value theory (EVT). The objective of EVT is to estimate the entire tail of a distribution from a finite sample by fitting to an appropriate distribution those values of the sample that fall in the tail. Two concepts play a crucial role in EVT: (1) the behavior of the upper order statistics (i.e., the largest k values in a sample) and, in particular, of the sample maxima; and (2) the behavior of the points where samples exceed a given threshold. We will explore the limit distributions of maxima and the distribution of the points of exceedances of a high threshold. Based on these concepts a number of estimators of the tail index in sequences of independent and identically distributed (IID) variables are presented.

Maxima

In the previous sections we explored the behavior of sums. The key result of the theory of sums is that the behavior of sums simplifies in the limit of properly scaled and centered infinite sums regardless of the shape of individual summands. If sums converge, their limit distributions can only be stable distributions. In addition, the normalized sums of finite-mean, finite-variance variables always converge to a normal variable.

A parallel theory can be developed for maxima, informally defined as the largest value in a sample. The limit distribution of maxima, if it exists, belongs to one of three possible distributions: *Frechet*, *Weibull*, or *Gumbel*. This result forms the basis of classical EVT. Each limit distribution of maxima has its own maximum domain of attraction. In addition, limit laws are

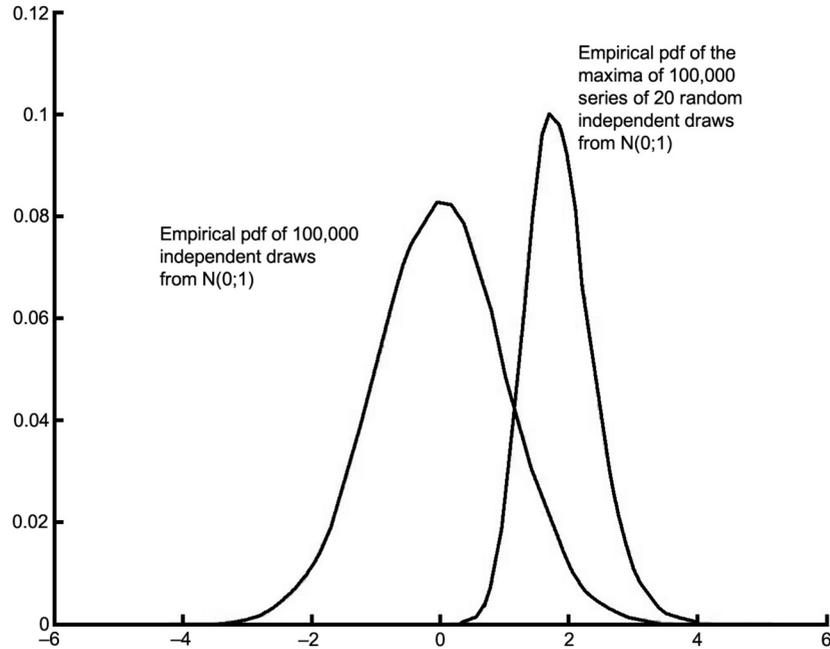


Figure 1 The Distribution of the Maxima of a Normal Variable

max-stable (i.e., they are closed with respect to maxima). However, the behavior of maxima is less robust than the behavior of sums. Maxima do not converge to limit distributions for important classes of distributions, such as Poisson or geometric distributions.

Consider a sequence of independent variables X_i with common, nondegenerate distribution F and the maxima of samples extracted from this sequence:

$$M_1 = X_1, M_n = \max(X_1, \dots, X_n)$$

The maxima M_n form a new sequence of random variables, which are not, however, independent.

As the variables of the sequence X_i are assumed to be independent, the distribution F_n of the maxima M_n can be immediately written down:

$$F(x)_n = P(X_1 \leq x \vee \dots \vee X_n \leq x) = F^n(x)$$

where \vee is the logical symbol for *and*.

If the distribution F , which is a nondecreasing function, reaches 1 at a finite point x_F —that is,

if $x_F = \sup\{x: F(x) < 1\} < \infty$, then

$$\lim_{n \rightarrow \infty} P(M_n < x) = \lim_{n \rightarrow \infty} F_n(x) = 0, \text{ for } x < x_F$$

If x_F is finite,

$$P(M_n < x) = F_n(x) = 1, \text{ for } x > x_F$$

The point x_F is called the right endpoint of the distribution F .

Figure 1 illustrates the behavior of maxima in the case of a normal distribution. Given a normal distribution with mean zero and variance one, 100,000 samples of 20 elements each are selected. For each sample, the maximum is chosen. The distribution of the maxima and the empirical distribution of independent draws from the same normal are illustrated in the figure.

A deeper understanding of the behavior of maxima can be obtained considering sequences of normalized and centered maxima. Consider the following sequence: $c_n^{-1}(M_n - d_n)$ where $c_n > 0$, $d_n \in \mathbb{R}$ are constants.

A fundamental result on the behavior of maxima is the Fisher-Tippett theorem, which can be stated as follows. Consider a sequence of IID

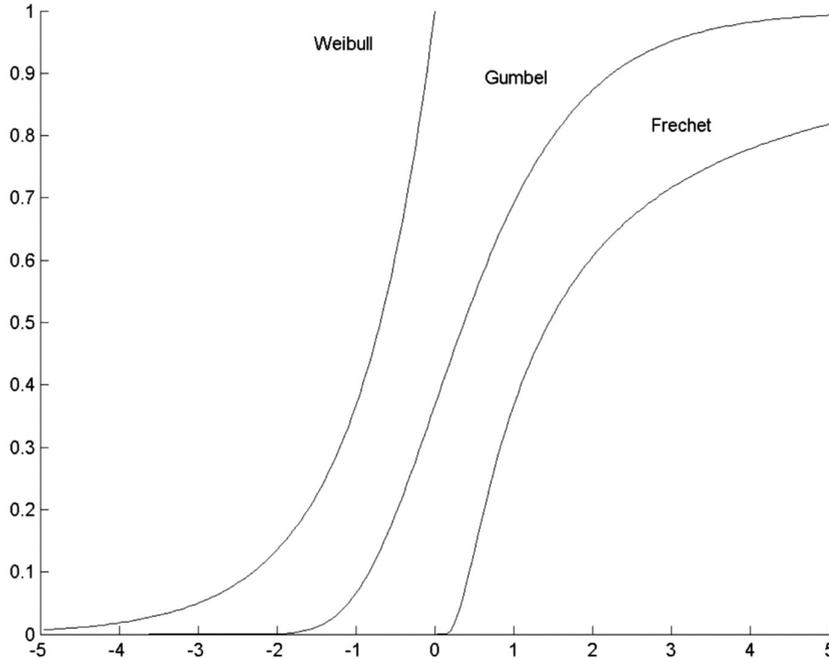


Figure 2 The Distribution of Fréchet, Gumbel, and Weibull

variables X_i and the relative sequence of maxima M_n . If there exist two sequences of constants $c_n > 0, d_n \in R$ and a nondegenerate distribution function H such that

$$c_n^{-1}(M_n - d_n) \xrightarrow{D} H$$

then H is one of the following distributions:

Fréchet: $\Phi_\alpha(x) = \begin{cases} 0 & x \leq 0 \\ \exp(-x^{-\alpha}) & x > 0 \end{cases} \quad \alpha > 0$

Weibull: $\Psi_\alpha(x) = \begin{cases} \exp[-(-x)^{-\alpha}] & x < 0 \\ 1 & x \geq 0 \end{cases} \quad \alpha > 0$

Gumbel: $\Lambda(x) = \exp\{-e^{-x}\}, x \in R$

The limit distribution H is unique, in the sense that different sequences of normalizing constants determine the same distribution.

The three above distributions—Fréchet, Weibull, and Gumbel—are called *standard extreme value distributions*. They are continuous functions for every real x . Random variables distributed according to one of the extreme value distributions are called *extremal random variables*.

As an example, consider a standard exponential variable X . As $F(x) = P(X \leq x) = 1 - e^{-x}, x \geq 0$ the distribution of the maxima is $P(M_n \leq x) = F^n(x) = (1 - e^{-x})^n, x \geq 0$. If we choose $d_n = \ln n$, we can write: $P(M_n - d_n \leq x) = P(M_n \leq \ln n + x) = (1 - n^{-1}e^{-x})^n, x \geq 0$. For any given $x, (1 - n^{-1}e^{-x})^n \rightarrow \exp(-e^{-x})$, which shows that the maxima of standard exponential variables centered with $d_n = \ln n$ tend to a Gumbel distribution. Figure 2 illustrates the three distributions: Fréchet, Gumbel, and Weibull.

We can now ask if there are conditions on the distribution F that ensure the existence of centering and scaling constants and the convergence to an extreme value distribution. To this end, let's first introduce the concept of the maximum domain of attraction (MDA) of an extreme value distribution H or $MDA(H)$.

A random variable X is said to belong to the $MDA(H)$ of the extreme value distribution H if there exist constants $c_n > 0, d_n \in R$ such that

$$c_n^{-1}(M_n - d_n) \xrightarrow{D} H$$

Two distribution functions F, G are said to be tail equivalent if they have the same right endpoints and the following condition holds:

$$\lim_{x \rightarrow \infty} \frac{\bar{F}(x)}{\bar{G}(x)} = c, 0 < c < \infty$$

Tail equivalence is an important concept for characterizing MDAs. In fact, it can be demonstrated that every $MDA(H)$ is closed with respect to tail equivalence (i.e., if two distribution functions F and G are tail equivalent $F \in MDA(H)$ if and only if $G \in MDA(H)$). Tail equivalence allows for a powerful characterization of the three MDAs.

Let's first define the quantile function. Given a distribution function F , the quantile function of F , written $F^{\leftarrow}(x)$, is defined as follows:

$$F^{\leftarrow}(x) = \inf\{s \in R : F(s) \geq x\}, 0 < x < 1$$

The MDA of the Frechet Distribution

The Frechet distribution is written as $\Phi_{\alpha}(x) = \exp(-x^{-\alpha})$. Let's start by observing that the tail of the Frechet distribution decays as an inverse power law. In fact, we can write $1 - \Phi_{\alpha}(x) = 1 - \exp(-x^{-\alpha}) \approx x^{-\alpha}$ for $x \rightarrow \infty$.

It can be demonstrated that a distribution function F belongs to the MDA of a Frechet distribution $\Phi_{\alpha}(x)$, $\alpha > 0$ if and only if there is a slowly varying function L such that $\bar{F}(x) = x^{-\alpha}L(x)$. In this case, the constants assume the values

$$c_n = (1/F^{\leftarrow})(n), d_n = 0$$

We can rewrite this condition more compactly as follows:

$$F \in MDA(\Phi_{\alpha}) \Leftrightarrow \bar{F} \in R_{-\alpha}$$

From the above definitions it can be demonstrated that the following five distributions belong to the MDA of the Frechet distribution: (1) Pareto; (2) Cauchy; (3) Burr; (4) stable laws with exponent $\alpha < 2$; or (5) log-gamma distribution.

The MDA of the Weibull Distribution

The Weibull distribution is written as follows:

$$\Psi_{\alpha} = \exp[-(-x^{-\alpha})]$$

The Weibull and the Frechet distributions are closely related to each other. In fact, it is clear from the definition that the following relationship holds:

$$\Psi_{\alpha}(x) = \Phi_{\alpha}(-x^{-1}), x > 0$$

One can therefore expect that the MDA of the two distributions are closely related. In fact, it can be demonstrated that a distribution function F belongs to the MDA of a Weibull distribution $\alpha > 0$ if and only if

$$x_F < \infty$$

and

$$\bar{F}(x_F - x^{-1}) = x^{-\alpha}L(x)$$

where L is a slowly varying function.

If

$$F \in MDA(\Psi_{\alpha})$$

then

$$c_n^{-1}(M_n - x_F) \xrightarrow{D} \Psi_{\alpha}$$

The MDA of the Weibull distribution includes important distributions such as the distribution uniform in $(0,1)$, power laws truncated to the right, and beta distributions.

The MDA of the Gumbel Distribution

The Gumbel distribution is written as $\Lambda(x) = \exp[-\exp(-x)]$. Observe that the Gumbel distribution has exponential tails. This fact can be easily ascertained through Taylor expansion. There is no simple characterization of the MDA of the Gumbel distribution.

The MDA of a Gumbel distribution encompasses a large class of distributions that includes the exponential distribution, the normal distribution, and the lognormal distribution. Though the Gumbel distribution has exponential tails, its MDA includes subexponential

distributions such as the Berkander distribution, as explained in Goldie and Resnick (1988).

Max-Stable Distributions

Stable distributions remain unchanged after summation; *max-stable distributions* remain unchanged after taking maxima. A nondegenerate random variable X and the relative distribution is called max-stable if there are constants $c_n > 0$, $d_n \in R$ such that the following conditions are satisfied

$$\max(X_1, \dots, X_n) \stackrel{D}{=} c_n X + d_n$$

where X, X_1, \dots, X_n are IID variables.

It can be demonstrated that the class of max-stable distributions coincides with the class of possible limit laws for normalized and centered maxima. In view of the previous discussions, the max-stable laws are the three possible limit laws: Frechet, Weibull, and Gumbel.

Generalized Extreme Value Distributions

The three extreme value distributions, Frechet, Weibull, and Gumbel, can be represented as a one-parameter family of distributions through the standard generalized extreme value distribution (GEV) of Jenkinson and Von Mises. Define the distribution function H_ξ as follows:

$$H_\xi = \begin{cases} \exp[-(1 + \xi x)^{-1/\xi}] & \text{for } \xi \neq 0 \\ \exp(-\exp(-x)) & \text{for } \xi = 0 \end{cases}$$

where $1 + \xi x > 0$. One can see from the definition that $\xi = \alpha^{-1} > 0$ corresponds to the Frechet distribution, $\xi = 0$ corresponds to the Gumbel distribution, and $\xi = -\alpha^{-1} < 0$ corresponds to the Weibull distribution. We can now introduce the related location-scale dependent family $H_{\xi;\mu,\psi}$ by replacing the argument x with $(x - \mu)/\psi$.

Order Statistics

The behavior of *order statistics* is a useful tool for characterizing fat-tailed distributions. For instance, the famous Zipf's law is an example of the behavior of order statistics. Consider a sample X_1, \dots, X_n made of n independent draws from the same distribution F . Let's arrange the sample in decreasing order:

$$X_{n,n} \leq \dots \leq X_{1,n}$$

The random variable $X_{k,n}$ is called the k th upper order statistic. It can be demonstrated that the distribution of the k th upper order statistic is

$$F_{k,n} = P(X_{k,n} < x) = \sum_{r=0}^{k-1} \bar{F}^r(x) F^{n-r}(x)$$

In addition, if F is continuous, it has a density with respect to F such that

$$f_{k,n} = \int_{-\infty}^x f_{k,n}(z) dF(z)$$

where

$$f_{k,n} = \frac{n!}{(k-1)!(n-k)!} \bar{F}^{k-1}(x) F^{n-k}(x)$$

The differences between two consecutive variables in a sample $X_{k,n} - X_{k+1,n}$ are random variables called spacings. In the case of variables with finite right endpoint x_F the zero-th spacing is defined as: $X_{0,n} - X_{1,n} = x_F - X_{1,n}$. The distribution of spacings depends on the distribution F . For instance, it can be demonstrated that the spacings of an exponential random variable are independent, exponential random variables with mean $1/n$ for an n -sample. Spacings are a key concept for the definition of the Hill estimator, as explained later in this section.

Another key concept, which is related to spacings, is that of quantile transformation. Let X_1, \dots, X_n be IID variables with distribution function F and let U_1, \dots, U_n be IID variables uniformly distributed on the interval $(0,1)$. Recall that, given a distribution function F , the

quantile function of F , written $F^{\leftarrow}(x)$, is defined as follows:

$$F^{\leftarrow}(x) = \inf\{s \in R : F(s) \geq x\}, \quad 0 < x < 1$$

It can be demonstrated that the following results hold:

- $F^{\leftarrow}(U_1) \stackrel{D}{=} X_1$
- $(X_{1,n}, \dots, X_{n,n}) \stackrel{D}{=} [F^{\leftarrow}(U_{1,n}), \dots, F^{\leftarrow}(U_{n,n})]$
- The random variable $F(X_1)$ has a uniform distribution on $(0,1)$ if and only if F is a continuous function.

To appreciate the importance of the quantile transformation, let's introduce first the notion of *empirical distribution function* and second the Glivenko-Cantelli theorem. The empirical distribution function F_n of a sample X_1, \dots, X_n is defined as follows:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x)$$

where I is the indicator function. In other words, for each x , the empirical distribution function counts the number of samples that are less than or equal to x .

The Glivenko-Cantelli theorem provides the theoretical underpinning of nonparametric statistics. It states that, if the samples X_1, \dots, X_n are independent draws from the distribution F , the empirical distribution function F_n tends to F for large n in the sense that

$$\Delta_n = \sup_{x \in R} |F_n(x) - F(x)| \xrightarrow{a.s.} 0, \quad \text{for } n \rightarrow \infty$$

The quantile transformation tells us that in cases where F is a Pareto distribution, if we approximate n random draws from a uniformly distributed variable as the sequence $1, 2, \dots, n$, then the corresponding values of the sample X_1, \dots, X_n will be

$$\frac{1}{1}, \frac{1}{2}, \dots, \frac{1}{n}$$

which is a statement of Zipf's law.

From the quantile transformation, the limit law of the ratio between two successive order statistics can also be inferred. Suppose that

an (infinite) population is distributed according to a distribution $F \in \mathfrak{N}(\alpha)$ with regularly varying tails. Suppose that n samples are randomly and independently drawn from this distribution and ordered in function of size: $X_{n,n} \geq X_{n-1,n} \geq \dots \geq X_{1,n}$. It can be demonstrated that the following property holds:

$$\frac{X_{k,n}}{X_{k+1,n}} = 1, \quad \frac{k}{n} \rightarrow 0$$

Point Process of Exceedances or Peaks over Threshold

We have now reviewed the behavior of sums, maxima, and upper order statistics of continuous random variables. Yet another approach to EVT is based on point processes; herein we will use point processes only to define the point process of exceedances.

Point processes can be defined in many different ways. To illustrate the mathematics of point processes, let's first introduce the homogeneous Poisson process. A *homogeneous Poisson process* is defined as a process $N(t)$ that starts at zero, i.e., $N(0) = 0$, and has independent stationary increments. In addition, the random variable $N(t)$ is distributed as a Poisson variable with parameter λt . $N(t)$ is therefore a time-dependent discrete variable that can assume nonnegative integer values. Figure 3 illustrates the distribution of a Poisson variable.

A homogeneous Poisson process can also be defined as a random sequence of points on the real line. Consider all discrete sequences of points on the real line separated by random intervals. Intervals are independent random variables with exponential distribution. This is the usual definition of a Poisson process. Call $N(t)$ the number of points that fall in the interval $[0, t]$. It can be demonstrated that $N(t)$ is a homogeneous Poisson process according to the previous definition.

This latter definition can be generalized to define point processes. Intuitively, a generic point process is a random collection of discrete points

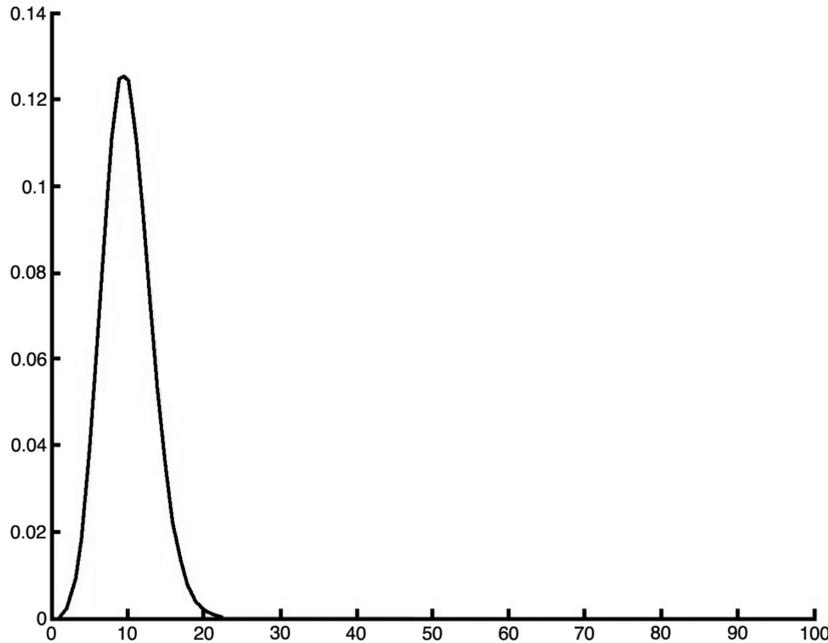


Figure 3 Distribution of a Poisson Variable

in some space. From a mathematical point of view, it is convenient to describe a point process through the distribution of the number of points that fall in an arbitrary set.⁶ In the case of homogeneous Poisson processes, we consider the number of points that fall in a given interval; for a generic point process, it is convenient to consider a wider class of sets.

Consider a subspace E of a finite dimensional Euclidean space of dimension n . Consider also the σ -algebra \mathfrak{B} of the Borel sets generated by open sets in E . The space E is called the state space. For each point x in E and for each set $A \in \mathfrak{B}$, define the *Dirac measure* ε_x as

$$\varepsilon_x = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

For any given sequence $x_i, i \geq 1$ of points in E , define the following set function:

$$m(A) = \sum_{i=1}^{\infty} \varepsilon_{x_i}(A) = \text{card}\{i : X_i \in A\}, A \in \mathfrak{B}$$

It can be verified that $m(A)$ is a measure \mathfrak{B} , called a counting measure. If a counting mea-

sure is finite on each compact set, then it is called a point measure. In other words, any given countable sequence in E generates a counting measure on \mathfrak{B} .

A point process is obtained associating to each family of sets $A_i \in \mathfrak{B}$ the joint probability distributions:

$$\Pr\{m(A_i) = n_i; i = 1, 2, \dots, k; k = 1, 2, \dots\}$$

To make this definition mathematically rigorous, a point process can be defined as a measurable map from some probability space to the set of all point measures equipped with an appropriate σ -algebra. Besides the mathematical details, it should be clear that point processes are defined by the probability distribution of the number of points that fall in each set A of some σ -algebra. The key ingredients of point processes are (1) counting measures that associate to each set A the number of points of each discrete sequence that falls in A with the additivity restrictions of measures and (2) probability distributions defined over the space of counting measures.

Equipped with the general concept of point processes, we can now define the *point process of exceedances*. Consider a threshold formed by any real number u and a sequence of random variables $X_i, i = 1, 2, \dots$. The point process of exceedances with state space $E = (0,1)$ counts the number of instances where the random variables X_i exceed the threshold u :

$$N_n(A) = \sum_{i=1}^{\infty} \varepsilon_{i/n}(A) = \text{card}\{i \leq n \text{ and } X_i > u\}$$

Note that in this case the state space specifies the size of the sample.

Estimation

In the previous sections we presented some key topics related to the probability structure of the tails of distributions, be they light- or fat-tailed. Let's now turn to the problem of estimation, which is the key practical task. The problem of estimation for EVT is essentially the problem of estimating the tail of a distribution from a finite sample. The key statistical idea of EVT from the point of view of estimation is to use only those sample data that belong to the tail and not the entire sample. This notion has to be made precise by finding criteria that allow one to separate the tail from the bulk of the distribution. Therefore, the estimation problem of EVT distribution can be broken down into three separate subproblems:

- Identify the beginning of the tail.
- Identify the shape of the tail, in particular discriminate if it is a power-law tail.
- Estimate the tail parameters, in particular the tail index in the case of a power-law tail.

It turns out that these three problems cannot be easily separated. In fact, there is no reliable constructive theory for solving all these problems automatically. In particular, the choice of the statistical model (i.e., the distribution that best describes data) is a classical problem of formulating and validating a scientific hypothesis in a probabilistic context. However, there

are many tools and tests to help the modeler in this endeavor.

The first fundamental tool is the graphical representation of data, in particular the *quantile plot* or *QQ-plot* defined as the following set:

$$\left\{ X_{k,n}, F \leftarrow \left(\frac{n-k+1}{n+1} \right) : k = 1, 2, \dots, n \right\}$$

The quantile transformation and the Glivenko-Cantelli theorem allow concluding that this plot must be approximately linear. Should F be a Pareto distribution, the linearity of the QQ-plot is another statement of Zipf's law. The quantile plot allows a quick verification of a statistical hypotheses by checking the approximate linearity of the plot. It also allows the modeler to form a preliminary opinion on where the tail begins and whether the model fails at the far end of the tail.

Though invaluable as an exploratory tool, graphics rely on human judgment and intuition. Rigorous tests are needed. A starting point is parameter estimation for the generalized extreme value (GEV) distribution that we write as

$$H_{\xi;\mu,\Psi}(x) = \exp \left\{ - \left(1 + \xi \frac{x - \mu}{\Psi} \right)^{-1/\xi} \right\},$$

$$1 + \xi \frac{x - \mu}{\Psi} > 0$$

with the convention that the case $\xi = 0$ corresponds to the Gumbel distribution:

$$H_{0;\mu,\Psi}(x) = \exp \left\{ -e^{-\frac{x-\mu}{\Psi}} \right\}, \quad x \in R$$

We saw above that these distributions are the limit distributions, if they exist, of the normalized maxima of IID sequences. Suppose that the data to be estimated are independent draws from some EGV. This is a rather strong assumption that we will progressively relax. This assumption might be justified in domains where long series of data are available so that the sample data are the maxima of blocks of consecutive data. Though this assumption is probably too strong in the domain of finance, it is useful to elaborate its consequences.

Standard methodologies exist for parameter estimation in this case. In particular, the usual maximum likelihood (ML) methodology can be used for fitting the best GEV to data. Note that if the above distributions fit maxima we have to divide data into blocks and consider the maxima of each block. To apply ML, we have to compute the likelihood function on the data and choose the parameters that maximize it. This can be done with numerical integration methods.

An estimation method alternative to ML is the method of moments, which consists in equating empirical moments with theoretical moments. An ample literature on various versions of the method of moments exists.⁷

Let's now release the assumption that the sequence of empirical data are independent draws from an exact GEV and replace this with the weaker assumption that empirical data are independent draws from $F \in \text{MDA}(H_\xi)$. If we assume that the limit distribution is a Frechet distribution, then data must be independent draws from some distribution F whose tail has the form:

$$\bar{F} = x^{-\alpha}L(x)$$

where L is a slowly varying function as described earlier in this entry. For this reason, estimation under this weaker assumption is semiparametric in nature. We will now introduce a number of estimators of the shape parameter ξ .

The Pickand Estimator

The *Pickand estimator* $\hat{\xi}_{k,n}^{(P)}$ for an n -sample of independent draws from a distribution $F \in \text{MDA}(H_\xi)$ is defined as

$$\hat{\xi}_{k,n}^{(P)} = \frac{1}{\ln 2} \ln \frac{X_{k,n} - X_{2k,n}}{X_{2k,n} - X_{4k,n}}$$

where the $X_{k,n}$ are upper order statistics.

It can be demonstrated that the Pickand estimator has the following properties:

- Weak consistency:

$$\hat{\xi}_{k,n}^{(P)} \xrightarrow{P} \xi, n \rightarrow \infty, k \rightarrow \infty, \frac{k}{n} \rightarrow 0$$

- Strong consistency:

$$\hat{\xi}_{k,n}^{(P)} \xrightarrow{a.s.} \xi, n \rightarrow \infty, \frac{k}{\ln(\ln n)} \rightarrow \infty, \frac{k}{n} \rightarrow 0$$

- Asymptotic normality under technical conditions.

The Pickand estimator is an estimator of the parameter ξ that does not require any assumption on the type of limit distribution. Let's now examine the Hill estimator, which requires the prior knowledge that sample data are independent draws from a Frechet distribution. Later in this entry we will see that the assumption of independence can be weakened.

The Hill Estimator

Suppose that X_1, \dots, X_n are independent draws from a distribution $F \in \text{MDA}(\Phi_\alpha)$, $\alpha > 0$ so that $\bar{F} = x^{-\alpha}L(x)$ where L is a slowly varying function. The *Hill estimator* can be obtained as an MLE based on the k upper order statistics. The Hill estimator takes the following form:

$$\hat{\alpha}^{(H)} = \hat{\alpha}_{k,n}^{(H)} = \left(\frac{1}{k} \sum_{j=1}^k \ln X_{j,n} - \ln X_{k,n} \right)^{-1}$$

The Hill estimator has the same weak and strong consistency property as well as asymptotic normality as the Pickand estimator. The Hill estimator is by far the most popular estimator of the tail index. It has the advantage of being robust to some dependency in the data but can perform very poorly in case of deviations from strict Pareto behavior. In addition, it is subject to a bias-variance trade-off in the following sense: The variance of the Hill estimator depends on the ratio k/n : It decreases for increasing k . However, using a large fraction of the data will introduce bias in the estimator.

As stated above, a critical tenet of EVT is the idea of fitting the tail rather than the entire distribution. A number of articles on the automatic

determination of the optimal subset of samples to be included in the tail have appeared. One approach to the automatic determination of the tail sample using the variance-bias trade-off was proposed by Drees and Kaufmann (2000), while Dacorogna et al. (1995), and Danielsson and de Vries (1977) proposed methods based on a bootstrap approach.

The *moment ratio estimator* is a generalization of the Hill estimator. Consider the following estimator of the second order moments of the k upper order statistic:

$$\hat{M}_{k,n} = \frac{1}{k} \left(\sum_{j=1}^k \ln X_{j,n} - \ln X_{k+1,n} \right)^2$$

The moment ratio estimator is defined as follows:

$$\hat{\alpha}_{k,n}^{(m)} = \frac{1}{2} \left(\frac{\hat{M}_{k,n}}{\hat{\alpha}_{k,n}^{(H)}} \right)$$

Wagner and Marsh (2000) did extensive simulation analysis of various estimators. Their finding is that the moment ratio estimator outperforms the Hill estimator in sequences with a dependence structure (this is discussed further in the next section).

The Hill estimator was extended by Dekkers and de Haan (1989) to cover the entire range of shape parameters ξ . A number of other estimators have been proposed. In particular, under the assumption that financial data follow a stable process, estimation procedures based on regression analysis have been suggested. In fact, the assumption of stable behavior, or at least of exact Pareto tail, naturally leads to fitting a linear model in a logarithmic scale. There is an ample literature on this topic with a number of useful discussions, though empirical studies based on Monte Carlo simulations are still limited.⁸

The estimation methods reviewed above are based on the behavior of maxima and upper order statistics; another methodology uses the points of exceedances of high thresholds. Estimation methodologies based on the points of

exceedances require an appropriate model for the point process of exceedances that was defined in general terms previously in this entry.

ELIMINATING THE ASSUMPTION OF IID SEQUENCES

In the previous sections we reviewed a number of mathematical tools that are used to describe fat-tailed processes under the key assumption of IID sequences. In this section we discuss the implications of eliminating this assumption. However, in finance theory the assumption of stationary sequences of independent variables is only a first approximation; it has been challenged in several instances. Consider individual price time series. The autocorrelation function of returns decays exponentially and goes to near zero at very short-time horizons while the autocorrelation function of volatility decays only hyperbolically and remains different from zero for long periods. In addition, if we consider portfolios made of many securities, price processes exhibit patterns of cross correlations at different time-lags and, possibly, cointegrating relationships. These findings offer additional reasons to consider the assumption of serial independence as only a first approximation.

If we now consider the question of stationarity, empirical findings are more delicate. The nonstationarity that can be removed by differencing is easy to handle and does not present a problem. The critical issue is whether financial time series can be modeled with a single data generation process (DGP) that remains the same for the entire period under consideration or if the model must be modified. Consider, for instance, the question of structural breaks. At a basic level, structural breaks entail nonstationarity as the model parameters change with time and thus the finite-dimension distributions change with time. However, at a higher level one might try to model structural

changes, for instance through state-space models or *Markov switching models*. In this way, stationarity is recovered but at the price of a more complex, serially autocorrelated model.

EVT for multivariate models with complex patterns of serial correlations loses its generality and becomes model-dependent. One has to evaluate each model in terms of its behavior as regards extremes. In this section we will explore a number of models that have been proposed for modeling financial time series: ARCH and GARCH models and, more in general, state-space models. First, however, a number of methodological considerations are in order.

In the context of IID sequences, EVT tries to answer the question of how to estimate a distribution with heavy tails given only a limited amount of data. The model is the simplest (i.e., a sequence of IID variables) and the question is how to extrapolate from finite samples to the entire tail. In the context of IID distributions, conditional and unconditional distributions coincide. However, if we release the IID assumption, we have to specify the model and to estimate the entire model—not just the tail of one variable. Conditional and unconditional distributions no longer coincide. For instance, there are families of models that are conditionally normal and unconditionally fat-tailed.

Here difficulties begin as model estimation might be complex. In addition, estimation of some specific tail might not be the primary concern in model estimation. In the context of variables with a dependence structure, EVT can be thought of as a methodology to estimate the tails of the unconditional distribution, leaving aside the question of full model estimation.

An important methodological question is whether fat-tailedness is generated by the transformation of a sequence of zero-mean, finite variance IID variables (i.e., white noise) or whether innovations themselves have fat tails (i.e., so-called colored noise). For instance, as we

will see, GARCH models entail fat-tailed return distributions as the result of the transformation of white noise. On the other hand, one might want to estimate an autoregressive moving average (ARMA) model under the assumption of innovations with infinite variance.

Understanding how power laws and, more in general, fat tails are generated from normal variables has been a primary concern of econometrics and econophysics. Given the universality of power laws in economics, it is clearly important to understand how they are generated. These questions go well beyond the statistical analysis of heavy-tailed processes and involve questions of economic theories. Essentially, one wants to understand how the decisions of a large number of economic agents do not average out but produce cascading and amplification phenomena.

The law of large numbers tells that if individual processes are independent and have finite variance, then phenomena average out in aggregate and tend to an average limit. However, if individual processes have fat tails, phenomena do not average out even in the infinite limit. The weight of individual tails prevails and drives the aggregate process. Philip W. Anderson, the corecipient of the 1997 Nobel Prize in Physics, remarked:

Much of the real world is controlled as much by the "tails" of distributions as by means or averages: by the exceptional, not the mean; by the catastrophe, not the steady drip; by the very rich, not the "middle class." We need to free ourselves from "average" thinking. (Anderson, 1997)

When and if fat-tailed drivers exist, they control the ensemble to which they belong. But what generates these powerful drivers? Models that generate fat tails from standard normal innovations attempt to answer this question. Different types of models have been proposed. One such category of models is purely geometric and exploits mathematical theories such as percolation and random graph. Others exploit phenomena of dynamic nonlinear self-reinforcing cascades of events.

Percolation models are based on the well known mathematical fact that in regular spatial structures of nodes connected by links, a uniform density of links produces connected subsets of nodes whose size is distributed according to power laws. Percolation models are time-transversal models: They model aggregation at any given time. They might be used to explain how fat-tailed IID sequences are generated.

Dynamic financial econometric models exploit cascading phenomena due to nonlinearities, in particular multiplicative noise. In a deterministic setting, it is well known that nonlinear chaotic models generate sequences that, when analyzed statistically, exhibit fat-tailed distributions. The same happens when noise is subject to nonlinear transformation. In the next sections, we explore simple ARMA models, ARCH-GARCH models, subordinated models, and state-space models, all examples of dynamic financial econometric models.

Before doing this, however, let's go back to the question of estimation. As observed above, if variables are not IID but can be considered generated by a DGP, the question of estimation is no longer the estimation of a variable but that of estimating a model or a theory. The estimation of the eventual tail index is part of a larger effort. However, empirical data are a sequence of samples characterized by an unconditional distribution. One might want to understand if estimation procedures used for IID sequences can be applied in this more general setting. For instance, one might want to understand if tail-index estimators such as the Hill estimator can be used in the case of serially correlated sequences generated by a generic DGP.

From a practical standpoint, this question is quite important as one wants to estimate the tails even if one does not know exactly what model generated the sequence. Clearly, there is no general answer to this problem. However, the behavior of a number of estimators under different DGPs has been explored through simulation as explained in the following section.

Heavy-Tailed ARMA Processes

Let's first consider the infinite moving average representation of a univariate stationary series:

$$x_t = \sum_{i=0}^{\infty} h_i \varepsilon_{t-i} + m$$

under the assumption that innovations are IID α -stable laws of tail index α . By the properties of stable distributions it can be demonstrated that the finite-dimensional distributions of the process x are α -stable. However, restrictions on the coefficients need to be imposed. It can be demonstrated that a sufficient condition to ensure that the process x exists and is stationary is the following:

$$\sum_{i=0}^{\infty} |h_i|^\alpha < \infty$$

A general univariate ARMA(p,q) model is written as follows:

$$X_t = \sum_{i=1}^p \alpha_i X_{t-i} + \sum_{j=1}^q \alpha_j Z_{t-j}$$

where the Z are IID variables.

Using the lag operator— L —notation, L^i represents the variable at i lags, the ARMA(p,q) model is written as follows:

$$X_t = \sum_{i=1}^p L^i X_t + \sum_{j=1}^q L^j Z_t$$

The theory of ARMA processes can be carried over at least partially to cover the case of fat-tailed innovations. In particular, an ARMA(p,q) process with IID α -stable innovations admits a stationary, infinite moving average representation under the same conditions as in the classical finite-variance case. The coefficients of the moving average satisfy the condition

$$\sum_{i=0}^{\infty} |h_i|^\alpha < \infty$$

In the case of fat-tailed innovations, covariances and autocovariances lose their meaning. It can also be demonstrated, however, that the

empirical autocorrelation function is meaningful and is asymptotically normal. It can be demonstrated that maximum likelihood estimates can be extended to the infinite variance case, though through a number of ad hoc processes.

ARCH/GARCH Processes

The simplest ARCH model can be written as follows. Suppose that X is the random variable to be modeled, Z is a sequence of independent standard normal variables, and σ is a hidden variable. The ARCH(1) model is written as

$$\begin{aligned} X_t &= \sigma_t Z_t \\ \sigma_t^2 &= \beta + \delta X_{t-1}^2 \end{aligned}$$

This basic model was extended by Bollerslev (1989), who proposed the GARCH(p,q) model written as

$$\begin{aligned} X_t &= \sigma_t Z_t \\ \sigma_t^2 &= \beta + \sum_{i=1}^p \gamma_i \sigma_{t-i}^2 + \sum_{i=1}^q \delta_i X_{t-i}^2 \end{aligned}$$

The IID variables Z can be standard normal variables or other symmetrical, eventually fat-tailed, variables.

Let's first observe that model parameters must be constrained in order to guarantee the stationarity of the model. Stationarity conditions depend on each model. No general simple expression for the stationarity conditions is available.

Due to the multiplicative nature of noise, GARCH models are able to generate fat-tailed distributions even if innovations have finite variance. This fact was established by Kesten (1973). The tail index can be theoretically computed at least in the case GARCH(1,1). Suppose a GARCH(1,1) stationary process with Gaussian innovation is given. It can be demonstrated that

$$P(X > x) \approx \frac{c}{2} x^{-2\kappa}$$

where κ is the solution of an integral equation. In the generic p, q case, the return process is still fat-tailed but no practical way to compute the index from model parameter is known.

Subordinated Processes

Subordinated processes allow the time scale to vary. Subordinated models are, in a sense, the counterpart of stochastic volatility models insofar as they model the change in volatility by contracting and expanding the time scale. The first model was proposed by Clark (1973). Subordinated models have been extensively studied by Ghysels, Gouriou, and Josiak (1995).

Subordinated models can be applied quite naturally in the context of trading. Individual trades are randomly spaced. In modern electronic exchanges, the time and size of trades are individually recorded, thus allowing for accurate estimates of the distributional properties of inter-trades intervals. Consideration of random spacings between trades naturally leads to the consideration of subordinated models. Subordinated models generate unconditional fat-tailed distributions.

Markov Switching Models

The GARCH family of models is not the only family of serially correlated models able to produce fat tails starting from normally distributed innovations. State-space models and Markov-switching models present the same feature. The basic ideas of state-space models and Markov switching models is to split the model into two parts: (1) a regressive model that regresses the model variable over a hidden variable and (2) an autoregressive model that describes the hidden variables.

In its simplest linear form, a state-space model is written as follows:

$$\begin{aligned} X_t &= \alpha Z_t + \varepsilon_t \\ Z_t &= \beta Z_{t-1} + \delta_t \end{aligned}$$

where ε_t, δ_t are normally distributed independent white noises. State-space models can also be written in a multiplicative form:

$$\begin{aligned} X_t &= \alpha Z_{t-1} + \varepsilon_t \\ \alpha_t &= \beta \alpha_{t-1} + \delta_t \end{aligned}$$

If the second equation is a Markov chain, the model is called a Markov-switching model. A well-known example of Markov-switching models is the Hamilton model in which a two-state Markov chain drives the switch between two different regressions.

Purely linear state-space models exhibit fat tails only if innovations are fat-tailed. However, multiplicative state-space models and Markov-switching models can exhibit fat tails even if innovations are normally distributed. There is a growing literature on Markov-switching and multiplicative state-space models and a relatively large number of different models, univariate as well as multivariate, have been proposed. Stochastic volatility models are the continuous-time version of multiplicative state-space models.

Estimation

Let's now go back to the question of model estimation in a non-IID framework. Suppose that we want to estimate the tail index of the unconditional distribution of a set of empirical observations in the general setting of non-IID variables. Note that if variables are fat-tailed, we cannot say that they are serially autocorrelated as moments of second order generally do not exist. Therefore we have to make some hypothesis on the DGP.

There is no general theory of estimation under arbitrary DGP. Both theoretical and simulation work are limited to specific DGPs. ARMA models have been extensively studied. EVT holds for ARMA models under general nonclustering conditions.⁹

Often only simulation results are available. A fairly ample set of results are available for

GARCH(1,1) models. For these models Resnick and Starica (1998) showed that the Hill estimator is a consistent estimator of the tail index. Wagner and Marsh compared the performance of the Hill estimator and of the moment ratio estimator for three model classes: IID α -stable returns, IID symmetric Student, and GARCH(1,1) with Student-t innovation. They found that, in an adoptive framework, the moment ratio estimator generally yields results superior to the Hill estimator.

Scaling and Self-Similarity

The concept of scaling is now quite frequently evoked in economics and finance. Let's begin by making a distinction between scaling and self-similarity and some of the properties associated with inverse power laws within or outside the Levy-stable scaling regime. These concepts have different, and not equivalent, definitions.

The concepts of scaling and self-similarity apply to distributions, processes, or structures. Self-similarity was introduced as a property that applies to geometrical self-similar objects (i.e., fractal structures). In this context, self-similarity means that a structure can be put into a one-to-one correspondence with a part of itself. Note that no finite structure can have this property; self-similarity is the mark of infinite structures. Self-similarity entails scaling: If a fractal structure is expanded by a given factor, its measure expands by a power of the same factor.¹⁰ The notion of scaling is often expressed as absence of scale, meaning that a scaling object looks the same at any scale, large or small: It is impossible to ascertain the size of a portion of a scaling object by looking at its shape. The classical illustration is a Norwegian coastline with its fjords and fjords within fjords that look the same regardless of the scale.

However, scaling can be defined without making reference to fractals. In its simplest form, the notion of scaling entails a variable x and an observable A , which is a function of $A = A(x)$. If the observable obeys a scaling

relationship, there is a constant factor between x and A in the sense that $A(\lambda x) = \lambda^s A(x)$, where s is the scaling exponent that does not depend on x . The only function $A(x)$ that satisfies this relationship is a power law. In the three-dimensional Euclidean space, volume scales as the third power of linear length and surface as the second power, while fractals scale according to their fractal dimension.

The same ideas can be applied in a random context, but require careful reasoning. A power-law distribution has a scaling property as multiplying the variable by a factor multiplies probabilities by a constant factor, regardless of the level of the variable. This means that the ratio between the probability of the events $X > x$ and $X > ax$ depends only on a power of a , not on x . As an inverse power law is not defined at zero, scaling in this sense is a property of the tails. The probabilistic interpretation of this property is the following: The probability that an observation exceeds ax conditional on the knowledge that the observation exceeds x does not depend on x but only on a .

There are, however, other meanings attached to scaling and these might be a source of confusion. In the context of physical phenomena, scaling is often intended as identity of distribution after aggregation. The same idea is also behind the theory of groups of renormalization and the notion of self-similarity applied to structures such as coastlines. In the latter case, the intuitive meaning of self-similarity is that if one aggregates portions of the coastline, approximating their shape with a straight line, and then rescales, the resulting picture is qualitatively similar to the original. The same idea applies to percolation structures: By aggregating “sites” (i.e., points in a percolation lattice) into supersites and carefully redefining links, one obtains the same distribution of connected clusters.

Applying the idea of aggregation in a random context, self-similarity seems to mean that, after rescaling, the distribution of the sum of independent copies of a random variable maintains

the same shape of the distribution of the variable itself. Note that this property holds only for the tails of subexponential distributions—and it holds strictly only for stable laws that have tails in the (0,2) range but whose shape is not a power law except, approximately, in the tails. It also holds for Gaussian distributions that do not have power-law tails.

Scaling acquires yet another meaning when applied to stochastic processes that are functions of time. The most common among the different meanings is the following: A stochastic process is said to have a scaling property if there is no natural scale for looking at its paths and distributions. Intuitively, this means that it is not possible to gauge the scale of a sample by looking at its distribution; there is absence of scale. An example from finance comes from price patterns. If a price pattern is generated by a process with the scaling property, the plots of average daily and monthly prices will appear to be perfectly similar in distribution; looking at the plot, it’s impossible to tell if it refers to daily or monthly prices.

Self-similarity is another way of expressing the same concept. A process is self-similar if a portion of the process is similar to the entire process. As we are considering a random environment, self-similarity applies to distributions, not to the actual realization of a process. Let’s now make these concepts more precise.

A stochastic process $X(t)$ is said to be self-similar (ss) of index H (H-ss) if all its finite-dimensional distributions obey the scaling relationship:

$$(X_{kt_1}, X_{kt_2}, \dots, X_{kt_m}) \stackrel{D}{=} k^{-H}(X_{t_1}, X_{t_2}, \dots, X_{t_m}) \forall k > 0$$

$$0 < H < 1, t_1, t_2, \dots, t_m > 0$$

The above expression means that the scaling of time by the factor k scales the variables X by the factor k^H . It gives precise meaning to the notion of self-similarity applied to stochastic processes.

There is a wide variety of self-similar processes that cannot be characterized in a simple way as scaling laws: The scaling property of

stochastic processes might depend upon the shape of distributions as well as the shape of correlations. Let's restrict our attention to processes that are self-similar with stationary increments (sssi) and with index H (H -sssi). These processes can be either Gaussian or non-Gaussian. Note that a Gaussian process is a process whose finite-dimensional distributions are all Gaussian.

Gaussian H -sssi processes might have independent increments or exhibit long-range correlations. The only Gaussian H -sssi process with independent increment is the Brownian motion, but there are an infinite number of fractional Brownian motions, which are Gaussian H -sssi processes with long-range correlations. Thus there are an infinite variety of Gaussian self-similar processes. Among the many non-Gaussian H -sssi processes with independent increments are the stable Levy processes, which are random walks whose increments follow a stable distribution.¹¹

There is another definition of self-similarity for stochastic processes that makes use of the concept of aggregation; it is closer, at least in spirit, to the theory of renormalization groups. Consider a stationary infinite sequence of independent and identically distributed variables X_i , $i \geq 1$. Create consecutive nonoverlapping blocks of m variables and define the corresponding aggregated sequence of level m averaging over each block as follows:

$$X_k^{(m)} = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X_i$$

A sequence is called exactly self-similar if, for any integer m the following relationship holds:

$$X \stackrel{D}{=} m^{1-H} X^{(m)}$$

A stationary sequence is called asymptotically self-similar if the above relationship holds only for $m \rightarrow \infty$.

When we apply the notion of scaling to stochastic processes—the natural setting for economics and finance—we have to abandon

the simple characterization of scaling as inverse power laws. Though the scaling property is in itself characterized through simple power laws, the scaling processes are complex and rich mathematical structures entailing a variety of distributions and correlation functions. In particular, the long-range correlation structure of the process plays a role as important as the distribution of its variables.

KEY POINTS

- Fat-tailed laws have been found in many economic variables.
- Fully approximating a finite economic system with fat-tailed laws depends on an accurate statistical analysis of the phenomena, but also on a number of the theoretical implications of subexponentiality and scaling.
- Modeling financial variables with stable laws implies the assumption of infinite variance, which seems to contradict empirical observations.
- Scaling laws might still be an appropriate modeling paradigm given the complex interaction of distributional shape and correlations in price processes.
- Scaling laws might help in understanding not only the sheer size of economic fluctuations but also the complexity of economic cycles.

NOTES

1. See Bamberg and Dorfleitner (2001).
2. See, for example, Sigman (1999).
3. See, for example, Goldie and Kluppelberg (1998) and Embrechts, Kluppelberg, and Mikosch (1999).
4. See Sigman (1999).
5. See Rachev and Mittnik (2000) and Rachev, Menn, and Fabozzi (2005).
6. Cox and Isham (1980).
7. For a discussion of the different methods, see Smith (1990). For a discussion of the method of probability-weighted moments, see Hosking, Wallis, and Wood (1985).

8. Diebold, Schuermann, and Stroughair (2000).
9. See Embrechts, Kluppelberg, and Mikosch (1999).
10. For an introduction to fractals, see Falconer (1990).
11. See Samorodnitsky and Taqqu (1994).

REFERENCES

- Anderson, R. W. (1997). Some thoughts about distribution in economics. In W. B. Arthur, S. N. Durlaf, and D. A. Lane (eds.), *The Economy as an Evolving Complex System II*. Reading, MA: Addison-Wesley.
- Bamberg, G., and Dorfleitner, D. (2001). Fat tails and traditional capital market theory. Working Paper, University of Augsburg, August.
- Bollerslev, T. (1989). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics* 31: 307–327.
- Bryson, M. C. (1982). Heavy-tailed distributions. In N. L. Kotz and S. Read (eds.), *Encyclopedia of Statistical Sciences: Volume 3*. New York: John Wiley & Sons, 598–601.
- Chistyakov, V. P. (1964). A theorem on sums of independent positive random variables and its applications to branching random processes. *Theory Probability Applic* 9: 640–648.
- Clark, P. K. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica* 4: 735–755.
- Cox, D. R., and Isham, V. (1980). *Point Processes*. London: Chapman and Hall.
- Dacorogna, M. M., Muller, U. A., Pictet, O. V., and de Vries, C. G. (1995). The distribution of extremal foreign exchange rate returns in extremely large data sets. Olsen & Associates, Zurich.
- Danielsson, J., and de Vries, C. G. (1977). Tail index and quantile estimation with very high frequency data. *Journal of Empirical Finance* 4: 241–257.
- Dekkers, A. L. M., and de Haan, L. (1989). On the estimation of the extreme-value index and large quantile estimation. *Annals of Statistics* 17: 1795–1832.
- Diebold, F. X., Schuermann, T., and Stroughair, J. D. (2000). Pitfalls and opportunities in the use of extreme value theory in risk management. *Journal of Risk Finance*, Winter: 30–36.
- Drees, H., and Kaufmann, E. (2000). Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and Their Application* 75: 254–274.
- Embrechts, P., Kluppelberg, C., and Mikosch, T. (1999). *Modelling Extremal Events for Insurance and Finance*. Berlin: Springer.
- Falconer, J. (1990). *Fractal Geometry*. Chichester, U.K.: John Wiley & Sons.
- Goldie, C. M., and Resnick, S. (1988). Distributions that are both subexponential and in the domain of attraction of an extreme-value distribution. *Advanced Applied Probability* 20: 706–718.
- Goldie, C. M., and Kluppelberg, C. (1998). Subexponential distributions. In R. J. Adler, R. E. Feldman, and M. S. Taqqu (eds.), *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*. Boston: Birkhauser, 435–459.
- Ghysels, E., Gouriéroux, C., and Josiak, J. (1995). Market time and asset price movement theory and estimation. Working Paper 95–32, Cyrano, Montreal.
- Hosking, J. R. M., Wallis, J. R., and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics* 27: 251–261.
- Kesten, H. (1973). Random difference equations and renewal theory for products of random matrices. *Acta Mathematica* 131: 207–248.
- Rachev, S. T., Menn, C., and Fabozzi, F. T. (2005). *Fat-Tailed and Skewed Asset Return Distributions: Implications for Risk Management*. Hoboken, NJ: John Wiley & Sons.
- Rachev, S. T., and Mittnik, S. (2000). *Stable Pareitian Models in Finance*. Chichester: John Wiley & Sons.
- Resnick, S., and Starica, C. (1998). Tail index estimation for dependent data. *Annals of Applied Probability* 8: 1156–1183.
- Samorodnitsky, G., and Taqqu, M. S. (1994). *Stable Non-Gaussian Random Processes*. New York: Chapman & Hall.
- Sigman, K. (1999). A primer on heavy-tailed distributions. *Queueing Systems* 33: 261–275.
- Smith, R. L. (1990). Extreme value theory. In W. Ledermann (ed.), *Handbook of Applicable Mathematics, Supplement*. Chichester, U.K.: John Wiley & Sons: 437–472.
- Wagner, N., and Marsh, T. (2000). On adaptive tail index estimation for financial return models. Research Program in Finance, Working Paper RPF-295, Haas School of Management, University of California, Berkeley, November.

Copulas

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair-Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

CHRISTIAN MENN, Dr. rer. pol.

Managing Partner, RIVACON

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Understanding dependences or functional links between variables is a key theme in financial modeling. In general terms, functional dependences are represented by dynamic models. Many important models are linear models whose coefficients are correlations coefficients. In many instances in financial modeling, it is important to arrive at a quantitative measure of the strength of dependencies. The correlation coefficient provides such a measure. In many instances, however, the correlation coefficient might be misleading. In particular, there are cases of nonlinear dependencies that result in a zero correlation coefficient. From the point of view of financial modeling, this situation is particularly dangerous as it leads to substantially underestimated risk. Different measures of dependence have been proposed, in particular copula functions.

Correlation is a widespread concept in financial modeling and stands for a measure of dependence between random variables. However, this term is very often incorrectly used to mean any notion of dependence. Actually *correlation* is one particular measure of *dependence* among many. In the world of multivariate normal distribution and, more generally, in the world of spherical and elliptical distributions, it is the accepted measure. This follows from a property of the multivariate normal distribution. In this entry, we discuss the limitations of correlation as a measure of the dependence between two random variables and introduce an alter-

native measure to overcome these limitations, *copulas*.¹

DRAWBACKS OF CORRELATION

In the general case, there are at least three major drawbacks of the correlation measure. Consider the case of two real-valued random variables X and Y . First, the variances of X and Y must be finite or the correlation is not defined. This assumption causes problems when working with heavy-tailed data because under certain circumstances the variances are infinite

and, for that reason, the correlation between them is not defined.

Second, independence of two random variables implies correlation equal to zero; however, generally speaking the opposite is not correct—zero correlation does not imply independence.² Only in the case of elliptical distribution are uncorrelatedness and independence interchangeable notions. This statement is not valid if only the marginal distributions are elliptical and the joint distribution is nonelliptical.

Lastly, a more technical point. The correlation is not invariant under nonlinear strictly increasing transformations, a serious disadvantage. In general $\text{corr}(T(X), T(Y)) \neq \text{corr}(X, Y)$. One example that explains this technical requirement is the following: Assume that X and Y represent the continuous return (log-return) of two financial assets over the period $[0, t]$, where t denotes some point of time in the future. If you know the correlation of these two random variables, this does not imply that you know the dependence structure between the asset prices itself because the asset prices (P and Q for asset X and Y , respectively) are obtained by $P_t = P_0 \cdot \exp(X)$ and $Q_t = Q_0 \cdot \exp(Y)$. The asset prices are strictly increasing functions of the return but the correlation structure is not maintained by this transformation. This observation implies that the return could be uncorrelated whereas the prices are strongly correlated and vice versa.

OVERCOMING THE DRAWBACKS OF CORRELATION: COPULAS

A more prevalent approach, which overcomes this disadvantage, is to model dependency using copulas. As noted by Patton (2004, p. 3): “The word *copula* comes from Latin for a ‘link’ or ‘bond’, and was coined by Sklar (1959), who first proved the theorem that a collection of marginal distributions can be ‘coupled’ together via a copula to form a multivariate distribution.” The idea is as follows. The

description of the joint distribution of a random vector is divided into two parts:

1. The specification of the marginal distributions.
2. The specification of the dependence structure by means of a special function, called *copula*.

The use of copulas offers the following advantages:

- The nature of dependency that can be modeled is more general. In comparison, only linear dependence can be explained by the correlation.
- Dependence of extreme events might be modeled.
- Copulas are indifferent to continuously increasing transformations (not only linear as it is true for correlations).

Because of these advantages, in recent years there has been increased application of copulas in asset and option pricing, portfolio selection, and risk management.

MATHEMATICAL DEFINITION OF COPULAS

From a mathematical viewpoint, a copula function C is nothing more than a probability distribution function on the d -dimensional hypercube $I_d = [0, 1] \times [0, 1] \times \dots \times [0, 1]$:

$$C : I_d \rightarrow [0, 1] \\ (x_1, \dots, x_d) \rightarrow C(x_1, \dots, x_d)$$

It has been shown³ that any multivariate probability distribution function F_Y of some random vector $Y = (Y_1, \dots, Y_d)$ can be represented with the help of a copula function C in the following form:

$$F_Y(y_1, \dots, y_d) = P(Y_1 \leq y_1, \dots, Y_d \leq y_d) \\ = C(P(Y_1 \leq y_1), \dots, P(Y_d \leq y_d)) \\ = C(F_{Y_1}(y_1), \dots, F_{Y_d}(y_d))$$

where the F_{Y_i} , $i = 1, \dots, d$ denote the marginal distribution functions of the random variables Y_i , $i = 1, \dots, d$.

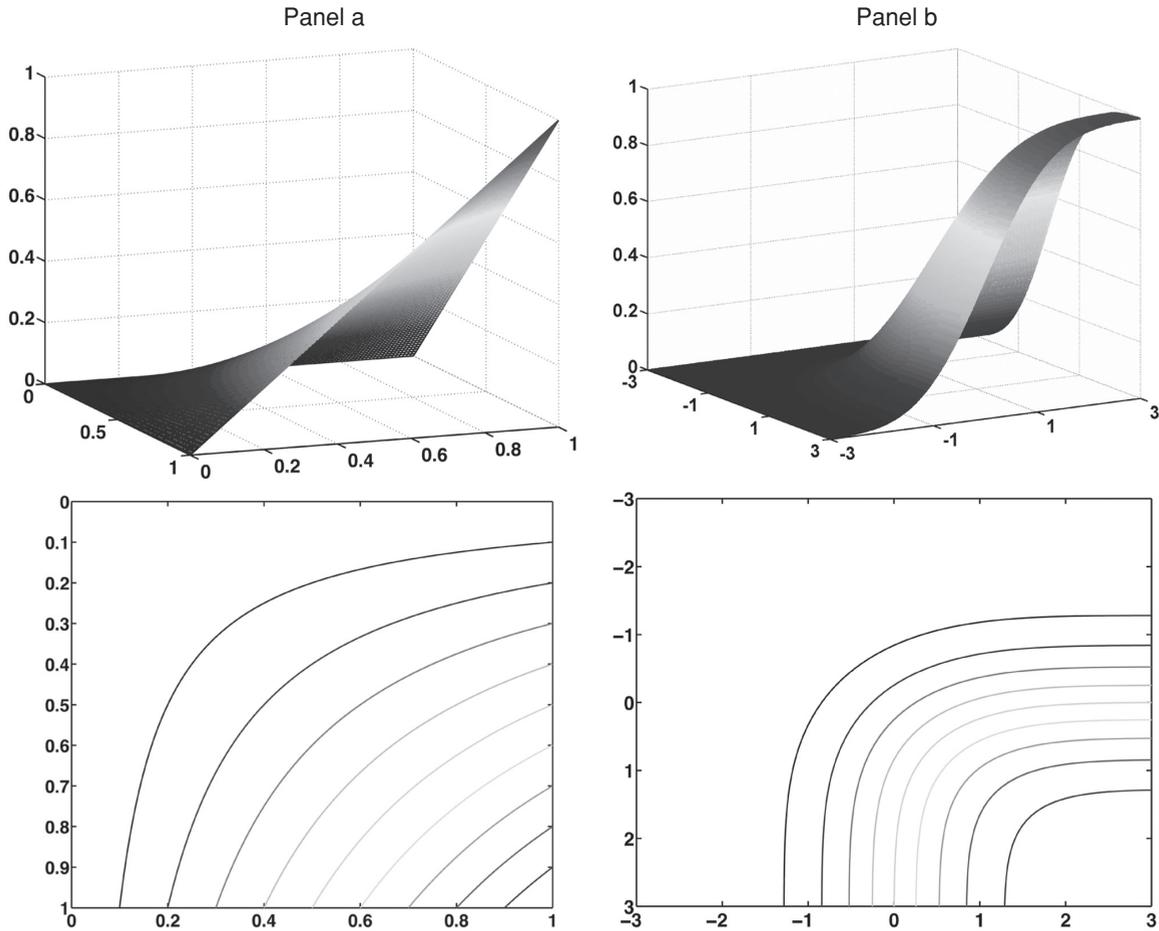


Figure 1 Visualization of the Copula for Bivariate Independence*
 Panel a: Uniform Marginal Distributions. Panel b: Standard Normal Marginal Distributions.
 *The graphs show the joint distribution function of a bivariate random vector for two different marginal distributions. Each panel consists of a surface and a corresponding contour plot.

The copula function makes the bridge between the univariate distribution of the individual random variables and their joint probability distribution. This justifies the fact that the copula function creates uniquely the dependence, whereas the probability distribution of the involved random variables is provided by their marginal distribution.

As an example we consider the following three bivariate copula functions:

- $C(x, y) = x \cdot y$
- $C(x, y) = \min(x, y)$

$$C(x, y) = \int_{-\infty}^{\Phi^{-1}(x)} \int_{-\infty}^{\Phi^{-1}(y)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left(\frac{s^2 - 2\rho st + t^2}{2(1-\rho^2)}\right) ds dt$$

The first represents the independent case as the joint probability distribution equals the product of their marginals. The second example represents a case of extreme dependence whereas the third example represents the general Gaussian copula function for the bivariate case.

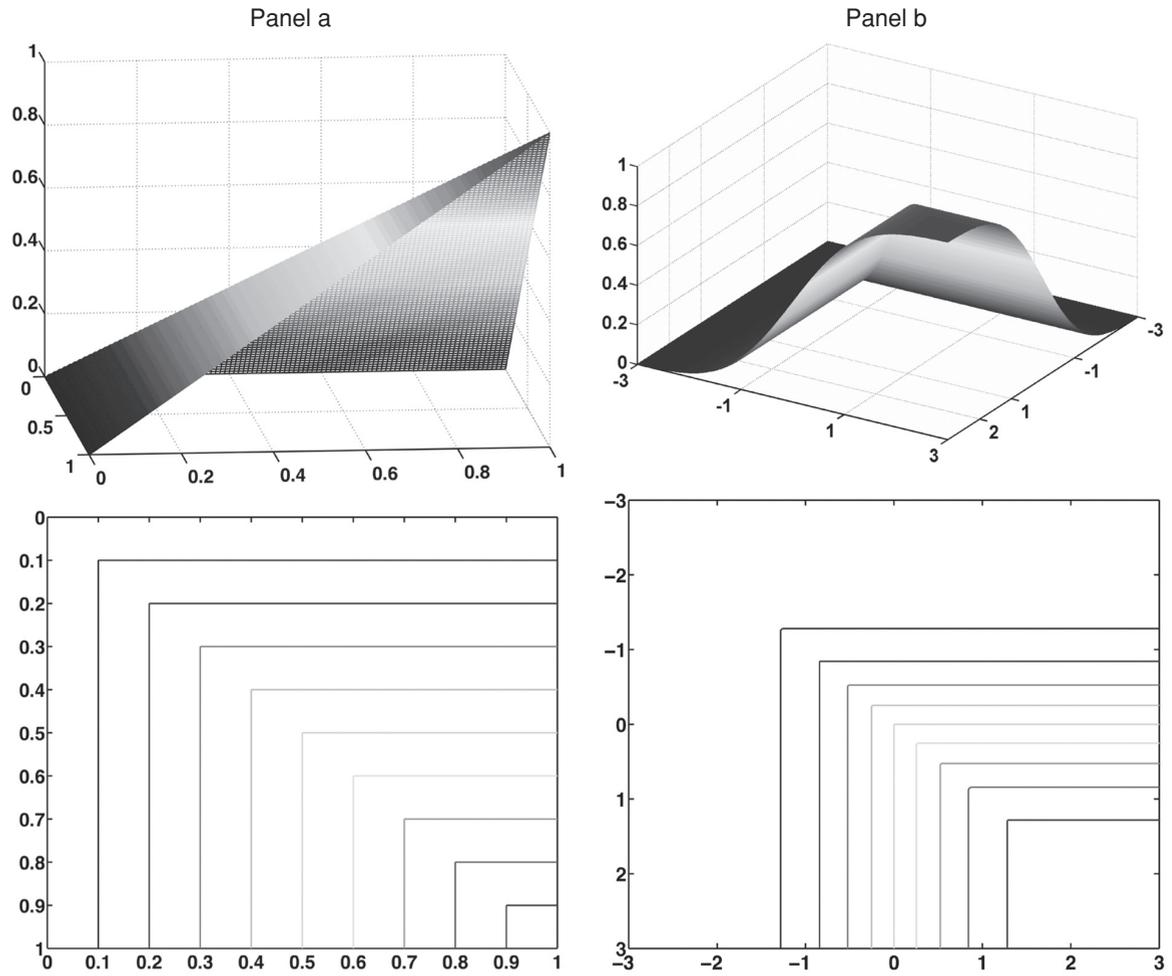


Figure 2 Visualization of the Bivariate Minimum Copula*

Panel a: Uniform Marginal Distributions. Panel b: Standard Normal Marginal Distributions.

*The graphs show the joint distribution function of a bivariate random vector for two different marginal distributions. Each panel consists of a surface and a corresponding contour plot.

We illustrate the effect of the different copulas by applying them to two different marginal distributions, namely (1) the uniform distribution on the interval $[0,1]$ and (2) the standard normal distribution. The results are presented in Figures 1, 2, and 3.

KEY POINTS

- In financial modeling, it is critical to understand dependencies or functional links be-

tween variables and have a quantitative measure of the strength of dependencies.

- The most commonly used measure of dependency in finance is the correlation coefficient. This measure might be misleading. In particular, there are cases of nonlinear dependencies that result in a zero correlation coefficient.
- The existence of finite variances is required for a correlation to be computed. Some return distributions, however, have fat tails, and the variances are infinite.

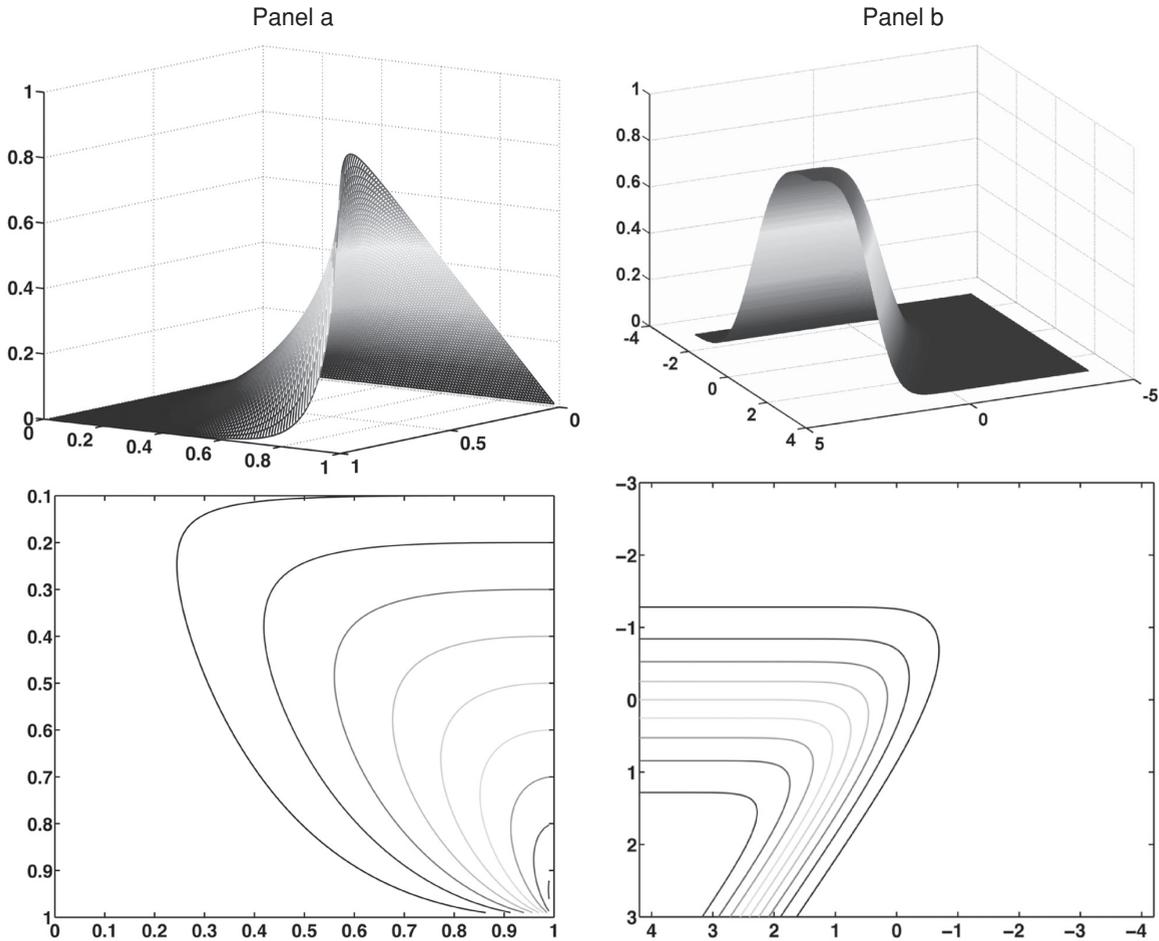


Figure 3 Visualization of the Gaussian Copula with Correlation $\rho = 0.8^*$
 Panel a: Uniform Marginal Distributions. Panel b: Standard Normal Marginal Distributions.
 *The graph shows the joint distribution function of a bivariate random vector for two different marginal distributions. Each panel consists of a surface and a corresponding contour plot.

- The correlation is not invariant under non-linear strictly increasing transformations, making the use of this measure a serious disadvantage.
- The copula overcomes the drawbacks of the correlation as a measure of dependency by allowing for a more general measure than linear dependence, allowing for the modeling of dependence for extreme events, and being indifferent to continuously increasing transformations.
- The copula function bridges the univariate distribution of the individual random vari-

ables and their joint probability distribution, thereby justifying the fact that the copula function creates the dependence uniquely, whereas the probability distribution of the involved random variables is provided by their marginal distribution.

NOTES

1. For a discussion of applications in finance and insurance, see Embrechts, McNeil, and Straumann (1999) and Patton (2003a, 2003b, 2004).

2. A simple example is the following: Let X be a standard normal distribution and $Y = X^2$. Because the third moment of the standard normal distribution is zero, the correlation between X and Y is zero despite the fact that Y is a function of X , which means that they are dependent.
3. The importance of copulas in the modeling of the distribution of multivariate random variables is provided by Sklar's theorem. The derivation was provided in Sklar (1959).

REFERENCES

- Embrechts, P., McNeil, A., and Straumann, D. (1999). Correlation and dependence properties in risk management: Properties and pitfalls. In *Risk Management: Value at Risk and Beyond*, ed. M. Dempster, 176–223. Cambridge: Cambridge University Press.
- Patton, A. J. (2003a). On the importance of skewness and asymmetric dependence for asset allocation. *Journal of Financial Econometrics* 2, 1: 130–168.
- Patton, A. J. (2003b). Estimation of copula models for time series of possibly different lengths. Working paper. London School of Economics, September.
- Patton, A. J. (2004). Modelling asymmetric exchange rate dependence. Working paper. London School of Economics, September.
- Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris* 8: 229–231.

Applications of Order Statistics to Risk Management Problems

RADU TUNARU, PhD

Professor of Quantitative Finance, Business School, University of Kent

Abstract: Value-at-risk (VaR) calculation based on parametric models is in essence an estimation problem. The point estimates should be interpreted accompanied by their confidence intervals. Risk management for complex portfolios may consider simultaneously two or more VaR confidence levels. The quantiles used for VaR estimation at different orders such as 1% and 5% are not independent and therefore should be analyzed jointly. Consequently, it would be useful to establish confidence regions for bivariate VaR estimates that will provide the risk managers with a valuable tool for verifying the accuracy of their estimation process, as requested by external audit. A trade-off between the complexity of probability distribution underlying the model and the degree of robustness achieved is recommended.

While there are many models used for calculations of risk management measures such as *value-at-risk* (VaR) and *expected tail loss* (ETL), there are not many tools available to a risk manager to verify whether the models chosen are very good in practice. In this entry, we highlight some practical aspects of VaR and ETL calculus that are underpinned by theoretical results on *order statistics*. More precisely, we show how to compute VaR and ETL based on quantile sample statistics and how to derive the probability distribution of this estimator. The most important development in this entry is that we illustrate how to control the backtesting of two risk measures, given by different specifications of confidence levels such as 99% and 95%. Usually there is a difference between the confidence level that a bank may use internally and the

confidence level required by a regulator. Then the risk manager should make sure that the *risk models* used perform well for both confidence levels.

PERFORMANCE OF VaR ESTIMATION

VaR is widely used in the financial industry as a measure for market risk in normal conditions. This concept has a strong influence on bank capital, some of the major implications of this estimation process being described in Jackson et al. (1997). The European Capital Adequacy Directive allows internal risk management models. Marshall and Siegel (1997) found

great errors in the estimation methods used in the industry. Berkowitz and O'Brien (2002) investigated the accuracy of value-at-risk models used by a sample of large commercial banks and their analysis revealed discrepancies in the performance of their models. Brooks and Persaud (2002) analyzed common methodologies for calculating VaR and concluded that simpler models provide better performance than very complex models. In the light of severe market disruptions and appeal for more stringent measures, the issue of how reliable is the model used for market risk is of paramount importance.

The estimation of VaR is a statistical exercise and the risk manager, trader, or quant analyst has to consider the reliability of the estimates proposed, especially when large amounts of money are involved. Although there is a plethora of models for VaR pointwise estimation, reviewed for example in Duffie and Pan (1997) and Jorion (1996, 1997), the literature on the confidence associated with these estimators is sparse. Jorion (1996) was among the first researchers to consider the uncertainty associated with VaR models leading to model risk. Kupiec (1995) suggested that it may be very hard to determine statistically the accuracy of VaR estimates. After his seminal paper, Pritsker (1997) and Dowd (2001) showed how to employ order statistics for assessing the VaR accuracy. Dowd (2000) described how to build confidence intervals for VaR estimates using simulations methods but his technique was illustrated only for some special cases linked to the Gaussian distribution.

Calibrating the models is not always easy and for auditing and backtesting purposes the pre-specified level of confidence can play an important role. The nonlinearity in results when calculating VaR at various levels of confidence means that, based on the same model, conclusions obtained in backtesting at one level cannot be extrapolated to other levels. In other words, we can have a model with very good forecasting power at 5% and quite bad results at 1%, or vice versa.

VaR AND DIFFERENT LEVELS OF CONFIDENCE

The starting point of VaR modeling is a time series Y_1, Y_2, \dots, Y_n of profit and loss observations (P/L); the time series consists of past returns or simulated returns. If the critical level (of confidence) for VaR is specified as α (e.g., 10%, 5%, 1%), for a given sample the VaR is determined from the empirical quantile at $\alpha\%$, which we shall denote by z_α . This means that, if $F(y) = \int_{-\infty}^y f(u)du$ is the cumulative density function of returns, then $F(z_\alpha) = \alpha$ and the probability area to the right of z_α is equal to $1 - \alpha$. One of the main assumptions made with many models for calculating VaR is that the returns Y_1, Y_2, \dots, Y_n are independent and identically distributed (IID). This is extremely important in supporting the idea that VaR (for future returns) can be forecasted based on past data. If the IID assumption is not true, then the empirical quantile cannot be simply calculated from a formula.

Let η be the number of times the realized losses exceed the VaR threshold. The risk manager expects *ex ante* that $E(\eta) = n\alpha$. However, *ex post* it is likely that $\eta \neq n\alpha$. For backtesting, the daily loss series implies a sequence of success or failure, depending whether the loss is greater than VaR threshold or not. The probability of failure is α and therefore, with n datapoints, the probability density function of η is given by the binomial distribution with parameters n and α

$$p(\eta = x) = \binom{n}{x} \alpha^x (1 - \alpha)^{n-x} \quad (1)$$

for $x \in \{0, 1, 2, \dots\}$. If the sample size n is large enough, the central limit theorem implies that $\frac{\eta - n\alpha}{\sqrt{n\alpha(1-\alpha)}}$ follows a standard Gaussian distribution. An asymptotic confidence interval for the number of losses that will be seen η can then be easily calculated. For example, a 95% asymptotic confidence interval for η is

$$\begin{aligned} -1.96\sqrt{n\alpha(1-\alpha)} + n\alpha &< \eta \\ &< 1.96\sqrt{n\alpha(1-\alpha)} + n\alpha \end{aligned} \quad (2)$$

From the probabilistic point of view the P/L values constitute a random sample $\{Y_1, Y_2, \dots, Y_n\}$ with cumulative distribution function

$$F(y_1, y_2, \dots, y_n; \theta) = \prod_{i=1}^n F_k(y_k; \phi_k) = \prod_{i=1}^n F(y; \phi)$$

where the last equality follows from the IID assumptions. For the empirical calculations of VaR the reordered sample $(Y_{[1]}, Y_{[2]}, \dots, Y_{[n]})$, with $Y_{[1]} \leq Y_{[2]} \leq \dots \leq Y_{[n]}$ is of interest because the VaR at level α is equal to the negative of the ν -th lowest value, where $\nu = 100\alpha + 1$. The statistic $Y_{[1]}$ is called the first order statistic, $Y_{[2]}$ is called the second order statistic, and so on. $Y_{[n]}$ is called the n -th order statistic, and they are all sample quantiles. The theory of order statistics allows making calculations on sample quantiles. This translates for empirical work based on the sample above into calculating the negative of the ν -th lowest value, where $\nu = n\alpha + 1$, or $Y_{[\nu]}$.

The portfolio losses can be analyzed through the empirical cumulative distribution function

$$\tilde{F}(y) = \begin{cases} 0 & \text{if } y < Y_{[1]} \\ \frac{i}{n} & \text{if } Y_{[i]} \leq y < Y_{[i+1]} \\ 1 & \text{if } y \geq Y_{[\nu]} \end{cases} \quad (3)$$

The inverse of this empirical cdf can be used as an estimator of VaR at α level. The VaR estimator is the order statistic $Y_{[j]}$ such that $\frac{i-1}{v} < \alpha \leq \frac{i}{v}$, which is slightly different from the upper empirical cumulative distribution function value calculated as the $Y_{[j]}$ such that $\frac{i-1}{v} \leq \alpha < \frac{i}{v}$. Mausser (2001) pointed out that with 100 IID P/L values, the VaR at 5% level would be estimated by the former estimator as $Y_{[5]}$ and by the latter as $Y_{[6]}$.

One major criticism in using VaR to quantify potential losses is the inability to gauge the size of extreme losses. To overcome this problem another risk measure called expected tail loss (ETL) has been introduced. The ETL is defined as the mean losses that exceed the VaR threshold. Hence, within the same framework proposed to calculate VaR, one can determine ETL by simply estimating the mean of the sam-

ple censored by the VaR estimate. If $Y_{[j]}$ is the order statistic estimator representing VaR, ETL can be estimated as the average of $(Y_{[1]}, Y_{[2]}, \dots, Y_{[j-1]})$. It is important to realize that while ETL may be more informative for gauging the potential losses than VaR, from an estimation point of view ETL will always depend on VaR.

The calculation of VaR and expected tail loss (ETL) with the order statistics methodology can be easily implemented in Matlab. Table 1 contains the VaR and ETL as estimated via the order statistics method for simulated samples using the Gaussian distribution and the t distribution for the series of P/L, at various confidence levels and sample sizes. In addition, the confidence intervals determined as the 0.025% and 0.975% percentiles of the distribution of each risk measure are also included. For a given sample size, the confidence intervals for both VaR and ETL are widening with the increase in the level of confidence, as shown in Figures 1 and 2. Similar results are obtained for larger sample sizes and other distributions. For a prespecified level of confidence, the confidence intervals tend to go narrower with the increase in the sample size.

JOINT PROBABILITY DISTRIBUTIONS FOR ORDER STATISTICS

If $F_{[i]}(u) = P(Y_{[i]} \leq u)$ is the cumulative distribution function of the i -th order statistic, then it is not difficult to see that $F_{[1]}(y) = 1 - [1 - F(y; \phi)]^n$ and $F_{[n]}(y) = F(y; \phi)^n$. Exploiting the fact that we use the quantile as a VaR estimator, Dowd (2001) suggested applying the following known result from order statistics for backtesting purposes

$$P(\text{exactly } j \text{ values from } Y_1, Y_2, \dots, Y_n \text{ are } \leq y) = \binom{n}{j} F(y; \phi)^j [1 - F(y; \phi)]^{n-j} \quad (4)$$

to derive the cumulative distribution function of this estimator

$$F_{[j]}(y) = P(Y_{[j]} \leq y) = \sum_{i=j}^n \binom{n}{i} F(y; \phi)^i [1 - F(y; \phi)]^{n-i} \quad (5)$$

Table 1 Order Statistics for VaR and ETL for One-Day Holding Period at 90%, 95% and 99% Confidence Levels and Various Sample Sizes Using Standard Normal Distribution and *t* Distribution

Sample size	Level	Measure	Normal			<i>t</i>		
			2.50%	Median	97.5%	2.50%	Median	97.5%
<i>n</i> = 100	90%	VaR	0.9299	1.2816	1.5874	0.9247	1.2770	1.5854
		ETL	1.4677	1.7535	2.0120	1.4671	1.7538	2.0198
	95%	VaR	1.2116	1.6449	2.0078	1.2068	1.6435	2.0130
		ETL	1.6956	2.0614	2.3788	1.6975	2.0670	2.3974
	99%	VaR	1.6031	2.3263	2.8160	1.6012	2.3407	2.8520
		ETL	2.0254	2.6640	3.1116	2.0335	2.6897	3.1677
<i>n</i> = 500	90%	VaR	1.1278	1.2816	1.4263	1.1268	1.2807	1.4256
		ETL	1.6269	1.7535	1.8748	1.6271	1.7515	1.8758
	95%	VaR	1.4543	1.6449	1.8218	1.4537	1.6446	1.8220
		ETL	1.8985	2.0614	2.2150	1.8996	2.0598	2.2176
	99%	VaR	1.9921	2.3263	2.6185	1.9930	2.3292	2.6236
		ETL	2.3650	2.6640	2.9299	2.3685	2.6653	2.9385
<i>n</i> = 1000	90%	VaR	1.1735	1.2816	1.3850	1.1731	1.2811	1.3847
		ETL	1.6644	1.7535	1.8401	1.6645	1.7513	1.8405
	95%	VaR	1.5110	1.6449	1.7719	1.5108	1.6447	1.7720
		ETL	1.9467	2.0614	2.1715	1.9473	2.0590	2.1727
	99%	VaR	2.0899	2.3263	2.5425	2.0906	2.3278	2.5447
		ETL	2.4519	2.6640	2.8604	2.4539	2.6623	2.8643
<i>n</i> = 5000	90%	VaR	1.2337	1.2816	1.3285	1.236	1.2815	1.3284
		ETL	1.7139	1.7535	1.7926	1.7140	1.7510	1.7927
	95%	VaR	1.5857	1.6449	1.7027	1.5856	1.6448	1.7027
		ETL	2.0105	2.0614	2.1114	2.0106	2.0583	2.1116
	99%	VaR	2.2214	2.3263	2.4274	2.2216	2.3266	2.4278
		ETL	2.5695	2.6640	2.7556	2.5700	2.6600	2.7562
<i>n</i> = 10000	90%	VaR	1.2478	1.2816	1.3148	1.2478	1.2815	1.3148
		ETL	1.7256	1.7535	1.7813	1.7256	1.7510	1.7813
	95%	VaR	1.6031	1.6449	1.6859	1.6031	1.6448	1.6859
		ETL	2.0256	2.0614	2.0968	2.0255	2.0582	2.0969
	99%	VaR	2.2524	2.3263	2.3984	2.2525	2.3265	2.3985
		ETL	2.5974	2.6640	2.7292	2.5976	2.6597	2.7296

Note: The number of degrees of freedom for *t* is chosen as the sample size minus 2.

In the following we shall denote $F(y; \phi)$ by $F(y)$, for simplicity. David (1981) pointed to the following useful result giving an analytical formula for the distribution function of the order statistic of order *j*.

$$F_{[j]}(y) = \mathcal{B}_{F(y)}(j, n - j + 1) \tag{6}$$

where $\mathcal{B}_U(a, b) = \frac{\int_0^U t^{a-1}(1-t)^{b-1}dt}{B(a, b)}$ is the incomplete beta function and $B(a, b)$ is the beta function. This helps to calculate the pdf function for those distributions that are absolute continuous with respect to a dominant probability measure.¹ The probability density function of

the *j*-th order statistics is

$$q_{[j]}(y) = \frac{1}{B(j, n-j+1)} F^{j-1}(y)[1-F(y)]^{n-j} f(y) \tag{7}$$

where $f(y) = \frac{dF}{dy}(y)$.

DISTRIBUTION-FREE CONFIDENCE INTERVALS FOR VaR

From a practical point of view, without any loss of generality, it is safe to assume that the cumulative distribution function *F* is strictly

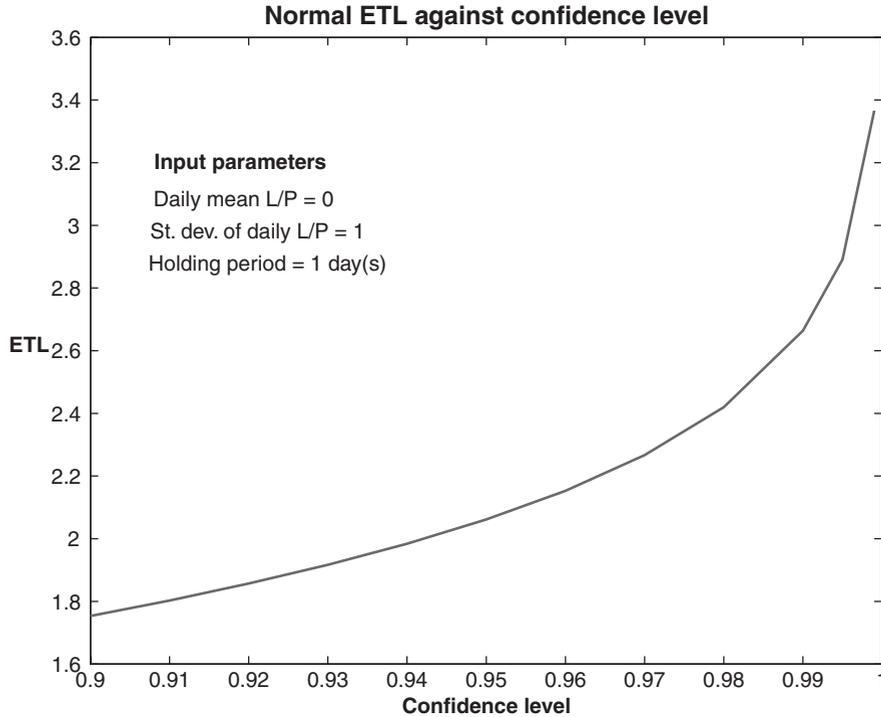


Figure 1 Expected Tail Loss for Normal P/L versus Level of Confidence When the Sample Size Is 100; Calculations Are Done with Order Statistics

increasing. Then, for any $\alpha \in (0, 1)$ the equation

$$F(y) = \alpha \tag{8}$$

has a unique solution. This solution refers to the entire population and it is called the quantile of order α , denoted by z_α . The 95% VaR is $z_{0.05}$.

The order statistics can provide a distribution-free confidence interval for the population quantiles. Thompson (1936) showed that

$$P(Y_{[i]} \leq z_\alpha \leq Y_{[j]}) = \sum_{k=i}^{j-1} \binom{n}{k} \alpha^k (1 - \alpha)^{n-k} \tag{9}$$

This powerful result allows the construction of distribution-free confidence intervals for VaR. For given sample size n and VaR level α , there are many combinations of i and j that make the quantity in (9) larger or equal to $1 - a$, the confidence level desired. There may be several combinations of order statistics $Y_{[i]}, Y_{[j]}$ that satisfy the relationship (9) and the risk manager

may decide to select the combination leading to the shortest confidence interval. Remark that choosing the degree of confidence $1 - a$ is independent of the level of confidence α for VaR point-estimation. In other words, a 95% confidence interval for the population quantile z_α can be calculated for 95% VaR or for 99% VaR.

BIVARIATE ORDER STATISTICS

The risk manager is faced with a dilemma. On one hand the regulators are asking usually for 99%-VaR calculation so that the banks are requested to set aside sufficient capital in order to absorb 99% of all losses. On the other hand, internal models may be used for day-to-day operations to forecast 95% Var. As explained by Brooks and Persand (2002) using an example from Kupiec (1995), the standard error of the

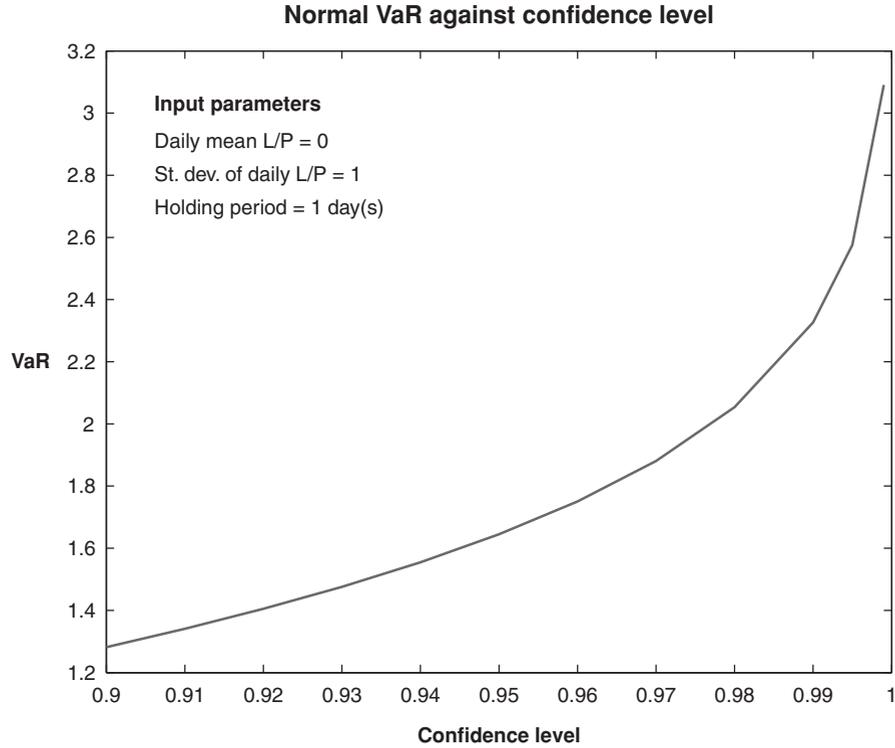


Figure 2 VaR for Normal P/L versus Level of Confidence When the Sample Size Is 100; Calculations Are Done with Order Statistics

99% VaR can be more than 50% larger than the corresponding standard error for the 95% VaR. This is the case for a model using the Gaussian distribution and it can be even worse for fat tail distributions, with the confidence intervals for the first percentile four times wider than confidence intervals for the fifth percentile. For back-testing purposes it would be ideal to do a joint analysis. Thus, the *bivariate* joint distribution of two order statistics will provide the confidence regions (two-dimensional sets) for pairs of VaR estimates. For example, the confidence regions for 1% VaR and 5% VaR are recovered from the bivariate joint distribution of $Y_{[v_1]}, Y_{[v_2]}$ where $v_1 = n \times 1/100 + 1$ and $v_2 = n \times 5/100 + 1$, respectively. This distribution is fully characterized by

$$F_{[i,j]}(x, y) = P(Y_{[i]} \leq x, Y_{[j]} \leq y) \quad (10)$$

with $1 \leq i < j \leq n$. The probability on the right side of equation (10) can be interpreted as the

probability that at least i values from the entire sample Y_1, Y_2, \dots, Y_n are not greater than x and at least j values from the same sample Y_1, Y_2, \dots, Y_n are not greater than y . Hence

$$F_{[i,j]}(x, y) = \sum_{k=j}^n \sum_{s=i}^k P(\text{exactly } i \text{ of } Y_1, Y_2, \dots, Y_n \text{ are } \leq x \text{ and exactly } j \text{ of } Y_1, Y_2, \dots, Y_n \text{ are } \leq y) \quad (11)$$

As in the univariate case, see David (1981), it follows that

$$F_{[i,j]}(x, y) = \sum_{k=j}^n \sum_{s=i}^k \frac{n!}{s!(k-s)!(n-k)!} \times [F(x)]^s [F(y) - F(x)]^{k-s} [1 - F(y)]^{n-k} \quad (12)$$

for any $x < y$. Since for $x \geq y$ the event $\{Y_{[j]} \leq y\}$ implies $Y_{[i]} \leq x$ then $F_{[i,j]}(x, y) = F_{[j]}(y)$.

An interesting corollary following from this result is that any two order statistics, and therefore VaR estimates at different levels, are not independent. This follows because the joint distribution in (12) cannot be factorized as a product of two factors, one depending only on x and the other only on y , up to a proportionality constant. In other words, if both 1% VaR and 5% VaR, for example, are needed for risk management purposes, then the quality of the VaR estimates should be investigated looking at the joint bivariate distribution like that in (12) rather than separate distributions of the type given in (5).

KEY POINTS

- Order statistics can be used as estimators of VaR and ETL and they are easy to compute.
- Banks may have to work with VaR measures at several levels of confidence because of regulatory requirements that may not coincide exactly with internal risk management decisions.
- ETL can be estimated easily with the framework based on order statistics, as the mean of the sample censored by the VaR threshold.
- For a given sample size, the confidence intervals for both VaR and ETL are widening with the increase in the level of confidence. For a prespecified level of confidence, the confidence intervals tend to go narrower with the increase in the sample size.
- There is a closed form solution for the density of any order statistic, which has been advocated here as a VaR estimator. Therefore, it would be easy to perform backtesting of VaR in this setup.
- The bivariate distribution of any two order statistics is known in closed form and therefore could be used for backtesting when banks have to work with two VaR measures simultaneously.

NOTE

1. For practical cases such as those encountered in finance we can safely assume that the random variables describing P/L series are continuous and they have probability density functions.

REFERENCES

- Berkowitz, J., and O'Brien, J. (2002). How accurate are value-at-risk models at commercial banks. *Journal of Finance* 57, 3: 1093–1111.
- Brooks, C., and Persaud, G. (2002). Model choice and value-at-risk performance. *Financial Analysts Journal* 58, 5: 87–97.
- David, H. (1981). *Order Statistics*, 2nd ed. New York: Wiley.
- Dowd, K. (2000). Assessing VaR accuracy. *Derivatives Quarterly* 6, 3: 61–63.
- Dowd, K. (2001). Estimating VaR with order statistics. *Journal of Derivatives* 8, 3: 23–30.
- Duffie, D., and Pan, J. (1997). An overview of value-at-risk. *Journal of Derivatives* 4, 3: 7–49.
- Jackson, P., Maude, D. J., and Perraudin, W. (1997). Bank capital and value-at-risk. *Journal of Derivatives* 4: 73–90.
- Jorion, P. (1996). Risk2: Measuring the risk in value-at-risk. *Financial Analysts Journal* 52: 47–56.
- Jorion, P. (1997). *Value-at-Risk: The New Benchmark for Controlling Market Risk*. Burr Ridge, IL: Irwin.
- Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives* 3: 73–84.
- Marshall, C., and Siegel, M. (1997). Value-at-risk: Implementing a risk measurement standard. *Journal of Derivatives* 4: 91–110.
- Mausser, H. (2001). Calculating quantile-based risk analytics with l-estimators. *ALGO Research Quarterly* 4, 4: 33–47.
- Pritsker, M. (1997). Evaluating VaR methodologies: Accuracy versus computational time. *Journal of Financial Services Research* 12: 201–241.
- Thompson, W. (1936). On confidence ranges for the median and other expectation distributions for populations of unknown distribution form. *Annals of Mathematical Statistics* 42: 268–269.

Risk Measures

Measuring Interest Rate Risk: Effective Duration and Convexity

GERALD W. BUETOW Jr., PhD, CFA
President and Founder, BFRC Services, LLC

ROBERT R. JOHNSON, PhD, CFA, CAIA
Independent Financial Consultant, Charlottesville, VA

Abstract: Modified duration and effective duration are two ways to measure the price sensitivity of a fixed income security. Both measure the percentage price change of a security from an absolute change in yields. Effective duration is a more complete measure of price sensitivity since it incorporates embedded optionality while modified duration does not. Combining effective duration with effective convexity is a superior risk management and measurement approach than using modified duration and convexity. In general, for fixed income securities with embedded options, numerical approaches (effective) to risk measurement are superior to analytic (modified) approaches.

Modified duration ignores any effect on cash flows that might take place as a result of changes in interest rates. *Effective duration* does not ignore the potential for such changes in cash flows. For example, bonds with embedded options will have very different cash flow properties as interest rates (or yields) change. Modified duration ignores these effects completely. In order to apply effective duration, an available interest rate model and corresponding pricing model are needed.¹ The example in this entry shows how to compute the effective duration of securities with cash flows that are dependent on changes in either the level or dynamics of the term structure of interest rates.

There is no difference between modified and effective duration for *option-free* or *straight bonds*. In fact, it can be shown that they are

mathematically identical when the change in rates (or yields) becomes very small. As shown in the example, even for bonds with embedded options, the differences between the two measures are minimal over certain ranges of yields. For example, when the embedded option is far out-of-the-money, the cash flows of the bond are not affected by small changes in yields, resulting in almost no difference in cash flows between the two measures.

Convexity and effective convexity measure the curvature of the price/yield relationship. Convexity (sometimes referred to as *standard convexity*) suffers the same limitations as modified duration and is therefore not generally useful for securities with embedded options. However, similar to the duration measures, in ranges of rates (or yields) where

the cash flows are not materially affected by small changes in yields, the two convexity measures are almost identical.

As with the duration measures, there is no difference between convexity and effective convexity for option-free or straight bonds. In fact, it can be shown that they are mathematically identical when the change in rates (or yields) becomes very small. As shown above, even for bonds with embedded options, the differences between the two measures are minimal over certain ranges of rates depending on the characteristics of the embedded option. For example, when the embedded option is far out-of-the-money, the cash flows of the bond are not affected by small changes in yields.

EFFECTIVE DURATION AND EFFECTIVE CONVEXITY—AN EXAMPLE

The following example illustrates how to calculate and interpret effective duration and effective convexity for straight bonds and bonds with embedded options.²

Suppose we need to measure the interest rate sensitivity of the following three securities:

1. A 5-year, 6.70% coupon straight (noncallable and nonputable) semiannual coupon bond, with a current price of 102.75% of par.
2. A 5-year, 6.25% coupon bond, callable at par in years 2 through 5 on the semiannual coupon dates, with a current price of 99.80% of par.
3. A 5-year, 5.75% coupon bond, putable at par in years 2 through 5 on the semiannual coupon dates, with a current price of 100.11% of par.

The cash flows of these securities are very different as interest rates change. Consequently, the sensitivities to changes in interest rates are also very different.

Using the *Black-Derman-Toy* interest rate model³ that is based on the existing term structure, the term structure of interest rates is shifted up and down by 10 basis points (bps) and the resulting price changes are recorded. P_- corresponds to the price after a downward shift in interest rates, P_+ corresponds to the price after an upward shift in interest rates, P is the current price, and S is the assumed shift in the term structure. (Note that shifting the term structure in a parallel manner will result in a change in yields equal to the shift for option-free bonds.) Table 1 shows these prices for each bond. The

Table 1 Original Prices and Resulting Prices from a Downward and Upward 10 Basis Point Interest Rate Shift and the Corresponding Effective Duration and Effective Convexity for Three Bonds Based on the Black-Derman-Toy Model

Variable	Price Changes Following 10 bp Shift		
	Original Price P	Upward Shift of 10 bp P_+	Downward Shift of 10 bp P_-
Straight Bond Price	102.7509029	102.3191235	103.1848805
Callable Bond Price	99.80297176	99.49321718	100.1085624
Putable Bond Price	100.1089131	99.84237604	100.3819059

<i>Effective Duration and Effective Convexity Measures Calculated from Using the Price Changes Resulting from the 10bp Shifts in the Term Structure</i>		
	Effective duration	Effective convexity
Straight Bond	4.21	21.39
Callable Bond	3.08	-41.72
Putable Bond	2.70	64.49

Table 2 Effective Duration and Effective Convexity for Various Shifts in the Term Structure for Three Bonds

Term Structure Shift (bps)	Straight Bond		Callable Bond		Putable Bond	
	Effective Duration	Effective Convexity	Effective Duration	Effective Convexity	Effective Duration	Effective Convexity
-500	4.40	23.00	1.91	4.67	4.46	23.46
-250	4.30	22.19	1.88	4.55	4.37	22.66
0	4.21	21.39	3.08	-41.72	2.70	64.49
250	4.12	20.62	4.15	20.85	1.87	7.07
500	4.03	19.87	4.07	20.10	1.81	4.23
1000	3.85	18.42	3.89	18.66	1.77	4.03

formulas for calculating effective duration and effective convexity are as follows:

$$\text{Effective duration} = \frac{(P_-) - (P_+)}{2PS} \quad (1)$$

$$\text{Effective convexity} = \frac{(P_-) - (P_+) - 2P}{PS^2} \quad (2)$$

It is critical to understand the importance of the pricing model in this exercise. The model must account for the change in cash flows of the securities as interest rates change. The callable and putable bonds have very different cash flow characteristics that depend on the level of interest rates. The pricing model used must account for this property.⁴

Straight Bond

The effective duration for the *straight bond* is found by recording the price changes from shifting the term structure up (P_+) and down (P_-) by 10 bps and then substituting these values into equation (1). The prices are shown in Table 1. Consequently, the computation is:

$$\begin{aligned} \text{Effective duration} &= \frac{103.1848805 - 102.3191235}{2(102.7509029)(0.001)} \\ &= 4.21 \end{aligned}$$

Similarly, the calculation for effective convexity is found by substituting the corresponding prices into equation (2):

$$\begin{aligned} \text{Effective convexity} &= \frac{103.1848805 + 102.3191235 - 2(102.7509029)}{102.7509029(0.001)^2} \\ &= 21.39 \end{aligned}$$

For the straight bond, the modified duration is 4.21 and the convexity is 21.40. These are very close to the effective measures shown in Table 1. This demonstrates that, for option-free bonds, the two measures are almost the same for small changes in yields.

Table 2 shows the effects of the term structure shifts on the effective duration and effective convexity of the straight bond. The effective duration increases as yields decrease because as yields decrease the slope of the price yield relationship for option-free bonds becomes steeper and effective duration (and modified duration) is directly proportional to the slope of this relationship. For example, the effective duration at very low yields (-500-bp shift) is 4.40 and decreases to 3.85 at very high rates (+1,000 bps), Figure 1 illustrates this phenomenon; as yields increase notice how the slope of the price/yield relationship decreases (becomes more horizontal or flatter).

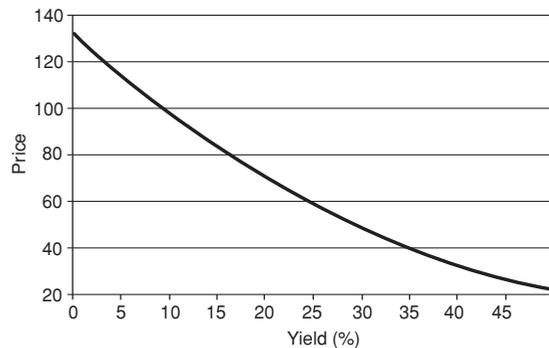


Figure 1 Price/Yield Relationship of the Straight Bond

As the term structure shifts up (that is, as rates rise), the yield to maturity on a straight bond increases by approximately the same amount. As the yield increases, its convexity decreases. Figure 1 illustrates this property. As yields increase, the curvature (or the rate of change of the slope) decreases. The results in Table 2 for the straight bond also bear this out. The effective convexity values become smaller as yields increase. For example, the effective convexity at very low yields (−500-bp shift) is 23.00 and decreases to 18.43 at very high rates (+1,000-bp shift).

These are both well-documented properties of option-free bonds. The modified duration and convexity numbers for the straight bond are almost identical to the effective measures for the straight bond shown in Table 2.

Callable Bond

The effective duration for the *callable bond* is found by recording the price changes from shifting the term structure up (P_+) and down (P_-) by 10 bps and then substituting these values into equation (1). The prices are shown in Table 1. Note that these prices take into account the changing cash flows resulting from the embedded call option. Consequently, the computation is:

$$\begin{aligned} \text{Effective duration} &= \frac{100.1085624 - 99.49321718}{2(99.800297)(0.001)} \\ &= 3.08 \end{aligned}$$

Similarly, the calculation for effective convexity is found by substituting the corresponding prices into equation (2):

$$\begin{aligned} \text{Effective convexity} &= \frac{100.1085624 + 99.49321718 - 2(99.80297176)}{99.80297176(0.001)^2} \\ &= -41.72 \end{aligned}$$

The relationship between the shift in rates and effective duration is shown in Table 2 and in Figure 2. As rates increase, the effective duration of

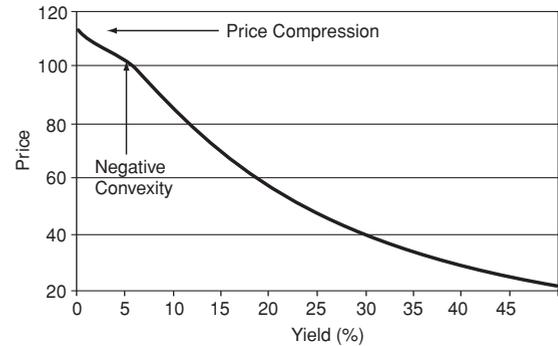


Figure 2 Price/Yield Relationship of the Callable Bond

the callable bond becomes larger. For example, the effective duration at very low yields (−500-bp shift) is 1.91 and increases to 3.89 at very high rates (+1,000 bps). This reflects the fact that as rates increase the likelihood of the bond being called decreases and, as a result, the bond behaves more like a straight bond; hence, its effective duration increases. Conversely, as rates drop, this likelihood increases and the bond and its effective duration behave more like a bond with a two-year maturity because of the call option becoming effective in two years. As rates decrease significantly, the likelihood of the issuer calling the bond in two years increases. Consequently, at very low and intermediate rates the difference between the effective duration measure and modified duration is large and at very high rates the difference is small.

As explained above, effective convexity measures the curvature of the price/yield relationship of bonds. Low values for effective convexity simply mean that the relationship is becoming linear (an effective convexity of zero represents a linear relationship). As shown in Table 2, the effective convexity values of the callable bond at extremely low interest rates (that is, for the −250-bp and −500-bp shifts in the term structure) are very small positive numbers (4.55 and 4.67, respectively). This means that the relationship is almost linear but exhibits slight convexity. This is due to the call option being delayed by two years. At these

extremely low interest rates, the callable bond exhibits slight positive convexity because the price compression at the call price is not complete for another two years. (*Price compression* for a callable bond refers to the property that a callable bond's price appreciation potential is severely limited as yields decline. As shown in Figure 2 as yields fall below a certain level (that is, where the yield corresponds to the call price), the price appreciation of the callable bond is being compressed). If this bond were immediately callable, the price/yield relationship would exhibit positive convexity at high yields and negative convexity at low yields. At the current level of interest rates, the effective convexity is negative as expected. At these rate levels, the embedded call option causes enough price compression to cause the curvature of the price/yield relationship to be *negatively convex* (that is, concave). Figure 2 illustrates these properties. It is at these levels that the embedded option has a significant effect on the cash flows of the callable bond.

Table 2 shows that for large positive yield curve shifts (that is, for the +250-bp, +500-bp, and +1,000-bp shifts in the term structure), the effective convexity of the callable bond becomes positive and very close to the effective convexity values of the straight bond. For example, the effective convexity at the +250-bp shift is 20.85 for the callable bond and 20.62 for the *straight bond*. The only reason they are not the same is because the coupon rates of the bonds are not equal. Consequently, at very low and intermediate rates the difference between effective convexity and the standard convexity is large and at very high rates the difference is small. The intuition behind these findings is straightforward. At low rates, the cash flows of the callable bond are severely affected by the likelihood of the embedded call option being exercised by the issuer. At high rates, the embedded call option is so far out-of-the-money that it has almost no effect on the cash flows of the callable bond and so the callable bond behaves like a straight bond.

Putable Bond

The effective duration for the *puttable bond* is found by recording the price changes from shifting the term structure up (P_+) and down (P_-) by 10 bps and then substituting these values into equation (1). The prices are shown in Table 1. Note that these prices take into account the changing cash flows resulting from the embedded put option. Consequently, the computation is:

$$\begin{aligned} \text{Effective duration} &= \frac{100.3819059 - 99.84237604}{2(100.1089131)(0.001)} \\ &= 2.70 \end{aligned}$$

Similarly, the calculation for effective convexity is found by substituting the corresponding prices into equation (2):

$$\begin{aligned} \text{Effective convexity} &= \frac{100.3819059 + 99.84237604 - 2(100.1089131)}{100.1089131(0.001)^2} \\ &= 64.49 \end{aligned}$$

Because the puttable bond behaves so differently from the other two bonds, the effective duration and effective convexity values are very different. As rates increase, the bond behaves more like a two-year bond because the owner will, in all likelihood, exercise the right to put the bond back at the put price as soon as possible. As a result, effective duration of the puttable bond is expected to decrease as rates increase. This is due to the embedded put option severely affecting the cash flows of the puttable bond. Conversely, as rates fall, the puttable bond behaves more like a five-year straight bond since the embedded put option is so far out-of-the-money and has little effect on the cashflows of the puttable bond. Effective duration should reflect these properties. Table 2 shows that this is indeed the case. For example, the effective duration at very low yields (−500-bp shift) is 4.46 and decreases to 1.77 at very high rates (+1,000 bps). Consequently, at very high rates and intermediate rates the difference between the effective duration and modified duration

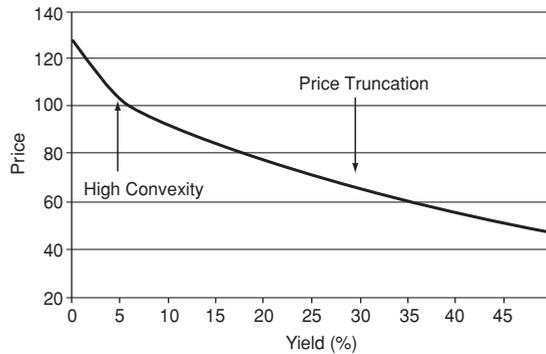


Figure 3 Price/Yield Relationship of the Puttable Bond

measures is large and at low rates the difference is small.

Table 2 shows that the effective convexity of the puttable bond is positive for all rate shifts as would be expected, but it becomes smaller as rates increase (that is, for the +250-bp, +500-bp, and +1,000-bp shifts in the term structure). As rates increase, the puttable bond price/yield relationship will become linear because of the bond's price truncation at the put price. (Price truncation for a puttable bond refers to the property that the puttable bond's price depreciation potential is severely limited as yields increase.) As shown in Figure 3 as yields rise above a certain level (that is, where the yield corresponds to the put price), the *price* depreciation of the puttable bond is *truncated*.) This is the reason for the small effective convexity values for the puttable bond for the three positive shifts in the term structure (7.07, 4.23, and 4.03, respectively). It is at these levels that the embedded put option has a significant effect on the cash flows of the puttable bond. Consequently, at very high rates and intermediate rates the difference between the effective convexity and standard convexity is very large. Figure 3 illustrates these properties.

At very low rates (that is, for the 250-bp and 500-bp downward shifts in the term structure), the puttable bond behaves like a 5-year straight bond because the put option is so far out-of-the-money. Therefore, as the term structure is

shifted downward, the puttable bond's effective convexity values approach those of a comparable 5-year straight bond. Comparing the effective convexity measures for the puttable bond and the straight bond illustrates this characteristic. For example, the effective convexity at the -250-bp shift is 22.66 for the puttable bond and 22.19 for the straight bond. The two convexity measures are almost identical. In fact, they would be identical if their coupon rates were equal.

Figure 2 illustrates these properties. Also notice how the transition from low yields to high yields forces the price/yield relationship to have a very high convexity at intermediate levels of yields. For example, the current effective convexity of the puttable bond is 64.49 compared to 21.39 for the straight bond and -41.72 for the callable bond. This is because of the price truncation of the puttable bond resulting from the embedded put option moving from out-of-the-money and having little influence over the cash flows to in-the-money and having a significant impact on cash flows.

PUTTING IT ALL TOGETHER

Notice in Table 2 how effective duration changes much more across yields for the callable and puttable bonds than it does for the straight bond. This is to be expected because the embedded options have such a significant influence over cash flows as yields change over a wide spectrum. Interestingly, at high (low) yields the callable (puttable) bond's effective duration is very close to the straight bond. This is where the embedded call (put) option is so far out-of-the-money that the two securities behave similarly. The same intuition holds for the effective convexity measures.

A common use of effective duration and effective convexity is to estimate the percentage price changes in fixed income securities for assumed changes in yield. In fact, it is not uncommon for effective duration and effective

Table 3 Percentage Price Changes Assuming an Increase in Yield of 100 bps and Effective Duration and Effective Convexity for Various Shifts in the Term Structure

Term Structure Shift (bp)	Straight Bond			Callable Bond			Putable Bond		
	% Price Change Using Effective Duration	% Price Change Using Effective Convexity	Total % Price Change	% Price Change Using Effective Duration	% Price Change Using Effective Convexity	Total % Price Change	% Price Change Using Effective Duration	% Price Change Using Effective Convexity	Total % Price Change
	-500	-4.40	0.11500	-4.28500	-1.91	0.02335	-1.88665	-4.46	0.11730
-250	-4.30	0.11095	-4.18905	-1.88	0.02275	-1.85725	-4.37	0.11330	-4.25670
0	-4.21	0.10695	-4.10305	-3.08	-0.20860	-3.28860	-2.70	0.32245	-2.37755
250	-4.12	0.10310	-4.01690	-4.15	0.10425	-4.04575	-1.87	0.03535	-1.83465
500	-4.03	0.09935	-3.93065	-4.07	0.10050	-3.96950	-1.81	0.02115	-1.78885
1000	-3.85	0.09210	-3.75790	-3.89	0.09330	-3.79670	-1.77	0.02015	-1.74985

convexity to be presented in terms of estimated percentage price change for a given change in yield (typically 100 bp): Tables 3 and 4 show this alternative presentation for a ±100 bp change in yield. These results are computed by substituting the values from Table 2 into the following relationship:

$$\begin{aligned} \text{\% Price change} &= \frac{\Delta P}{P} \approx -(ED)(\Delta y)(100) \\ &\quad + \frac{1}{2}(EC)(\Delta y)^2(100) \end{aligned} \quad (3)$$

where *ED* is the effective duration, *EC* is the effective convexity, and Δy is the assumed change in yield (e.g., 100 bp). Equation (3) is the result of a Taylor Series expansion on the bond price function. Also, note that the effective duration (ED) and effective convexity (EC) terms can be

replaced by modified duration and standard convexity, respectively, for option-free bonds.

Table 3 illustrates the resulting percentage price changes resulting from an increase in yield of 100 bps at various levels of the term structure. For example, the percentage price change for the callable bond at the current term structure (0-bp shift) is calculated using the values from Table 2 and substituting them into equation (3) as follows:

$$\begin{aligned} \text{\% Price change} &\approx -(3.08)(0.01)(100) \\ &\quad + \frac{1}{2}(-41.72)(0.01)^2(100) \\ &\approx -3.08 - 0.2086 = -3.2886\% \end{aligned}$$

This example shows that the estimated total percentage price change from effective convexity (-0.2086%) is much smaller than the

Table 4 Percentage Price Changes Assuming a Decrease in Yield of 100 bps and Effective Duration and Effective Convexity for Various Shifts in the Term Structure

Term Structure Shift (bp)	Straight Bond			Callable Bond			Putable Bond		
	% Price Change Using Effective Duration	% Price Change Using Effective Convexity	Total % Price Change	% Price Change Using Effective Duration	% Price Change Using Effective Convexity	Total % Price Change	% Price Change Using Effective Duration	% Price Change Using Effective Convexity	Total % Price Change
	-500	4.40	0.1150	4.5150	1.91	0.0234	1.9334	4.46	0.1173
-250	4.30	0.1110	4.4110	1.88	0.0228	1.9028	4.37	0.1133	4.4833
0	4.21	0.1070	4.3170	3.08	-0.2086	2.8714	2.70	0.3225	3.0225
250	4.12	0.1031	4.2231	4.15	0.1043	4.2543	1.87	0.0354	1.9054
500	4.03	0.0994	4.1294	4.07	0.1005	4.1705	1.81	0.0212	1.8312
1000	3.85	0.0921	3.9421	3.89	0.0933	3.9833	1.77	0.0202	1.7902

percentage price change from effective duration (−3.08).

Table 4 illustrates the resulting percentage price changes resulting from a decrease in yield of 100 bp at the various levels of the term structure. For example, the percentage price change for the callable bond at the current term structure (0-bp shift) is calculated using the values from Table 2 and substituting them into equation (3) as follows:

$$\begin{aligned} \% \text{ Price change} &\approx -(3.08)(-0.01)(100) \\ &\quad + \frac{1}{2}(-41.72)(-0.01)^2(100) \\ &\approx 3.08 - 0.2086 = 2.8714\% \end{aligned}$$

KEY POINTS

- Duration and convexity are measures for estimating the price sensitivity of a security to changes in interest rates.
- Modified duration and effective duration are two ways to measure the price sensitivity of a fixed income security. Both measure the percentage price change of a security from an absolute change in yields.
- There are important differences between effective duration and modified duration and effective convexity and convexity. The differences are due to changing cash flows of the security being evaluated.
- The effective measures account for changing cash flows and the traditional measures do not. The differences between the two are very significant whenever the cash flows are greatly affected by the level of interest rates. However, to properly compute the effective measures both an interest rate and a valuation model are required. Consequently, they are more computationally intensive than the traditional measures.
- The effective and traditional measures are identical for option-free bonds.
- Combining effective duration with effective convexity is a superior risk management and measurement approach than using modified duration and convexity.
- Investors would be best served by always using the effective measures since they properly account for the cash flow characteristics of a security.

NOTES

1. For the impact of interest rate models on duration and convexity, see Buetow, Hanke, and Fabozzi (2001).
2. For an illustration of how duration and convexity are computed for mortgage-backed securities, see Golub (2006) and Fabozzi (1999).
3. Black, Derman, and Toy, 1990.
4. Note that when calculating the measures, users are cautioned to not round values. Since the denominators of both the duration and convexity terms are very small, any rounding will have a significant impact on results.

REFERENCES

- Black, F., Derman, E., and Toy, W. (1990). A one-factor model of interest rates and its application to Treasury bond options. *Financial Analysts Journal* (January–February): 24–32.
- Buetow, G. W. Jr., Hanke, B., and Fabozzi, F. J. (2001). The impact of different interest rate models on effective duration, effective convexity, and option-adjusted spreads. *Journal of Fixed Income* (December): 41–53.
- Fabozzi, F. J. (1999). *Duration, Convexity, and Other Bond Risk Measures*. Hoboken, NJ: John Wiley & Sons.
- Golub, B. W. (2006). Approaches for measuring duration of mortgage-related securities. In F. J. Fabozzi (ed.), *The Handbook of Mortgage-Backed Securities*, 6th ed. (pp. 823–856). New York: McGraw-Hill.

Yield Curve Risk Measures

FRANK J. FABOZZI, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

STEVEN V. MANN, PhD
Professor of Finance, Moore School of Business, University of South Carolina

Abstract: Duration is a useful metric for assessing a bond portfolio's sensitivity to a parallel shift in the reference yield curve (e.g., the Treasury yield curve). When the yield curve shift is not parallel, however, two bond portfolios with the same duration will not generally experience the same return performance. To evaluate differences in expected performance across portfolios, it is therefore necessary to quantify the price impact due to changes in the shape, as opposed to a parallel shift, of the yield curve. The risk exposure of a portfolio to changes in the yield curve is called yield curve risk. Several approaches have been suggested for measuring yield curve risk.

Duration and convexity are useful measures for approximating how the value of a bond portfolio or a bond index will change for a parallel shift in interest rates. Yet, empirically, both published studies¹ and proprietary studies by asset management firms have found that yield curve changes are not parallel. The exposure of a bond portfolio or a bond index to changes in the shape of the yield curve is called *yield curve risk*.

There are several approaches for measuring yield curve risk. In this entry, we describe some of the more common approaches: cash-flow distribution analysis versus a benchmark, key rate duration, slope elasticity measure, yield curve reshaping duration, and analysis of likely shifts in the yield curve. We begin the entry with an illustration of the drawback of using duration

and convexity measures when the yield curve does not shift in a parallel fashion.

DURATION, CONVEXITY, AND NONPARALLEL YIELD CURVE SHIFTS

To illustrate the limitations of duration and convexity, let's first look at how two portfolios consisting of hypothetical Treasury securities with the same portfolio duration will perform if the yield curve does not shift in a parallel fashion. Consider the three hypothetical Treasury securities shown in Table 1. Security A is the short-term Treasury, security B is the long-term Treasury, and security C is the intermediate-term Treasury. Each Treasury security is selling

Table 1 Three Hypothetical Treasury Securities to Illustrate the Limitations of Duration and Convexity

Information on three Treasury securities:				
Treasury Issue	Coupon Rate (%)	Price	Yield to Maturity (%)	Maturity (years)
A	6.5	100	6.5	5
B	8.0	100	8.0	20
C	7.5	100	7.5	10

Calculation of duration and convexity (shock rates by 10 basis points):				
Treasury issue	Value if rate changes by			
	+10 bp	-10 bp	Duration	Convexity
A	99.5799	100.4222	4.21122	10.67912
B	99.0177	100.9970	9.89681	73.63737
C	99.3083	100.6979	6.49821	31.09724

at par, and it is assumed that the next coupon payment is six months from now. The duration and convexity for each security are calculated in the exhibit. Since all the securities are trading at par value, the durations and convexities are the dollar duration and dollar convexity per \$100 of par value.

Suppose that the following two Treasury portfolios are constructed. The first portfolio consists of only security C, the 10-year issue, and shall be referred to as the "bullet portfolio." The second portfolio consists of 51.86% of security A and 48.14% of security B, and this portfolio shall be referred to as the "barbell portfolio."

The dollar duration of the bullet portfolio is 6.49821. Recall that dollar duration is a measure of the dollar price sensitivity of a security or a portfolio. The dollar duration of the barbell is the weighted average of the dollar duration of the two Treasury securities in the portfolio and is computed below:

$$0.5186(4.21122) + 0.4814(9.89681) = 6.94821$$

The dollar duration of the barbell is equal to the dollar duration of the bullet. In fact, the barbell portfolio was designed to produce this result.

Duration is just a first approximation of the change in price resulting from a change in in-

terest rates. The convexity measure provides a second approximation. The dollar convexity measure of the two portfolios is not equal. The dollar convexity measure of the bullet portfolio is 31.09724. The dollar convexity measure of the barbell is a weighted average of the dollar convexity measure of the two Treasury securities in the portfolio. That is,

$$0.5186(10.67912) + 0.4814(73.63737) = 40.98658$$

Thus, the bullet has a dollar convexity measure that is less than that of the barbell portfolio. Below is a summary of the dollar duration and dollar convexity of the two portfolios:

Parameter	Bullet Portfolio	Barbell Portfolio
Dollar duration	6.49821	6.49821
Dollar convexity	31.09724	40.98658

The better Treasury portfolio depends on the portfolio manager's investment objectives and investment horizon. Let's assume a six-month investment horizon. The last column of Table 2

Table 2 Performance of Bullet and Barbell Treasury Portfolios over a Six-Month Horizon Assuming a Parallel Yield Curve Shift: Scenario Analysis

Yield Change (in bps)	Total Return (%)		
	Bullet Portfolio	Barbell Portfolio	Difference ^a
-300	53.47	55.79	-2.32
-250	44.95	46.38	-1.43
-200	36.79	37.55	-0.76
-150	28.99	29.26	-0.27
-100	21.51	21.47	0.05
-50	14.35	14.13	0.22
-25	10.89	10.63	0.26
0	7.50	7.22	0.28
25	4.18	3.92	0.27
50	0.93	0.70	0.23
100	-5.36	-5.45	0.09
150	-11.39	-11.28	-0.11
200	-17.17	-16.79	-0.38
250	-22.71	-22.01	-0.70
300	-28.03	-26.96	-1.06

^aA positive sign indicates that the bullet portfolio outperformed the barbell portfolio; a negative sign indicates that the barbell portfolio outperformed the bullet portfolio.

shows the difference in the total return over a six-month investment horizon for the two Treasury portfolios, assuming that the yield curve shifts in a “parallel” fashion. By parallel it is meant that the yield for the short-term security (A), the intermediate-term security (C), and the long-term security (B) changes by the same number of basis points, shown in the first column of the table. The total return reported in the second column of Table 2 is:

$$\begin{aligned} & \text{Bullet portfolio's total return} \\ & - \text{Barbell portfolio's total return} \end{aligned}$$

Thus, a positive value in the last column means that the bullet portfolio outperformed the barbell portfolio, while a negative sign means that the barbell portfolio outperformed the bullet portfolio. Note that no assumption is needed for the reinvestment rate since the

three securities comprising the portfolios are assumed to be trading right after a coupon payment has been made and therefore there is no accrued interest.

Which portfolio is the better investment alternative if the yield curve shifts in a parallel fashion and the investment horizon is six months? The answer depends on the amount by which yields change. Notice in the last column that if yields change by less than 100 basis points, the bullet portfolio will outperform the barbell portfolio. The reverse is true if yields change by more than 100 basis points.

Now let's look at what happens if the yield curve does not shift in a parallel fashion. The last column of Tables 3 and 4 show the relative performance of the two Treasury portfolios for a nonparallel shift of the yield curve. Specifically, in Table 3 it is assumed that if the yield on

Table 3 Performance of Bullet and Barbell Treasury Portfolios over a Six-Month Horizon Assuming a Flattening of the Yield Curve: Scenario Analysis

Yield change for C (in bps)	Total return (%)		
	Bullet Portfolio	Barbell Portfolio	Difference ^a
-300	53.47	58.98	-5.51
-250	44.95	49.26	-4.31
-200	36.79	40.15	-3.36
-150	28.99	31.60	-2.62
-100	21.51	23.58	-2.06
-50	14.35	16.03	-1.67
-25	10.89	12.42	-1.53
0	7.50	8.92	-1.42
25	4.18	5.53	-1.35
50	0.93	2.23	-1.30
100	-5.36	-4.09	-1.27
150	-11.39	-10.06	-1.33
200	-17.17	-15.70	-1.47
250	-22.71	-21.04	-1.67
300	-28.03	-26.11	-1.92

Assumptions:
 Change in yield of security C results in a change in the yield of security A plus 30 basis points.
 Change in yield of security C results in a change in the yield of security B minus 30 basis points.
^aA positive sign indicates that the bullet portfolio outperformed the barbell portfolio; a negative sign indicates that the barbell portfolio outperformed the bullet portfolio.

Table 4 Performance of Bullet and Barbell Treasury Portfolios over a Six-Month Horizon Assuming a Steepening of the Yield Curve: Scenario Analysis

Yield Change for C (in bps)	Total Return (%)		
	Bullet Portfolio	Barbell Portfolio	Difference ^a
-300	53.47	52.82	0.65
-250	44.95	43.70	1.24
-200	36.79	35.14	1.65
-150	28.99	27.09	1.89
-100	21.51	19.52	1.99
-50	14.35	12.39	1.97
-25	10.89	8.98	1.91
0	7.50	5.66	1.84
25	4.18	2.44	1.74
50	0.93	-0.69	1.63
100	-5.36	-6.70	1.34
150	-11.39	-12.38	0.99
200	-17.17	-17.77	0.60
250	-22.71	-22.88	0.17
300	-28.03	-27.73	-0.30

Assumptions:
 Change in yield of security C results in a change in the yield of security A minus 30 basis points.
 Change in yield of security C results in a change in the yield of security B plus 30 basis points.
^aA positive sign indicates that the bullet portfolio outperformed the barbell portfolio; a negative sign indicates that the barbell portfolio outperformed the bullet portfolio.

C (the intermediate-term security) changes by the amount shown in the first column, A (the short-term security) will change by the same amount plus 30 basis points, whereas B (the long-term security) will change by the same amount shown in the first column less 30 basis points. That is, the nonparallel shift assumed is a flattening of the yield curve. For this yield curve shift, the barbell will outperform the bullet for the yield changes assumed in the first column. While not shown in the table, for changes greater than 300 basis points for C, the opposite would be true.

In Table 4, the nonparallel shift assumes that for a change in C's yield, the yield on A will change by the same amount less 30 basis points, whereas the yield on B will change by the same amount plus 30 basis points. That is, it assumes that the yield curve will steepen. In this case, the bullet portfolio would outperform the barbell portfolio for all but a change in yield greater than 250 basis points for C.

The key point here is that looking at duration or convexity tells us little about performance over some investment horizon because performance depends on the magnitude of the change in yields and how the yield curve shifts.

CASH-FLOW DISTRIBUTION ANALYSIS VERSUS A BENCHMARK

The most straightforward approach to assessing a portfolio's risk exposure to yield curve shifts is by looking at the distribution of the present value of the cash flows for the portfolio being managed versus a benchmark. The benchmark will be either a bond index or a liability structure. The steps are as follows:

Step 1: Determine the discrete time periods for the analysis. The shortest and longest time is determined by the shortest and longest cash flows for the portfolio and the benchmark. Each time period is referred to as a cash-flow vertex.

Step 2: Compute the cash flows for the portfolio and the benchmark for each cash-flow vertex.

Step 3: Compute the present value of the cash flows for the portfolio and the benchmark for each cash-flow vertex. The spot rate used to compute the present value is the spot rate for the cash-flow vertex. For example, if the cash-flow vertex is year 5, the 5-year spot rate is used.

Step 4: Compute the duration contribution at each cash flow vertex for the portfolio and the benchmark.

Step 5: Compute the duration contribution as a percentage of duration for both the portfolio and the benchmark for each cash-flow vertex.

Step 6: Compute the difference in the portfolio percentage and benchmark percentage computed in Step 5 for each cash-flow vertex.

In practice, the application is not straightforward because of the inclusion of bonds with embedded options and mortgage-backed and asset-backed securities. Suppose a bond is a 7-year bond that is callable in three years. The cash flows for this bond depend on the portfolio manager's assessment of the probability that it will be called in three years. For mortgage-backed and asset-backed securities, the cash flows depend on the prepayment assumption.

Another difficulty in the implementation process is the allocation of cash flows to the cash-flow vertices when a cash flow is not exactly on a cash-flow vertex date. For example, consider a bond whose coupon payment of \$1 million is to be received 4.75 years from now and that there is a 4-year and 5-year cash-flow vertex. How should the \$1 million coupon payment be allocated? The procedure would be to allocate 25% to the 4-year cash-flow vertex and 75% to the 5-year cash-flow vertex.

Despite its simplicity, the cash-flow distribution analysis is commonly used as a measure of yield curve risk for index fund managers (see Volpert, 2000).

KEY RATE DURATION

One approach to measure yield curve risk is to change the yield for a particular maturity of the yield curve and determine the sensitivity of a security or portfolio to this change, holding all other yields constant. The sensitivity of the change in value to a particular change in yield is called *rate duration*. There is a rate duration for every point on the yield curve. Consequently, there is not one rate duration, but a vector of durations representing each maturity on the yield curve. The total change in value if all rates change by the same number of basis points is simply the duration of a security or portfolio to a parallel shift in rates.

This approach was first suggested by Chambers and Carleton (1988), who called it *duration vectors*. Reitano (1992) suggested a similar approach and referred to these durations as *partial durations*. The most popular version of this approach is that developed by Ho (1992). This approach examines how changes in Treasury yields at different points on the spot curve affect the value of a bond portfolio. Ho's methodology has three basic steps. The first step is to select several key maturities or "key rates" of the spot rate curve. Ho's approach focuses on 11 key maturities on the spot rate curve. These rate durations are called *key rate durations*. The specific maturities on the spot rate curve for which a key rate duration is measured are 3 months, 1 year, 2 years, 3 years, 5 years, 7 years, 10 years, 15 years, 20 years, 25 years, and 30 years. However, in order to illustrate Ho's methodology, we will select only three key rates: 1 year, 10 years, and 30 years.

The next step is to specify how other rates on the spot curve change in response to key rate changes. Ho's rule is that a key rate's effect on neighboring rates declines linearly and reaches zero at the adjacent key rates. For example, suppose the 10-year key rate increases by 40 basis points. All spot rates between 10 years and 30 years will increase but the amount each

changes will be different and the magnitude of the change diminishes linearly. Specifically, there are 40 semiannual periods between 10 and 30 years. Each spot rate starting with 10.5 years increases by 1 basis point less than the spot rate to its immediate left (that is, 39 basis points) and so forth. The 30-year rate which is the adjacent key rate is assumed to be unchanged. Thus, only one key rate changes at a time. Spot rates between 1 year and 10 years change in an analogous manner such that all rates change but by differing amounts. Changes in the 1-year key rate affect spot rates between 1 and 10 years, while spot rates 10 years and beyond are assumed to be unaffected by changes in the 1-year spot rate. In a similar vein, changes in the 30-year key rate affect all spot rates between 30 years and 10 years while spot rates shorter than 10 years are assumed to be unaffected by changes in the 30-year rate. This process is illustrated in Figure 1. Note that if we add the three rate changes together, we obtain a parallel yield curve shift of 40 basis points.

The third and final step is to calculate the percentage change in the bond's portfolio value when each key rate and neighboring spot rates are changed. There will be as many key rate durations as there are preselected key rates. Let's illustrate this process by calculating the key rate duration for a coupon bond. Our hypothetical 6% coupon bond has a maturity value of \$100 and matures in five years. The bond delivers coupon payments semiannually. Valuation is accomplished by discounting each cash flow using the appropriate spot rate. The bond's current value is \$107.32 and the process is illustrated in Table 5. The initial hypothetical (and short) spot curve is contained in column (3). (Note that the spot rates are annual rates and are reported as bond-equivalent yields. When present values are computed, we use the appropriate semiannual rates that are taken to be one half the annual rate.) The present values of each of the bond's cash flows are presented in the last column.

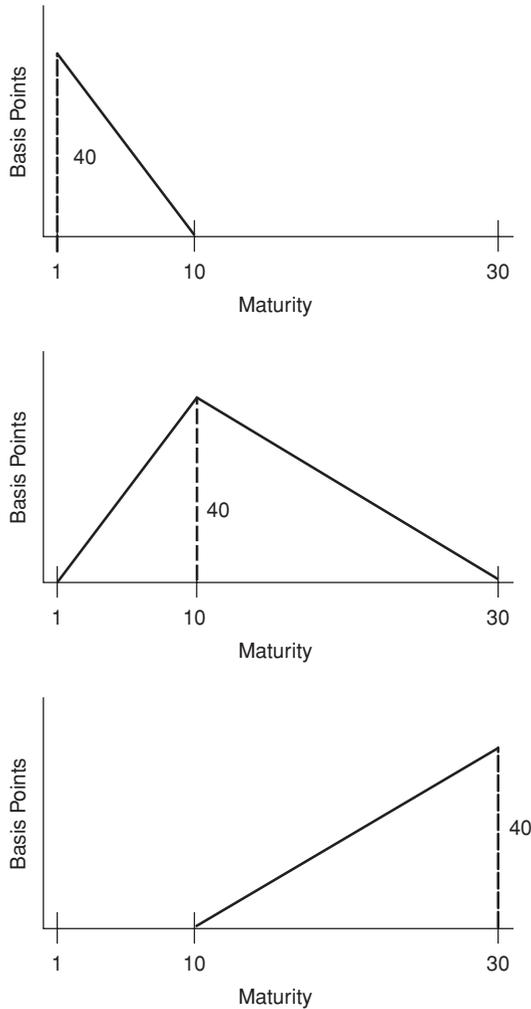


Figure 1 Graph of How Spot Rates Change when Key Rates Change

Table 5 Valuation of 5-Year 6% Coupon Bond Using Spot Rates

Years	Period	Spot Rate (in percent)	Cash Flow (in dollars)	Present Value (in dollars)
0.5	1	3.00	3.0	2.96
1.0	2	3.25	3.0	2.90
1.5	3	3.50	3.0	2.85
2.0	4	3.75	3.0	2.79
2.5	5	4.00	3.0	2.72
3.0	6	4.10	3.0	2.66
3.5	7	4.20	3.0	2.59
4.0	8	4.30	3.0	2.53
4.5	9	4.35	3.0	2.47
5.0	10	4.40	103.0	82.86
			Total	107.32

To compute the key rate duration of the 5-year bond, we must select some key rates. We assume the key rates are 0.5, 3, and 5 years. To compute the 0.5-year key rate duration, we shift the 0.5-year rate upwards by 20 basis points and adjust the neighboring spot rates between 0.5 and 3 years as described earlier. (The choice of 20 basis points is arbitrary.) Figure 2 shows the initial spot curve and the spot curve after the 0.5-year key rate and neighboring rates are shifted. The next step is to compute the bond's new value as a result of the shift. This calculation is shown in Table 6. The bond's value to the shift is \$107.30. To estimate the 0.5-year key rate duration, we divide the percentage change in the bond's price as a result of the shift in the spot curve by the change in the 0.5-year key rate. Accordingly, we employ the following formula:

$$\text{Key rate duration} = \frac{P_0 - P_1}{P_0(\Delta y)}$$

where

P_0 = the bond's value using the initial spot curve

P_1 = the bond's value after the shift in the spot curve

Δy = shift in the key rate (in decimal)

Substituting in numbers from our illustration presented above, we can compute the 0.5-year

Table 6 Valuation of the 5-Year 6% Coupon Bond after 0.5-Year Key Rate and Neighboring Spot Rates Change

Years	Period	Spot Rate (in percent)	Cash Flow (in dollars)	Present Value (in dollars)
0.5	1	3.20	3.0	2.95
1.0	2	3.41	3.0	2.90
1.5	3	3.62	3.0	2.84
2.0	4	3.83	3.0	2.78
2.5	5	4.04	3.0	2.71
3.0	6	4.10	3.0	2.66
3.5	7	4.20	3.0	2.59
4.0	8	4.30	3.0	2.53
4.5	9	4.35	3.0	2.47
5.0	10	4.40	103.0	82.86
			Total	107.30

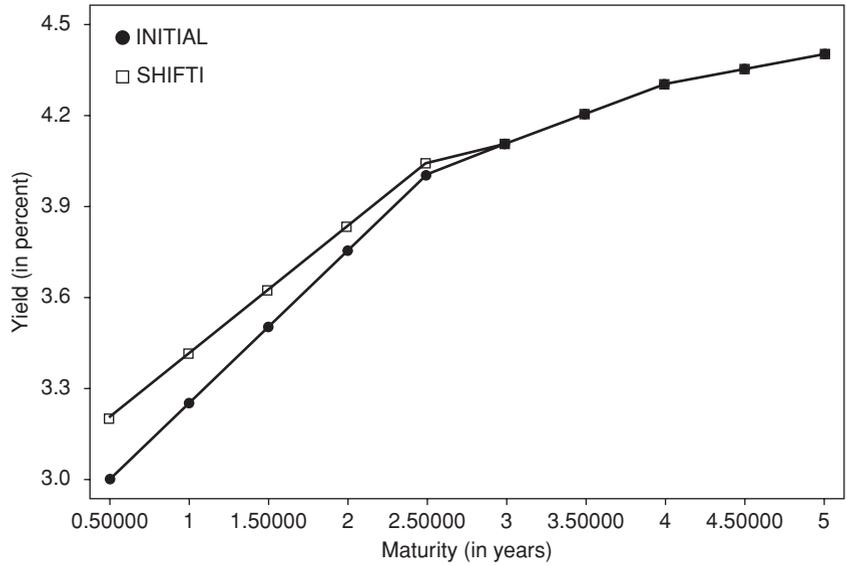


Figure 2 Graph of the Initial Spot Curve and the Spot Curve after the 0.5-Year Key Rate Shift

key rate duration as follows:

$$P_0 = 107.32$$

$$P_1 = 107.30$$

$$\Delta y = 0.002$$

$$\begin{aligned} \text{0.5-year key rate duration} &= \frac{107.32 - 107.30}{107.32(0.002)} \\ &= 0.0932 \end{aligned}$$

To compute the 3-year key rate duration, we repeat this process. We shift the 3-year rate by 20 basis points and adjust the neighboring spot rates as described earlier. Figure 3 shows the initial spot curve and the spot curve after the 3-year key rate and neighboring rates are shifted. Note that in this case the only two spot rates that do not change are the 0.5-year and the

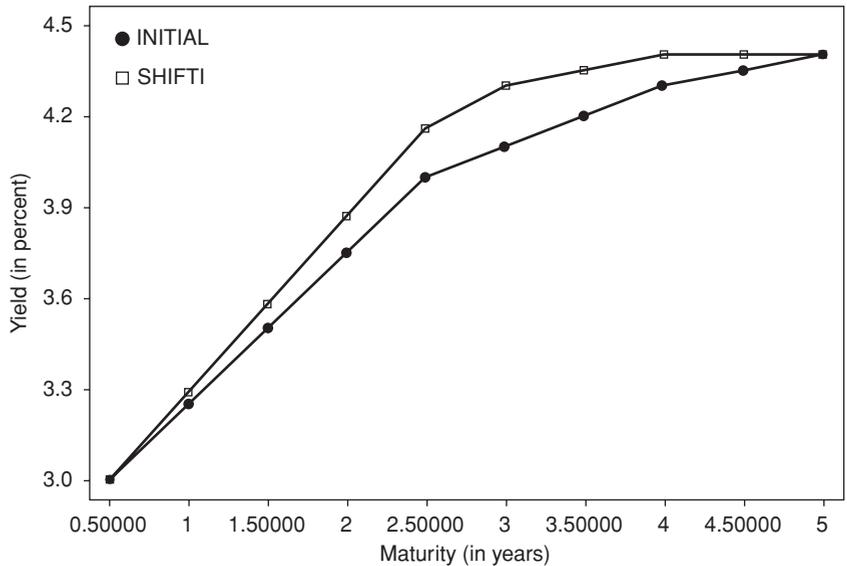


Figure 3 Graph of the Initial Spot Curve and the Spot Curve after the 3-Year Key Rate Shift

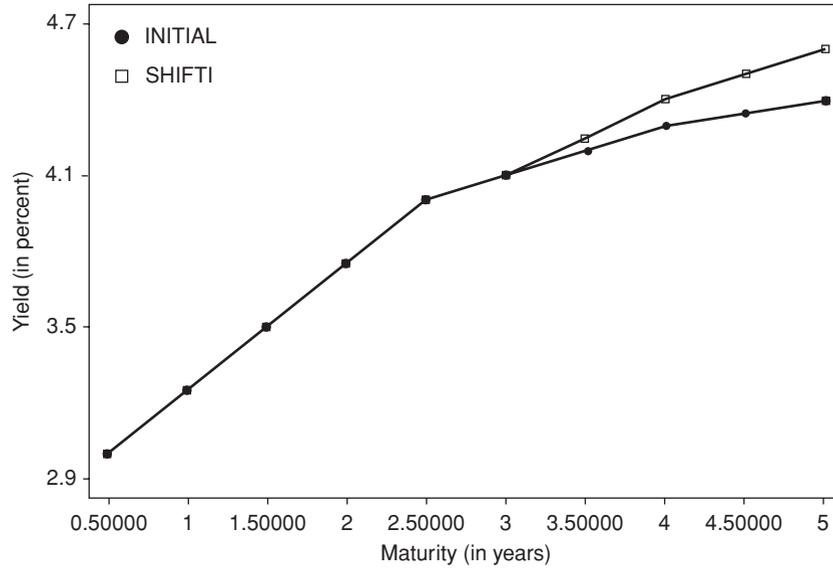


Figure 4 Graph of the Initial Spot Curve and the Spot Curve after the 5-Year Key Rate Shift

5-year key rates. Then, we compute the bond’s new value as a result of the shift. The bond’s postshift value is \$107.25 and the calculation appears in Table 7. Accordingly, the 3-year key rate duration is computed as follows:

$$\begin{aligned} \text{3-year key rate duration} &= \frac{107.32 - 107.25}{107.32(0.002)} \\ &= 0.3261 \end{aligned}$$

The final step is to compute the 5-year key duration. We shift the 5-year rate by 20 basis points

and adjust the neighboring spot rates. Figure 4 presents a graph of the initial spot curve and the spot curve after the 5-year key rate and neighboring rates are shifted. The bond’s postshift value is \$106.48 and the calculation appears in Table 8. Accordingly, the 5-year key rate duration is computed as follows:

$$\begin{aligned} \text{5-year key rate duration} &= \frac{107.32 - 106.48}{107.32(0.002)} \\ &= 3.9135 \end{aligned}$$

Table 7 Valuation of the 5-Year 6% Coupon Bond After 3-Year Key Rate and Neighboring Spot Rates Change

Years	Period	Spot Rate (in percent)	Cash Flow (in dollars)	Present Value (in dollars)
0.5	1	3.00	3.0	2.96
1.0	2	3.29	3.0	2.90
1.5	3	3.58	3.0	2.84
2.0	4	3.87	3.0	2.78
2.5	5	4.16	3.0	2.71
3.0	6	4.30	3.0	2.64
3.5	7	4.35	3.0	2.58
4.0	8	4.40	3.0	2.52
4.5	9	4.40	3.0	2.47
5.0	10	4.40	103.0	82.86
			Total	107.25

Table 8 Valuation of the 5-Year 6% Coupon Bond after 5-Year Key Rate and Neighboring Spot Rates Change

Years	Period	Spot Rate (in percent)	Cash Flow (in dollars)	Present Value (in dollars)
0.5	1	3.00	3.0	2.96
1.0	2	3.25	3.0	2.90
1.5	3	3.50	3.0	2.85
2.0	4	3.75	3.0	2.79
2.5	5	4.00	3.0	2.72
3.0	6	4.10	3.0	2.66
3.5	7	4.25	3.0	2.59
4.0	8	4.40	3.0	2.52
4.5	9	4.50	3.0	2.46
5.0	10	4.60	103.0	82.05
			Total	106.48

What information can be gleaned from these key rate durations? Each key rate duration by itself means relatively little. However, the distribution of the bond's key rate durations helps us assess its exposure to yield curve risk. Intuitively, the sum of the key rate durations is approximately equal to a bond's duration. (The reason it is only approximate is that modified duration assumes a flat yield curve, whereas key rate duration takes the spot curve as given.)

As a result, it is useful to think of a set of key rate durations as a decomposition of duration into sensitivities to various portions of the yield curve. In our illustration, it is not surprising that the lion's share of the yield curve risk exposure of the coupon bond in our illustration is due to the bond's terminal cash flow, so the 5-year key rate duration is the largest of the three. Simply put, the 5-year bond's value is more sensitive to movements in longer spot rates and less sensitive to movements in shorter spot rates.

Key rate durations are most useful when comparing two (or more) bond portfolios that have approximately the same duration. If the spot curve is flat and experiences a parallel shift, these two bond portfolios can be expected to experience approximately the same percentage change in value. However, the performance of the two portfolios will generally not be the same for a nonparallel shift in the spot curve. The key rate duration profile of each portfolio will give the portfolio manager some clues about the relative performance of the two portfolios when the yield curve changes shape and slope.

SLOPE ELASTICITY MEASURE

The *slope elasticity measure*, introduced by Schumacher, Dektar, and Fabozzi (1994) for managing the yield curve risk of portfolios of collateralized mortgage obligation bonds, also looks at the sensitivity of a position or portfolio to changes in the slope of the yield curve. They define the yield curve slope as the spread between the 30-year on-the-run Treasury yield

and the 3-month Treasury bill yield (that is, basically the longest and the shortest points on the Treasury yield curve).

They find that while this is not a perfect definition, it captures most of the effect of changes in yield curve slope. They then define changes in the yield curve as follows: Half of any basis point change in the yield curve slope results from a change in the 3-month yield and half from a change in the 30-year yield. For example, with a 200-basis-point steepening of the yield curve, the assumption is that 100 basis points of that steepening come from a rise in the 30-year yield, and another 100 basis points come from a fall in the 3-month yield.

The sensitivity of a bond's price to changes in the yield curve is simply its slope elasticity. They define slope elasticity as the approximate negative percentage change in a bond's price resulting from a 100-basis-point change in the slope of the curve. Slope elasticity is calculated as follows: Increase and decrease the yield curve slope, calculate the price change for these two scenarios after adjusting for the price effect of a change in the level of yields, and compare the prices to the initial price. More specifically, the slope elasticity for each scenario is calculated as follows:

$$\frac{\text{Price effect of a change in slope} / \text{Base price}}{\text{Change in yield curve slope}}$$

The slope elasticity is then the average of the slope elasticity for the two scenarios.

A bond or bond portfolio that benefits when the yield curve flattens is said to have positive slope elasticity; a bond or a bond portfolio that benefits when the yield curve steepens is said to have negative slope elasticity. The definition of yield curve risk follows from that of slope elasticity. It is defined as the exposure of the bond to changes in the slope of the yield curve.

YIELD CURVE RESHAPING DURATION

Yield curve reshaping duration, introduced by Klaffky, Ma, and Nozari (1992), focuses on three

points on the yield curve: 2-year, 10-year, and 30-year, and the spread between the 10-year and 2-year issues and the spread between the 30-year and 10-year issues. The former spread is referred to as the short end of the yield curve, and the latter spread the long end of the yield curve. Klaffky, Ma, and Nozari refer to the sensitivity of a portfolio to changes in the short end of the yield curve as *short-end duration (SEDUR)* and to changes in the long end of the yield curve as *long-end duration (LEDUR)*. These concepts, however, are applicable to other points on the yield curve.

To calculate the SEDUR of each security in the portfolio, the percentage change in the security's price is calculated for (1) a steepening of the yield curve at the short end by 50 basis points, and (2) a flattening of the yield curve at the short end of the yield curve by 50 basis points. Then the security's SEDUR is computed as follows:

$$\text{SEDUR} = \frac{P_s - P_f}{2P_0(\Delta y)}$$

where

- P_s = security's price if the short end of the yield curve steepens by 50 basis points
- P_f = security's price if the short end of the yield curve flattens by 50 basis points
- P_0 = security's current market price
- Δy = number of basis points by which the yield curve is changed

To calculate the LEDUR, the same procedure is used for each security in the portfolio: Calculate the price for (1) a flattening of the yield curve at the long end by 50 basis points, and (2) a steepening of the yield curve at the long end of the yield curve by 50 basis points. Then the security's LEDUR is computed as follows:

$$\text{LEDUR} = \frac{P_f - P_s}{2P_0(\Delta y)}$$

For an illustration, see Fabozzi (1999).

ANALYSIS OF LIKELY YIELD CURVE SHIFTS

While key rate duration is a useful measure for identifying the exposure of a portfolio to different potential shifts in the yield curve, it is difficult to employ this approach to yield curve risk in hedging a portfolio. An alternative approach is to investigate how yield curves have changed historically and incorporate typical yield curve change scenarios into the hedging process. This approach of using likely yield curve changes obtained from principal component analysis has been suggested by Richard and Gord (1997), Golub and Tilman (1997), and Axel and Vankudre (2000).

Empirically, studies have found that yield curve changes are not parallel. Rather, when the level of interest rates changes, studies have found that short-term rates move more than longer-term rates. Some firms develop their own proprietary models that decompose historical movements in the rate changes of Treasury strips with different maturities in order to analyze typical or likely rate movements. The statistical technique used to decompose rate movements is principal component analysis.

KEY POINTS

- When using a portfolio's duration and convexity to measure the exposure to interest rates, it is assumed that the yield curve shifts in a parallel fashion.
- For a nonparallel shift in the yield curve, duration and convexity may not provide adequate information about the risk exposure to changes in interest rates.
- Yield curve risk is the exposure of a portfolio to a change in the shape of the yield curve. There are several approaches that have been proposed for measuring a portfolio's yield curve risk.
- A simple approach to measuring yield curve risk, an approach commonly used by index managers, is an analysis of the cash

flow distribution of a portfolio relative to a benchmark.

- Key rate duration measures how changes in Treasury yields at different points on the spot rate curve affect the value of a bond.
- Slope elasticity looks at the sensitivity of a position or portfolio to changes in the slope of the yield curve and is defined as the approximate negative percentage change in a bond's price resulting from a 100-basis-point change in the slope of the curve.
- Yield curve reshaping duration decomposes the yield curve into a short end and a long end. The sensitivity of a portfolio to changes in the short end of the yield curve is called short-end duration (SEDUR) and to changes in the long end of the yield curve is called long-end duration (LEDUR).
- Using principal component analysis, a portfolio manager can determine likely yield curve shifts and use those shifts to assess the exposure of a portfolio to yield curve risk.

NOTE

1. See, e.g., Litterman and Scheinkman (1991) and Jones (1991).

REFERENCES

- Axel, R., and Vankudre, P. (2000). *Managing the yield curve with principal component analysis*. In F. J. Fabozzi (ed.), *Professional Perspectives on Fixed Income Portfolio Management*, Volume 3 (pp. 37–49). Hoboken, NJ: John Wiley & Sons.
- Dattatreya, R. E., and Fabozzi, F. J. (1995). The risk point method for measuring and controlling yield curve risk. *Financial Analysts Journal*, July–August: 45–54.
- Fabozzi, F. J. (1999). *Duration, Convexity, and Other Bond Risk Measures*. Hoboken, NJ: John Wiley & Sons.
- Golub, B. W., and Tilman, L. M. (1997). Measuring plausibility of hypothetical interest rate shocks. In F. J. Fabozzi (ed.), *Managing Fixed Income Portfolios* (pp. 73–86). Hoboken, NJ: John Wiley & Sons.
- Ho, T. S. Y. (1992). Key rate durations: Measures of interest rate risk. *Journal of Fixed Income*, September: 29–44.
- Jones, F. J. (1991). Yield curve strategies. *Journal of Fixed Income*, September: 43–51.
- Litterman, R., and Scheinkman, J. (1991). Common factors affecting bond returns. *Journal of Fixed Income*, June: 54–61.
- Richard, S. F., and Gord, B. J. (1997). Measuring and managing interest-rate risk. In F. J. Fabozzi (ed.), *Managing Fixed Income Portfolios* (pp. 19–30). Hoboken, NJ: John Wiley & Sons.
- Schumacher, M. P., Dektar, D. C., and Fabozzi, F. J. (1994). Yield curve risk of CMO bonds. In F. J. Fabozzi (ed.), *CMO Portfolio Management* (pp. 271–310). Hoboken, NJ: John Wiley & Sons.
- Volpert, K. E. (2000). Managing indexed and enhanced indexed bond portfolios. In F. J. Fabozzi (ed.), *Fixed Income Readings for the Chartered Financial Analyst Program* (pp. 85–100). New Hope, PA: Frank J. Fabozzi Associates.

Value-at-Risk

STOYAN V. STOYANOV, PhD

Professor of Finance at EDHEC Business School and Head of Research for EDHEC Risk Institute-Asia

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair-Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: A risk measure that has been widely accepted since the 1990s is the value-at-risk (VaR). In the late 1980s, it was integrated by JP Morgan on a firmwide level into its risk-management system. In the mid-1990s, the VaR measure was approved by regulators as a valid approach to calculating capital reserves needed to cover market risk. The Basel Committee on Banking Supervision released a package of amendments to the requirements for banking institutions, allowing them to use their own internal systems for risk estimation. In this way, capital reserves, which financial institutions are required to keep, could be based on the VaR numbers computed internally by an in-house risk management system. Generally, regulators demand that the capital reserve equal the VaR number multiplied by a factor between 3 and 4. Thus, regulators link the capital reserves for market risk directly to the risk measure. In practice, there are several approaches for estimating VaR.

In this entry, we cover the most commonly used risk measure used by financial institutions: *value-at-risk* (VaR). We comment on its properties and different calculation methods. Where possible, the definitions and equations are geometrically interpreted, making the ideas more intuitive and understandable.

VALUE-AT-RISK DEFINED

VaR is defined as the minimum level of loss at a given, sufficiently high, confidence level for

a predefined time horizon. The recommended confidence levels are 95% and 99%. Suppose that we hold a portfolio with a 1-day 99% VaR equal to \$1 million. This means that over the horizon of 1 day, the portfolio may lose more than \$1 million with probability equal to 1%.

The same example can be constructed for percentage returns. Suppose that the present value of a portfolio we hold is \$10 million. If the 1-day 99% VaR of the return distribution is 2%, then over the time horizon of 1 day, we lose more than 2% (\$200,000) of the portfolio present value with probability equal to 1%.

Denote by $(1 - \epsilon)100\%$ the confidence level parameter of the VaR. As we explained, losses larger than the VaR occur with probability ϵ . The probability ϵ , we call *tail probability*. Depending on the interpretation of the random variable, VaR can be defined in different ways. Formally, the VaR at confidence level $(1 - \epsilon)100\%$ (tail probability ϵ) is defined as the negative of the lower ϵ -quantile of the return distribution,

$$VaR_\epsilon(X) = - \inf_x \{x | P(X \leq x) \geq \epsilon\} = -F_X^{-1}(\epsilon) \tag{1}$$

where $\epsilon \in (0,1)$ and $F_X^{-1}(\epsilon)$ is the inverse of the distribution function. If the random variable X describes random returns, then the VaR number is given in terms of a return figure. The definition of VaR is illustrated in Figure 1.

If X describes random payoffs, then VaR is a threshold in dollar terms below which the portfolio value falls with probability ϵ ,

$$VaR_\epsilon(X) = \inf_x \{x | P(X \leq x) \geq \epsilon\} = F_X^{-1}(\epsilon) \tag{2}$$

where $\epsilon \in (0,1)$ and $F_X^{-1}(\epsilon)$ is the inverse of the distribution function of the random payoff. VaR

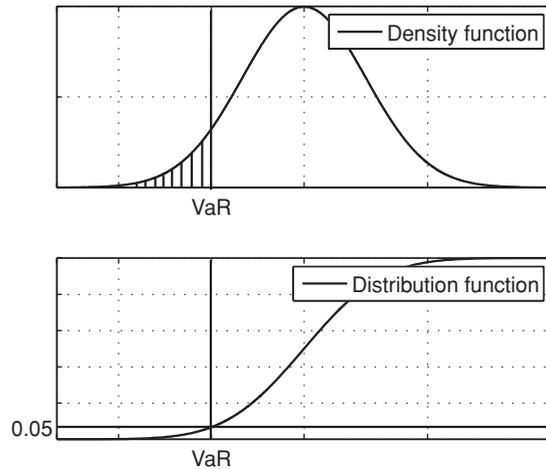


Figure 1 The VaR at 95% Confidence Level of a Random Variable X
Note: The top plot shows the density of X , the marked area equals the tail probability, and the bottom plot shows the distribution function.

can also be expressed as a distance to the present value when considering the profit distribution. The random profit is defined as $X - P_0$ where X is the payoff and P_0 is the present value. The VaR of the random profit equals

$$\begin{aligned} VaR_\epsilon(X - P_0) &= - \inf_x \{x | P(X - P_0 \leq x) \geq \epsilon\} \\ &= P_0 - VaR_\epsilon(X) \end{aligned}$$

in which $VaR_\epsilon(X)$ is defined according to (2) since X is interpreted as a random payoff. In this case, the definition of VaR is essentially given by equation (1).

According to the definition in equation (1), VaR may become a negative number. If $VaR_\epsilon(X)$ is a negative number, then this means that at tail probability ϵ we do not observe losses but profits. Losses happen with even smaller probability than ϵ . If for any tail probability $VaR_\epsilon(X)$ is a negative number, then no losses can occur and, therefore, the random variable X bears no risk as no exposure is associated with it. In this entry, we assume that random variables describe either random returns or random profits and we adopt the definition in equation (1).

We illustrate one aspect in which VaR differs from the deviation measures and all uncertainty measures. As a consequence of the definition, if we add to the random variable X a nonrandom profit C , the resulting VaR can be expressed by the VaR of the initial variable in the following way

$$VaR_\epsilon(X + C) = VaR_\epsilon(X) - C \tag{3}$$

Thus, adding a nonrandom profit decreases the risk of the portfolio. Furthermore, scaling the return distribution by a positive constant λ scales the VaR by the same constant,

$$VaR_\epsilon(\lambda X) = \lambda VaR_\epsilon(X) \tag{4}$$

It turns out that these properties characterize not only VaR. They are identified as key features of a risk measure.

To illustrate, let's use an example. Suppose initially we have a portfolio that consists of a common stock with random monthly return denoted by r_X . We rebalance the portfolio so

that it becomes an equally weighted portfolio of the stock and a default-free government bond with a nonrandom monthly return of 5.26%, $r_B = 5.26\%$. Thus, the portfolio return can be expressed as

$$r_p = r_X(1/2) + r_B(1/2) = r_X/2 + 0.0526/2$$

Using equations (3) and (4), we calculate that if $VaR_\epsilon(r_X) = 12\%$, then $VaR_\epsilon(r_p) \approx 3.365\%$, which is by far less than 6%—half of the initial risk. Any deviation measure would indicate that the dispersion (or the uncertainty) of the portfolio return r_p would be twice as small as the uncertainty of r_X .

A very important remark has to be made with respect to the performance of VaR and, as it turns out, of any other risk measure. It is heavily dependent on the assumed probability distribution of the variable X . An unrealistic hypothesis may result in underestimation or overestimation of true risk. If we use VaR to build reserves in order to cover losses in times of crises, then underestimation may be fatal and overestimation may lead to inefficient use of capital. An inaccurate model is even more dangerous in an optimal portfolio problem in which we minimize risk subject to some constraints, as it may adversely influence the optimal weights and therefore not reduce the true risk.

Even though VaR has been largely adopted by financial institutions and approved by regulators, it turns out that VaR has important deficiencies. While it provides an intuitive description of how much a portfolio may lose, generally, it should be abandoned as a risk measure. The most important drawback is that, in some cases, the reasonable diversification effect that every portfolio manager should expect to see in a risk measure is not present; that is, the VaR of a portfolio may be greater than the sum of the VaRs of the constituents

$$VaR_\epsilon(X + Y) > VaR_\epsilon(X) + VaR_\epsilon(Y) \quad (5)$$

in which X and Y stand for the random payoff of the instruments in the portfolio. This shows that VaR cannot be a true risk measure.

We give a simple example, which shows that VaR may satisfy (5). Suppose that X denotes a bond that either defaults with probability 4.5% and we lose \$50 or it does not default and in this case the loss is equal to zero. Let Y be the same bond but assume that the defaults of the two bonds are independent events. The VaR of the two bonds at 95% confidence level (5% tail probability) is equal to zero,

$$VaR_{0.05}(X) = VaR_{0.05}(Y) = 0$$

Being the 5% quantile of the payoff distribution in this case, VaR fails to recognize losses occurring with probability smaller than 5%. A portfolio of the two bonds has the following payoff profile: It loses \$100 with probability of about 0.2%, loses \$50 with probability of about 8.6%, and the loss is zero with probability 91.2%. Thus, the corresponding 95% VaR of the portfolio equals \$50 and clearly,

$$\begin{aligned} \$50 &= VaR_{0.05}(X + Y) > VaR_{0.05}(X) \\ &\quad + VaR_{0.05}(Y) = 0 \end{aligned}$$

What are the consequences of using a risk measure that may satisfy property (5)? It is going to mislead portfolio managers that there is no diversification effect in the portfolio and they may make the irrational decision to concentrate it only into a few positions. As a consequence, the portfolio risk actually increases.

Besides being sometimes incapable of recognizing the diversification effect, another drawback is that VaR is not very informative about losses beyond the VaR level. It only reports that losses larger than the VaR level occur with probability equal to ϵ but it does not provide any information about the likely magnitude of such losses, for example.

Nonetheless, VaR is not a useless concept to be abandoned altogether. For example, it can be used in risk reporting only as a characteristic of the portfolio return (payoff) distribution since it has a straightforward interpretation. The criticism of VaR is focused on its wide application by practitioners as a true risk measure, which, in view of the deficiencies described

above, is not well grounded and should be reconsidered.

COMPUTING PORTFOLIO VaR IN PRACTICE

In this section, we provide three approaches for portfolio VaR calculation that are used in practice. We assume that the portfolio contains common stocks, which is only to make the description easier to grasp; this is not a restriction of any of the approaches.

Suppose that a portfolio contains n common stocks and we are interested in calculating the daily VaR at 99% confidence level. Denote the random daily returns of the stocks by X_1, \dots, X_n and by w_1, \dots, w_n the weight of each stock in the portfolio. Thus, the portfolio return r_p can be calculated as

$$r_p = w_1 X_1 + w_2 X_2 + \dots + w_n X_n$$

The portfolio VaR is derived from the distribution of r_p . The three approaches vary in the assumptions they make.

The Approach of RiskMetrics

The approach of *RiskMetrics Group* is centered on the assumption that asset returns have a multivariate normal distribution. Under this assumption, the distribution of the portfolio return is also normal. Therefore, in order to calculate the portfolio VaR, we only have to calculate the expected return of r_p and the standard deviation of r_p . The 99% VaR will appear as the negative of the 1% quantile of the $N(Er_p, \sigma_{r_p}^2)$ distribution.

The portfolio expected return can be directly expressed through the expected returns of the assets

$$\begin{aligned} Er_p &= w_1 EX_1 + w_2 EX_2 + \dots + w_n EX_n \\ &= \sum_{k=1}^n w_k EX_k \end{aligned} \quad (6)$$

where E denotes mathematical expectation. Similarly, the variance of the portfolio return

$\sigma_{r_p}^2$ can be computed through the variances of the asset returns and their covariances,

$$\begin{aligned} \sigma_{r_p}^2 &= w_1^2 \sigma_{X_1}^2 + w_2^2 \sigma_{X_2}^2 + \dots + w_n^2 \sigma_{X_n}^2 \\ &\quad + \sum_{i \neq j} w_i w_j \text{cov}(X_i, X_j) \end{aligned}$$

in which the last term appears because we have to sum up the covariances between all pairs of asset returns. There is a more compact way of writing down the expression for $\sigma_{r_p}^2$ using matrix notation,

$$\sigma_{r_p}^2 = w' \Sigma w \quad (7)$$

in which $w = (w_1, \dots, w_n)$ is the vector of portfolio weights and Σ is the covariance matrix of asset returns,

$$\Sigma = \begin{pmatrix} \sigma_{X_1}^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{X_2}^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{X_n}^2 \end{pmatrix}$$

in which σ_{ij} , $i \neq j$, is the covariance between X_i and X_j , $\sigma_{ij} = \text{cov}(X_i, X_j)$. As a result, we obtain that the portfolio return has a normal distribution with mean given by equation (6) and variance given by equation (7).

The standard deviation is the scale parameter of the normal distribution and the mean is the location parameter. Due to the normal distribution properties, if $r_p \in N(Er_p, \sigma_{r_p}^2)$, then

$$\frac{r_p - Er_p}{\sigma_{r_p}} \in N(0, 1)$$

Thus, because of the properties (3) and (4) of the VaR, the 99% portfolio VaR can be represented as

$$\text{VaR}_{0.01}(r_p) = q_{0.99} \sigma_{r_p} - Er_p \quad (8)$$

where the standard deviation of the portfolio return σ_{r_p} is computed from equation (7), the expected portfolio return Er_p is given in (6), and $q_{0.99}$ is the 99% quantile of the standard normal distribution.

Note that $q_{0.99}$ is a quantity independent of the portfolio composition; it is merely a constant

that can be calculated in advance. The parameters that depend on the portfolio weights are the standard deviation of portfolio returns σ_{r_p} and the expected portfolio return. As a consequence, VaR under the assumption of normality is symmetric even though, by definition, VaR is centered on the left tail of the distribution; that is, VaR is asymmetric by construction. This result appears because the normal distribution is symmetric around the mean.

The approach of RiskMetrics can be extended for other types of distributions. Lamantia et al. (2006a) and Lamantia et al. (2006b) provide such extensions and comparisons for Student's t and stable distributions.

The Historical Method

The *historical method* does not impose any distributional assumptions; the distribution of portfolio returns is constructed from historical data. Hence, sometimes the historical simulation method is called a nonparametric method. For example, the 99% daily VaR of the portfolio return is computed as the negative of the empirical 1% quantile of the observed daily portfolio returns. The observations are collected from a predetermined time window such as the most recent business year.

While the historical method seems to be more general as it is free of any distributional hypotheses, it has a number of major drawbacks.

1. It assumes that the past trends will continue in the future. This is not a realistic assumption because we may experience extreme events in the future, for instance, which have not happened in the past.
2. It treats the observations as independent and identically distributed (IID), which is not realistic. The daily returns data exhibits clustering of the volatility phenomenon, autocorrelations and so on, which are sometimes a significant deviation from the IID assumption.
3. It is not reliable for estimation of VaR at very high confidence levels. A sample of one year of daily data contains 250 observations, which is a rather small sample for the purpose of the 99% VaR estimation.

The Hybrid Method

The *hybrid method* is a modification of the historical method in which the observations are not regarded as IID but certain weights are assigned to them depending on how close they are to the present. The weights are determined using the exponential smoothing algorithm. The exponential smoothing accentuates the most recent observations and seeks to take into account the time-varying volatility phenomenon.

The algorithm of the hybrid approach consists of the following steps.

1. Exponentially declining weights are attached to historical returns, starting from the current time and going back in time. Let $r_{t-k+1}, \dots, r_{t-1}, r_t$ be a sequence of k observed returns on a given asset, where t is the current time. The i -th observation is assigned a weight

$$\theta_i = c * \lambda^{t-i}$$

where $0 < \lambda < 1$, and $c = \frac{1-\lambda}{1-\lambda^k}$ is a constant chosen such that the sum of all weights is equal to one, $\sum \theta_i = 1$.

2. Similarly to the historical simulation method, the hypothetical future returns are obtained from the past returns and sorted in increasing order.
3. The VaR measure is computed from the empirical c.d.f. in which each observation has probability equal to the weight θ_i .

Generally, the hybrid approach is appropriate for VaR estimation of heavy-tailed time series. It overcomes, to some degree, the first and the second deficiency of the historical method but it is also not reliable for VaR estimation of very high confidence levels.

The Monte Carlo Method

In contrast to the historical method, the *Monte Carlo method* requires specification of a statistical model for asset returns. The statistical model is multivariate, hypothesizing both the behavior of the asset returns on a stand-alone basis and their dependence. For instance, the multivariate normal distribution assumes normal distributions for the asset returns viewed on a stand-alone basis and describes the dependencies by means of the covariance matrix. The multivariate model can also be constructed by specifying explicitly the one-dimensional distributions of the asset returns, and their dependence through a copula function.

The Monte Carlo method consists of the following basic steps.

- Step 1. *Selection of a statistical model.* The statistical model should be capable of explaining a number of observed phenomena in the data, for example, heavy tails, clustering of the volatility, and so on, which we think influence the portfolio risk.
- Step 2. *Estimation of the statistical model parameters.* A sample of observed asset returns is used from a predetermined time window, for instance the most recent 250 daily returns.
- Step 3. *Generation of scenarios from the fitted model.* Independent scenarios are drawn from the fitted model. Each scenario is a vector of asset returns, which depend on each other according to the presumed dependence structure of the statistical model.
- Step 4. *Calculation of portfolio risk.* Compute portfolio risk on the basis of the portfolio return scenarios obtained from the previous step.

The Monte Carlo method is a very general numerical approach to risk estimation. It does not require any closed-form expressions and, by choosing a flexible statistical model, accurate risk numbers can be obtained. A disadvantage is that the computed portfolio VaR is dependent on the generated sample of scenarios and will

fluctuate a little if we regenerate the sample. This side effect can be reduced by generating a larger sample. An illustration is provided in the following example.

Suppose that the daily portfolio return distribution is standard normal and, therefore, at Step 4 of the algorithm we have scenarios from the standard normal distribution. Under the assumption of normality, we can use the approach of RiskMetrics and compute the 99% daily VaR directly from formula (8). Nevertheless, we will use the Monte Carlo method to gain more insight into the deviations of the VaR based on scenarios from the VaR computed according to formula (8).

In order to investigate how the fluctuations of the 99% VaR change about the theoretical value, we generate samples of different sizes: 500, 1,000, 5,000, 10,000, 20,000, and 100,000 scenarios. The 99% VaR is computed from these samples and the numbers are stored. We repeat the experiment 100 times. In the end, we have 100 VaR numbers for each sample size. We expect that as the sample size increases, the VaR values will fluctuate less about the theoretical value which is $VaR_{0.01}(X) = 2.326$, $X \in N(0,1)$.

Table 1 contains the result of the experiment. From the 100 VaR numbers, we calculate the 95% confidence interval for the true value given in the third column. The confidence intervals cover the theoretical value 2.326 and also we notice that the length of the confidence interval

Table 1 The 99% VaR of the Standard Normal Distribution Computed from a Sample of Scenarios

Number of Scenarios	99% VaR	95% Confidence Interval
500	2.067	[1.7515, 2.3825]
1,000	2.406	[2.1455, 2.6665]
5,000	2.286	[2.1875, 2.3845]
10,000	2.297	[2.2261, 2.3682]
20,000	2.282	[2.2305, 2.3335]
50,000	2.342	[2.3085, 2.3755]
100,000	2.314	[2.2925, 2.3355]

Note: The 95% confidence interval is calculated from 100 repetitions of the experiment. The true value is $VaR_{0.01}(X) = 2.326$.

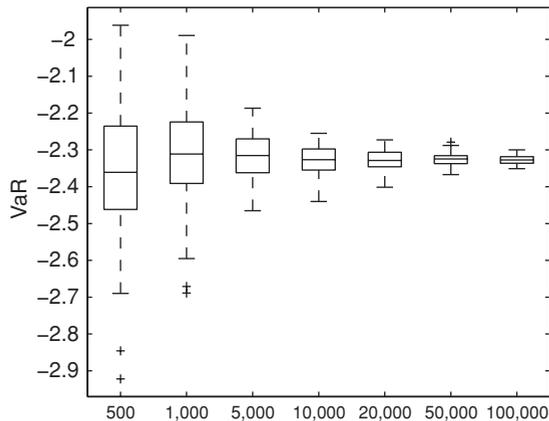


Figure 2 Boxplot Diagrams of the Fluctuation of the 99% VaR of the Standard Normal Distribution Based on Scenarios

Note: The horizontal axis shows the number of scenarios and the boxplots are computed from 100 independent samples.

decreases as the sample size increases. This effect is best illustrated with the help of the boxplot diagrams¹ shown in Figure 2. A sample of 100,000 scenarios results in VaR numbers that are tightly packed around the true value while a sample of only 500 scenarios may give a very inaccurate estimate.

This simple experiment shows that the number of scenarios in the Monte Carlo method has to be carefully chosen. The approach we used to determine the fluctuations of the VaR based on scenarios is a statistical method called parametric bootstrap. The bootstrap methods in general are powerful statistical methods that are used to compute confidence intervals when the problem is not analytically tractable but the calculations may be quite computationally intensive.

The true merits of the Monte Carlo method can only be realized when the portfolio contains complicated instruments such as derivatives. In this case, it is no longer possible to use a closed-form expression for the portfolio VaR (and any risk measure in general) because the distribution of portfolio return (or payoff) becomes quite arbitrary. The Monte Carlo method provides the general framework to generate scenar-

ios for the risk-driving factors, then reevaluates the financial instruments in the portfolio under each scenario, and, finally, estimates portfolio risk on the basis of the computed portfolio returns (or payoffs) in each state of the world.

While it may seem a straightforward approach, the practical implementation is a very challenging endeavor from both the software development and financial modeling points of view. The portfolios of big financial institutions often contain products that require yield curve modeling, development of fundamental and statistical factor models, and, on top of that, a probabilistic model capable of describing the heavy tails of the risk-driving factor returns, the autocorrelation, clustering of the volatility, and the dependence between these factors. Processing large portfolios is related to manipulation of colossal data structures, which requires excellent skills of software developers in order to be efficiently performed.

BACK-TESTING OF VaR

If we adopt VaR for analysis of portfolio exposure, then a reasonable question is whether the VaR calculated according to any of the methods discussed in the previous section is realistic. Suppose that we calculate the 99% daily portfolio VaR. This means that according to our assumption for the portfolio return (payoff) distribution, the portfolio loses more than the 99% daily VaR with 1% probability. The question is whether this estimate is correct; that is, does the portfolio really lose more than this amount with 1% probability? This question can be answered by back-testing of VaR.

Generally, the procedure consists of the following steps.

- Step 1. Choose a time window for the back-testing. Usually the time window is the most recent one or two years.
- Step 2. For each day in the time window, calculate the VaR number.

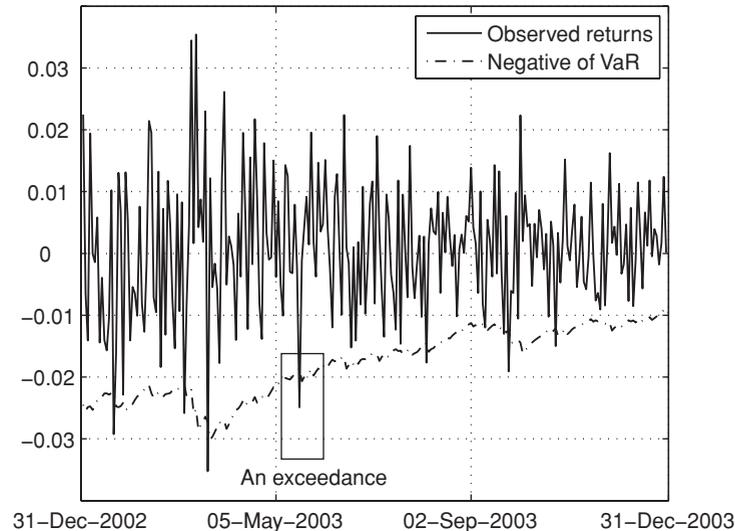


Figure 3 The Observed Daily Returns of the S&P 500 Index between December 31, 2002 and December 31, 2003 and the Negative of VaR

Note: The marked observation is an example of an exceedance.

Step 3. Check if the loss on a given day is below or above the VaR number computed the day before. If the observed loss is larger, then we say that there is a case of an exceedance. Figure 3 provides an example.

Step 4. Count the number of exceedances. Check if there are too many or too few of them by verifying if the number of exceedances belong to the corresponding 95% confidence interval.

If in Step 4 we find out that there are too many exceedances, then the VaR numbers produced by the model are too optimistic. Losses exceeding the corresponding VaR happen too frequently. If capital reserves are determined on the basis of VaR, then there is a risk of being incapable of covering large losses. Conversely, if we find out that there are too few exceedances, then the VaR numbers are too pessimistic. This is also an undesirable situation. Note that the actual size of the exceedances is immaterial; we only count them.

The confidence interval for the number of exceedances is constructed on the basis of the indicator-type events “we observe an exceedance,” “we do not observe an exceedance”

on a given day. If we consider the 99% VaR, then the probability of the first event, according to the model, is 1%. Let us associate a number with each of the events similar to a coin tossing experiment. If we observe an exceedance on a given day, then we say that the number 1 has occurred, otherwise 0 has occurred. If the backtesting time window is two years, then we have a sequence of 500 zeros and ones and the expected number of exceedances is 5. Thus, finding the 95% confidence interval for the number of exceedances reduces to finding an interval around 5 such that the probability of the number of ones belonging to this interval is 95%.

If we assume that the corresponding events are independent, then there is a complete analogue of this problem in terms of coin tossing. We toss an unfair coin independently 500 times with probability of success equal to 1%. What is the range of the number of success events with 95% probability? In order to find the 95% confidence interval, we can resort to the normal approximation to the binomial distribution. The formula is

$$\begin{aligned} \text{left bound} &= N\epsilon - F^{-1}(1 - 0.05/2)\sqrt{N\epsilon(1 - \epsilon)} \\ \text{right bound} &= N\epsilon + F^{-1}(1 - 0.05/2)\sqrt{N\epsilon(1 - \epsilon)} \end{aligned}$$

where N is the number of indicator-type events, ϵ is the tail probability of the VaR, and $F^{-1}(t)$ is the inverse distribution function of the standard normal distribution. In the example, $N = 500$, $\epsilon = 0.01$, and the 95% confidence interval for the number of exceedances is $[0, 9]$. Similarly, if we are back-testing the 95% VaR, under the same circumstances the confidence interval is $[15, 34]$.

Note that the statistical test based on the back-testing of VaR at a certain tail probability cannot answer the question if the distributional assumptions for the risk-driving factors are correct in general. For instance, if the portfolio contains only common stocks, then we presume a probabilistic model for stocks returns. By back-testing the 99% daily VaR of portfolio return, we verify if the probabilistic model is adequate for the 1% quantile of the portfolio return distribution; that is, we are back-testing if a certain point in the left tail of the portfolio return distribution is sufficiently accurately modeled. This should not be confused with statistical tests such as the Kolmogorov test or the Kolmogorov-Smirnov test, which concern accepting or rejecting a given distributional hypothesis.

COHERENT RISK MEASURES

Even though VaR has an intuitive interpretation and has been widely adopted as a risk measure, it does not always satisfy the important property that the VaR of a portfolio should not exceed the sum of the VaRs of the portfolio positions. This means that VaR is not always capable of representing the diversification effect.

This fact raises an important question. Can we find a set of desirable properties that a risk measure should satisfy? An answer is given by Artzner et al. (1998). They provide an axiomatic definition of a functional, which they call a *coherent risk measure*. The axioms follow with remarks given below each axiom. We denote the risk measure by the functional $\rho(X)$ assigning a real-valued number to a random

variable. Usually, the random variable X is interpreted as a random payoff and the motivation for the axioms in Artzner et al. (1998) follows this interpretation. In the remarks below each axiom, we provide an alternative interpretation, which holds if X is interpreted as random return.

The Monotonicity Property

Monotonicity $\rho(Y) \leq \rho(X)$,
if $Y \geq X$ in almost sure sense

Monotonicity states that if investment A has random return (payoff) Y , which is not less than the return (payoff) X of investment B at a given horizon in all states of the world, then the risk of A is not greater than the risk of B. This is quite intuitive but it really does matter whether the random variables represent random return or profit because an inequality in an almost sure sense between random returns may not translate into the same inequality between the corresponding random profits and vice versa.

Suppose that X and Y describe the random percentage returns on two investments A and B and let $Y = X + 3\%$. Apparently, $Y > X$ in all states of the world. The corresponding payoffs are obtained according to the equations

$$\text{Payoff}(X) = I_A(1 + X)$$

$$\text{Payoff}(Y) = I_B(1 + Y) = I_B(1 + X + 3\%)$$

where I_A is the initial investment in opportunity A and I_B is the initial investment in opportunity B. If the initial investment I_A is much larger than I_B , then $\text{Payoff}(X) > \text{Payoff}(Y)$ irrespective of the inequality $Y > X$. In effect, investment A may seem less risky than investment B in terms of payoff but in terms of return, the converse may hold.

The Positive Homogeneity Property

Positive Homogeneity $\rho(0) = 0$, $\rho(\lambda X) = \lambda\rho(X)$,
for all X and all $\lambda > 0$

The *positive homogeneity property* states that scaling the return (payoff) of the portfolio by

a positive factor scales the risk by the same factor. The interpretation for payoffs is obvious—if the investment in a position doubles, so does the risk of the position. We give a simple example illustrating this property when X stands for a random percentage return.

Suppose that today the value of a portfolio is I_0 and we add a certain amount of cash C . The value of our portfolio becomes $I_0 + C$. The value tomorrow is random and equals $I_1 + C$ in which I_1 is the random payoff. The return of the portfolio equals

$$\begin{aligned} X &= \frac{I_1 + C - I_0 - C}{I_0 + C} = \frac{I_1 - I_0}{I_0} \left(\frac{I_0}{I_0 + C} \right) \\ &= h \frac{I_1 - I_0}{I_0} = hY \end{aligned}$$

where $h = I_0/(I_0 + C)$ is a positive constant. The axiom positive homogeneity property implies that $\rho(X) = h\rho(Y)$; that is, the risk of the new portfolio will be the risk of the portfolio without the cash but scaled by h .

The Subadditivity Property

$$\begin{aligned} \text{Subadditivity} \quad \rho(X + Y) &\leq \rho(X) + \rho(Y), \\ &\text{for all } X \text{ and } Y \end{aligned}$$

If X and Y describe random payoffs, then the subadditivity property states that the risk of the portfolio is not greater than the sum of the risks of the two random payoffs.

The positive homogeneity property and the subadditivity property imply that the functional is convex

$$\begin{aligned} \rho(\lambda X + (1 - \lambda)Y) &\leq \rho(\lambda X) + \rho((1 - \lambda)Y) \\ &= \lambda\rho(X) + (1 - \lambda)\rho(Y) \end{aligned}$$

where $\lambda \in [0, 1]$. If X and Y describe random returns, then the random quantity $\lambda X + (1 - \lambda)Y$ stands for the return of a portfolio composed of two financial instruments with returns X and Y having weights λ and $1 - \lambda$ respectively. Therefore, the convexity property states that the risk of a portfolio is not greater than the sum of the risks of its constituents, meaning that it is the

convexity property that is behind the diversification effect that we expect in the case of X and Y denoting random returns.

The Invariance Property

$$\begin{aligned} \text{Invariance} \quad \rho(X + C) &= \rho(X) - C, \\ &\text{for all } X \text{ and } C \in \mathbb{R} \end{aligned}$$

The invariance property has various labels. Originally, it was called translation invariance while in other texts it is called cash invariance.² If X describes a random payoff, then the invariance property suggests that adding cash to a position reduces its risk by the amount of cash added. This is motivated by the idea that the risk measure can be used to determine capital requirements. As a consequence, the risk measure $\rho(X)$ can be interpreted as the minimal amount of cash necessary to make the position free of any capital requirements

$$\rho(X + \rho(X)) = 0$$

The invariance property has a different interpretation when X describes random return. Suppose that the random variable X describes the return of a common stock and we build a long-only portfolio by adding a government bond yielding a risk-free rate r_B . The portfolio return equals $wX + (1 - w)r_B$, where $w \in [0, 1]$ is the weight of the common stock in the portfolio. Note that the quantity $(1 - w)r_B$ is nonrandom by assumption. The invariance property states that the risk of the portfolio can be decomposed as

$$\begin{aligned} \rho(wX + (1 - w)r_B) &= \rho(wX) - (1 - w)r_B \\ &= w\rho(X) - (1 - w)r_B \end{aligned} \tag{9}$$

where the second equality appears because of the positive homogeneity property. In effect, the risk measure admits the following interpretation: Assume that the constructed portfolio is equally weighted, that is, $w = 1/2$, then the risk measure equals the level of the risk-free rate such that the risk of the equally weighted

portfolio consisting of the risky asset and the risk-free asset is zero. The investment in the risk-free asset will be, effectively, the reserve investment.

Alternative interpretations are also possible. Suppose that the present value of the position with random percentage return X is I_0 . Assume that we can find a government security earning return r_B^* at the horizon of interest. Then we can ask the question in the opposite direction: How much should we reallocate from I_0 and invest in the government security in order to hedge the risk $\rho(X)$? The needed capital C should satisfy the equation

$$\frac{I_0 - C}{I_0} \rho(X) - \frac{C}{I_0} r_B^* = 0$$

which is merely a restatement of equation (9) with the additional requirement that the risk of the resulting portfolio should be zero. The solution is

$$C = I_0 \frac{\rho(X)}{\rho(X) + r_B^*}$$

Note that if in the invariance property the constant is nonnegative, $C \geq 0$, then it follows that $\rho(X + C) \leq \rho(X)$. This result is in agreement with the monotonicity property as $X + C \geq X$. In fact, the invariance property can be regarded as an extension of the monotonicity property when the only difference between X and Y is in their means.

According to the discussion in the previous section, VaR is not a coherent risk measure because it may violate the subadditivity property.

An example of a coherent risk measure is the Average Value-at-Risk (AVaR), defined as the average of the VaRs that are larger than the VaR at a given tail probability ϵ . The accepted notation is $AVaR_\epsilon(X)$ in which ϵ stands for the tail probability level. A larger family of coherent risk measures is the family of spectral risk measures, which includes the AVaR as a representative. The spectral risk measures are defined as weighted averages of VaRs.

KEY POINTS

- VaR is defined as the minimum level of loss at a given, sufficiently high confidence level for a predefined time horizon.
- The performance of VaR, as well as any other risk measure, is heavily dependent on the assumed probability distribution for the economic measure of interest.
- Despite VaR's wide acceptance in the finance industry, it has important deficiencies so that, in general, it should be abandoned as a risk measure. However, it is not a useless concept to be abandoned altogether. For example, it can be used in risk reporting only as a characteristic of the portfolio return (pay-off) distribution since it has a straightforward interpretation.
- The most important drawback of VaR is that, in some cases, the reasonable diversification effect that every portfolio manager should expect to see in a risk measure is not present.
- The criticism of VaR is focused on its wide application by practitioners as a true risk measure, which, in view of its deficiencies, is not well grounded and should be reconsidered.
- Three approaches for portfolio VaR calculation that are used in practice are the Risk-Metrics approach, the historical method approach, and the Monte Carlo approach.

NOTES

1. A boxplot, or a box-and-whiskers diagram, is a convenient way of depicting several statistical characteristics of the sample. The size of the box equals the difference between the third and the first quartile (75% quantile–25% quantile), also known as the interquartile range. The line in the box corresponds to the median of the data (50% quantile). The lines extending out of the box are called whiskers and each of them is long up to 1.5 times the interquartile range. All

observations outside the whiskers are labeled outliers and are depicted by a plus sign.

2. This label can be found in Föllmer and Schied (2002).

REFERENCES

- Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1998). Coherent measures of risk. *Math. Fin.* 6: 203–228.
- Föllmer, H., and Schied, A. (2002). *Stochastic finance: An introduction in discrete time*, 2nd rev. and extended ed. Berlin: Walter de Gruyter.
- Lamantia, F., Ortobelli, S., and Rachev, S. (2006a). An empirical comparison among VaR models and time rules with elliptical and stable distributed returns. *Investment Management and Financial Innovations* 3: 8–29.
- Lamantia, F., Ortobelli, S., and Rachev, S. (2006b). VaR, CVaR and time rules with elliptical and asymmetric stable distributed returns. *Investment Management and Financial Innovations* 4: 19–39.

Average Value-at-Risk

STOYAN V. STOYANOV, PhD

Professor of Finance at EDHEC Business School and Head of Research for EDHEC Risk Institute-Asia

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair-Professor, Department of Applied Mathematics & Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Despite the fact that the value-at-risk (VaR) measure has been adopted as a standard risk measure in the financial industry, it has a number of deficiencies recognized by financial professionals. It is not a coherent risk measure. This is because it does not satisfy the subadditivity property requirement of a coherent risk measure. That is, there are cases in which the portfolio VaR is larger than the sum of the VaRs of the portfolio constituents, supporting the view that VaR cannot be used as a true risk measure. Unlike VaR, the average value-at-risk measure (AVaR)—also referred to as conditional value-at-risk and expected shortfall—is a coherent risk measure and has other advantages that result in its greater acceptance in risk modeling.

The *average value-at-risk* (AVaR) is a risk measure that is a superior alternative to VaR. Not only does it lack the deficiencies of VaR, but it also has an intuitive interpretation. There are convenient ways for computing and estimating AVaR, which allows its application in optimal portfolio problems. Moreover, it satisfies all axioms of *coherent risk measures* and it is consistent with the preference relations of risk-averse investors.

In this entry, we explore in detail the properties of AVaR and illustrate its superiority to VaR. We develop new geometric interpretations of AVaR and the various calculation methods.

We also provide closed-form expressions for the AVaR of the normal distribution, Student's t distribution, and a practical formula for Lévy stable distributions. Finally, we describe different estimation methods and remark on potential pitfalls.

AVERAGE VALUE-AT-RISK DEFINED

A disadvantage of VaR is that it does not give any information about the severity of losses beyond the VaR level. Consider the following example. Suppose that X and Y describe the

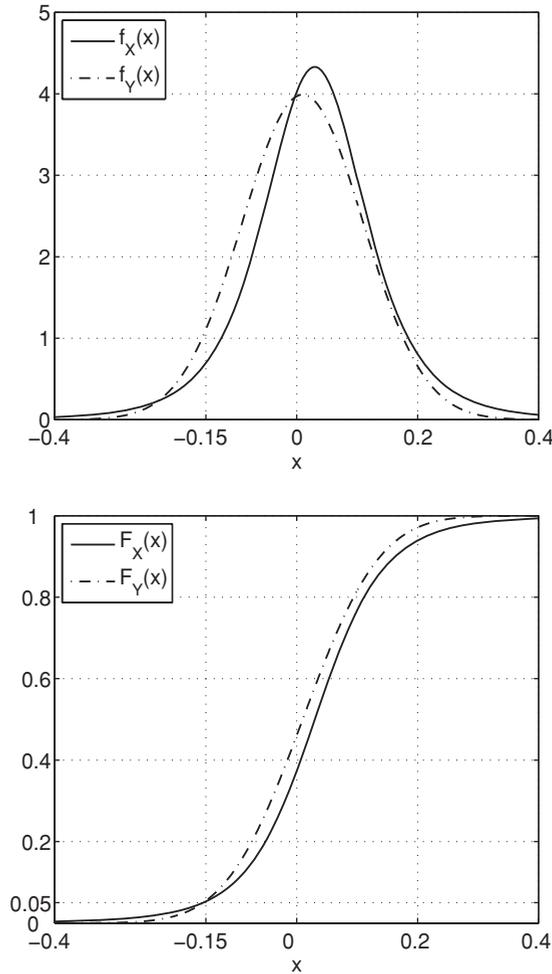


Figure 1 Note: The top plot shows the densities of X and Y and the bottom plot shows their c.d.f.s. The 95% VaRs of X and Y are equal to 0.15 but X has a thicker tail and is more risky.

random returns of two financial instruments with densities and distribution functions such as the ones in Figure 1. The expected returns are 3% and 1%, respectively. The standard deviations of X and Y are equal to 10%.¹ The cumulative distribution functions (c.d.f.s) $F_X(x)$ and $F_Y(x)$ cross at $x = -0.15$ and $F_X(-0.15) = F_Y(-0.15) = 0.05$. According to the definition of VaR, the 95% VaRs of both X and Y are equal to 15%. That is, the two financial instruments lose more than 15% of their present values with probability of 5%. In effect, we may conclude

that their risks are equal because their 95% VaRs are equal.

This conclusion is wrong because we pay no attention to the losses that are larger than the 95% VaR level. It is visible in Figure 1 that the left tail of X is heavier than the left tail of Y .² Therefore, it is more likely that the losses of X will be larger than the losses of Y , on condition that they are larger than 15%. Thus, looking only at the losses occurring with probability smaller than 5%, the random return X is riskier than Y . Note that both X and Y have equal standard deviations. If we base the analysis on the standard deviation and the expected return, we would conclude that not only is the uncertainty of X equal to the uncertainty of Y , but X is actually preferable because of the higher expected return. In fact, we realize that it is exactly the opposite, which shows how important it is to ground the reasoning on a proper risk measure.

The disadvantage of VaR, that it is not informative about the magnitude of the losses larger than the VaR level, is not present in the risk measure known as *average value-at-risk*. In the literature, it is also called *conditional value-at-risk*³ or *expected shortfall* but we will use average value-at-risk (AVaR) as it best describes the quantity it refers to.

The AVaR at tail probability ϵ is defined as the average of the VaRs, which are larger than the VaR at tail probability ϵ . Therefore, by construction, the AVaR is focused on the losses in the tail, which are larger than the corresponding VaR level. The average of the VaRs is computed through the integral

$$AVaR_\epsilon(X) = \frac{1}{\epsilon} \int_0^\epsilon VaR_p(X) dp \quad (1)$$

where $VaR_p(X)$ is defined by $VaR_\epsilon(X) = -\inf_x \{x \mid P(X \leq x) \geq \epsilon\} = -F_X^{-1}(\epsilon)$. As a matter of fact, the AVaR is not well defined for all real-valued random variables but only for those with finite mean; that is $AVaR_\epsilon(X) < \infty$ if $E|X| < \infty$. This should not be disturbing because random variables with infinite mathematical

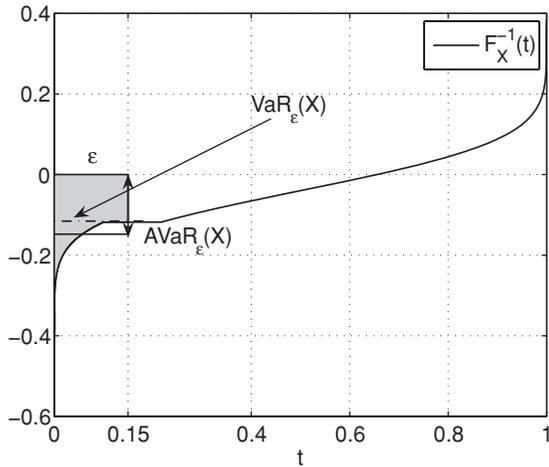


Figure 2 Geometrically, $AVaR_\epsilon(X)$ is the height for which the area of the drawn rectangle equals the shaded area closed between the graph of the inverse c.d.f. and the horizontal axis for $t \in [0, \epsilon]$. The $VaR_\epsilon(X)$ value is shown by a dash-dotted line.

expectation have limited application in the field of finance. For example, if such a random variable is used for a model of stock returns, then it is assumed that the common stock has infinite expected return, which is not realistic.

The AVaR satisfies all the axioms of coherent risk measures. One consequence is that, unlike VaR, it is convex for all possible portfolios, which means that it always accounts for the diversification effect.

A geometric interpretation of the definition in equation (1) is provided in Figure 2. In this figure, the inverse c.d.f. of a random variable X is plotted. The shaded area is closed between the graph of $F_X^{-1}(t)$ and the horizontal axis for $t \in [0, \epsilon]$ where ϵ denotes the selected tail probability. $AVaR_\epsilon(X)$ is the value for which the area of the drawn rectangle, equal to $\epsilon \times AVaR_\epsilon(X)$, coincides with the shaded area, which is computed by the integral in equation (1). The $VaR_\epsilon(X)$ value is always smaller than $AVaR_\epsilon(X)$. In Figure 2, $VaR_\epsilon(X)$ is shown by a dash-dotted line and is indicated by an arrow.

Let us revisit the example developed at the beginning of this section. We concluded that even

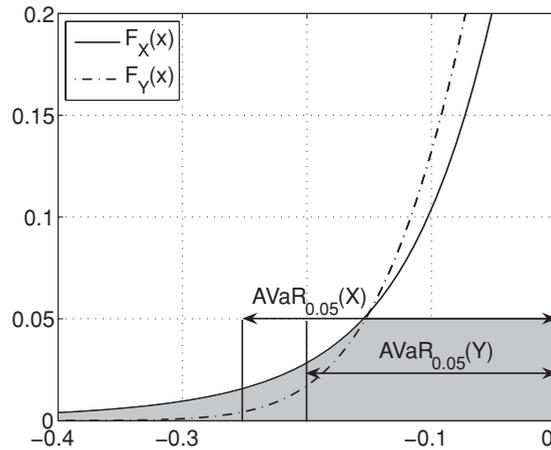


Figure 3 The AVaRs of the Return Distributions from Figure 1 in Line with the Geometric Intuition *Note:* Even though the 95% VaRs are equal, the AVaRs at 5% tail probability differ, $AVaR_{0.05}(X) > AVaR_{0.05}(Y)$.

though the VaRs at 5% tail probability of both random variables are equal, X is riskier than Y because the left tail of X is heavier than the left tail of Y ; that is, the distribution of X is more likely to produce larger losses than the distribution of Y on condition that the losses are beyond the VaR at the 5% tail probability. We apply the geometric interpretation illustrated in Figure 2 to this example. First, notice that the shaded area in Figure 2, which concerns the graph of the inverse of the c.d.f., can also be identified through the graph of the c.d.f. This is done in Figure 3, which shows a magnified section of the left tails of the c.d.f.s plotted in Figure 1. The shaded area appears as the intersection of the area closed below the graph of the distribution function and the horizontal axis, and the area below a horizontal line shifted at the tail probability above the horizontal axis. In Figure 3, we show the area for $F_X(x)$ at 5% tail probability. The corresponding area for $F_Y(x)$ is smaller because $F_Y(x) \leq F_X(x)$ to the left of the crossing point of the two c.d.f.s, which is exactly at 5% tail probability.

In line with the geometric interpretation, the $AVaR_{0.05}(X)$ is a number such that if we draw

a rectangle with height 0.05 and width equal to $AVaR_{0.05}(X)$, the area of the rectangle ($0.05 \times AVaR_{0.05}(X)$) equals the shaded area in Figure 3. The same exercise for $AVaR_{0.05}(Y)$ shows that $AVaR_{0.05}(Y) < AVaR_{0.05}(X)$ because the corresponding shaded area is smaller and both rectangles share a common height of 0.05.

Besides the definition in equation (1), AVaR can be represented through a minimization formula,⁴

$$AVaR_{\epsilon}(X) = \min_{\theta \in \mathbb{R}} \left(\theta + \frac{1}{\epsilon} E(-X - \theta)_+ \right) \quad (2)$$

where $(x)_+$ denotes the maximum between x and zero, $(x)_+ = \max(x, 0)$ and X describes the portfolio return distribution. It turns out that this formula has an important application in optimal portfolio problems based on AVaR as a risk measure. In the appendix to this entry, we provide an illuminating geometric interpretation of equation (2), which shows the connection to the definition of AVaR.

How can we compute the AVaR for a given return distribution? Throughout this section, we assume that the return distribution function is a continuous function, that is, there are no point masses. Under this condition, after some algebra and using the fact that VaR is the negative of a certain quantile, we obtain that the AVaR can be represented in terms of a conditional expectation,

$$\begin{aligned} AVaR_{\epsilon}(X) &= -\frac{1}{\epsilon} \int_0^{\epsilon} F_X^{-1}(t) dt \\ &= -E(X | X < -VaR_{\epsilon}(X)) \end{aligned} \quad (3)$$

which is called expected tail loss (ETL) and is denoted by $ETL_{\epsilon}(X)$. The conditional expectation implies that the AVaR equals the average loss provided that the loss is larger than the VaR level. In fact, the average of VaRs in equation (1) equals the average of losses in equation (3) only if the c.d.f. of X is continuous at $x = VaR_{\epsilon}(X)$. If there is a discontinuity, or a point mass, the relationship is more involved. The general formula is given in the appendix to this entry.

Equation (3) implies that AVaR is related to the conditional loss distribution. In fact, under certain conditions, it is the mathematical expectation of the conditional loss distribution, which represents only one characteristic of it. In the appendix to this entry, we introduce several sets of characteristics of the conditional loss distribution, which provide a more complete picture of it. Also, in the appendix, we introduce the more general concept of higher-order AVaR.

For some continuous distributions, it is possible to calculate explicitly the AVaR through equation (3). We provide the closed-form expressions for the normal distribution and Student's t distribution. In the appendix to this entry, we give a semi-explicit formula for the class of stable distributions.

1. The normal distribution

Suppose that X is distributed according to a normal distribution with standard deviation σ_X and mathematical expectation EX . The AVaR of X at tail probability ϵ equals

$$AVaR_{\epsilon}(X) = \frac{\sigma_X}{\epsilon \sqrt{2\pi}} \exp\left(-\frac{(VaR_{\epsilon}(Y))^2}{2}\right) - EX \quad (4)$$

where Y has the standard normal distribution, $Y \in N(0,1)$.

2. The Student's t distribution

Suppose that X has Student's t distribution with ν degrees of freedom, $X \in t(\nu)$. The AVaR of X at tail probability ϵ equals

$$AVaR_{\epsilon}(X) = \begin{cases} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{\sqrt{\nu}}{(\nu-1)\epsilon\sqrt{\pi}} \left(1 + \frac{(VaR_{\epsilon}(X))^2}{\nu}\right)^{\frac{1-\nu}{2}}, & \nu > 1 \\ \infty, & \nu = 1 \end{cases}$$

where the notation $\Gamma(x)$ stands for the gamma function. It is not surprising that for $\nu = 1$ the AVaR explodes because the Student's t distribution with one degree of freedom, also known as the Cauchy distribution, has infinite mathematical expectation.⁵

Note that equation (4) can be represented in a more compact way,

$$AVaR_\epsilon(X) = \sigma_X C_\epsilon - E X \tag{5}$$

where C_ϵ is a constant which depends only on the tail probability ϵ . Therefore, the AVaR of the normal distribution has the same structure as the normal VaR—the difference between the properly scaled standard deviation and the mathematical expectation. In effect, similar to the normal VaR, the normal AVaR properties are dictated by the standard deviation. Even though AVaR is focused on the extreme losses only, due to the limitations of the normal assumption, it is symmetric.

Exactly the same conclusion holds for the AVaR of Student's t distribution. The true merits of AVaR become apparent if the underlying distributional model is skewed.

AVaR ESTIMATION FROM A SAMPLE

Suppose that we have a sample of observed portfolio returns and we are not aware of their distribution. Provided that we do not impose any distributional model, the AVaR of portfolio return can be estimated from the sample of observed portfolio returns. Denote the observed portfolio returns by r_1, r_2, \dots, r_n at time instants t_1, t_2, \dots, t_n . The numbers in the sample are given in order of observation. Denote the sorted sample by $r_{(1)} \leq r_{(2)} \leq \dots \leq r_{(n)}$. Thus, $r_{(1)}$ equals the smallest observed portfolio return and $r_{(n)}$ is the largest. The AVaR of portfolio returns at tail probability ϵ is estimated according to the formula⁶

$$\widehat{AVaR}_\epsilon(r) = -\frac{1}{\epsilon} \left(\frac{1}{n} \sum_{k=1}^{\lceil n\epsilon \rceil - 1} r_{(k)} + \left(\epsilon - \frac{\lceil n\epsilon \rceil - 1}{n} \right) r_{(\lceil n\epsilon \rceil)} \right) \tag{6}$$

where the notation $\lceil x \rceil$ stands for the smallest integer larger than x .⁷ The “hat” above AVaR denotes that the number calculated by equation (6) is an estimate of the true value because it is

based on a sample. This is a standard notation in statistics.

We demonstrate how equation (6) is applied in the following example. Suppose that the sorted sample of portfolio returns is -1.37% , -0.98% , -0.38% , -0.26% , 0.19% , 0.31% , 1.91% and our goal is to calculate the portfolio AVaR at 30% tail probability. In this case, the sample contains 7 observations and $(n\epsilon) = (7 \times 0.3) = 3$. According to equation (6), we calculate

$$\begin{aligned} \widehat{AVaR}_{0.3}(r) &= -\frac{1}{0.3} \left(\frac{1}{7} (-1.37\% - 0.98\%) \right. \\ &\quad \left. + (0.3 - 2/7)(-0.38\%) \right) \\ &= 1.137\%. \end{aligned}$$

Formula (6) can be applied not only to a sample of empirical observations. We may want to work with a statistical model for which no closed-form expressions for AVaR are known. Then we can simply sample from the distribution and apply formula (6) to the generated simulations.

Besides formula (6), there is another method for calculation of AVaR. It is based on the minimization formula (2) in which we replace the mathematical expectation by the sample average,

$$\widehat{AVaR}_\epsilon(r) = \min_{\theta \in \mathbb{R}} \left(\theta + \frac{1}{n\epsilon} \sum_{i=1}^n \max(-r_i - \theta, 0) \right) \tag{7}$$

Even though it is not obvious, equations (6) and (7) are completely equivalent.

The minimization formula in equation (7) is appealing because it can be calculated through the methods of linear programming. It can be restated as a linear optimization problem by introducing auxiliary variables d_1, \dots, d_n , one for each observation in the sample,

$$\begin{aligned} &\min_{\theta, d} \quad \theta + \frac{1}{n\epsilon} \sum_{k=1}^n d_k \\ &\text{subject to} \quad -r_k - \theta \leq d_k, \quad k = 1, n \\ &\quad \quad \quad d_k \geq 0, \quad k = 1, n \\ &\quad \quad \quad \theta \in \mathbb{R} \end{aligned} \tag{8}$$

The linear problem (8) is obtained from (7) through standard methods in mathematical programming. We briefly demonstrate the equivalence between them. Let us fix the value of θ to θ^* . Then the following choice of the auxiliary variables yields the minimum in (8). If $-r_k - \theta^* < 0$, then $d_k = 0$. Conversely, if it turns out that $-r_k - \theta^* \geq 0$, then $-r_k - \theta^* = d_k$. In this way, the sum in the objective function becomes equal to the sum of maxima in equation (7).

Applying (8) to the sample in the example above, we obtain the optimization problem

$$\begin{aligned} \min_{\theta, d} \quad & \theta + \frac{1}{7 \times 0.3} \sum_{k=1}^7 d_k \\ \text{subject to} \quad & 0.98\% - \theta \leq d_1 \\ & -0.31\% - \theta \leq d_2 \\ & -1.91\% - \theta \leq d_3 \\ & 1.37\% - \theta \leq d_4 \\ & 0.38\% - \theta \leq d_5 \\ & 0.26\% - \theta \leq d_6 \\ & -0.19\% - \theta \leq d_7 \\ & d_k \geq 0, \quad k = 1, 7 \\ & \theta \in \mathbb{R} \end{aligned}$$

The solution to this optimization problem is the number 1.137%, which is attained for $\theta = 0.38\%$. In fact, this value of θ coincides with the VaR at 30% tail probability and this is not by chance but a feature of the problem, which is demonstrated in the appendix to this entry. We verify that the solution of the problem is indeed the number 1.137% by calculating the objective in equation (7) for $\theta = 0.38\%$,

$$\begin{aligned} AVaR_\epsilon(r) &= 0.38\% + \frac{0.98\% - 0.38\% + 1.37\% - 0.38\%}{7 \times 0.3} \\ &= 1.137\% \end{aligned}$$

Thus, we obtain the number calculated through equation (6).

COMPUTING PORTFOLIO AVaR IN PRACTICE

The ideas behind the approaches of VaR estimation are applied to AVaR. We assume that there

are n assets with random returns described by the random variables X_1, \dots, X_n . Thus, the portfolio return is represented by

$$r_p = w_1 X_1 + \dots + w_n X_n$$

where w_1, \dots, w_n are the weights of the assets in the portfolio.

The Multivariate Normal Assumption

If the asset returns are assumed to have a multivariate normal distribution, then the portfolio return has a normal distribution with variance $w' \Sigma w$, where w is the vector of weights and Σ is the covariance matrix between stock returns. The mean of the normal distribution is

$$Er_p = \sum_{k=1}^n w_k E X_k$$

where E stands for the mathematical expectation. Thus, under this assumption the AVaR of portfolio return at tail probability ϵ can be expressed in closed-form through equation (4),

$$\begin{aligned} AVaR_\epsilon(r_p) &= \frac{\sqrt{w' \Sigma w}}{\epsilon \sqrt{2\pi}} \exp\left(-\frac{(VaR_\epsilon(Y))^2}{2}\right) - Er_p \\ &= C_\epsilon \sqrt{w' \Sigma w} - Er_p \end{aligned} \quad (9)$$

where C_ϵ is a constant independent of the portfolio composition and can be calculated in advance. In effect, due to the limitations of the multivariate normal assumption, the portfolio AVaR appears symmetric and is representable as the difference between the properly scaled standard deviation of the random portfolio return and portfolio expected return.

The Historical Method

The historical method is not related to any distributional assumptions. We use the historically observed portfolio returns as a model for the future returns and apply formula (6) or (7).

The historical method has several drawbacks. It is very inaccurate for low tail probabilities, for example, 1% or 5%. Even with one year of daily

returns, which amounts to 250 observations, in order to estimate the AVaR at 1% probability, we have to use the 3 smallest observations, which is quite insufficient. What makes the estimation problem even worse is that these observations are in the tail of the distribution; that is, they are the smallest ones in the sample. The implication is that when the sample changes, the estimated AVaR may change a lot because the smallest observations tend to fluctuate a lot.

The Hybrid Method

According to the hybrid method, different weights are assigned to the observations by which the more recent observations get a higher weight. The rationale is that the observations far back in the past have less impact on the portfolio risk at the present time.

The hybrid method can be adapted for AVaR estimation. The weights assigned to the observations are interpreted as probabilities and, thus, the portfolio AVaR can be estimated from the resulting discrete distribution according to the formula

$$\widehat{AVaR}_\epsilon(r) = -\frac{1}{\epsilon} \left(\sum_{j=1}^{k_\epsilon} p_j r_{(j)} + \left(\epsilon - \sum_{j=1}^{k_\epsilon} p_j \right) r_{(k_\epsilon+1)} \right) \tag{10}$$

where $r_{(1)} \leq r_{(2)} \leq \dots \leq r_{(k_m)}$ denotes the sorted sample of portfolio returns or payoffs and p_1, p_2, \dots, p_{k_m} stand for the probabilities of the sorted observations; that is, p_1 is the probability of $r_{(1)}$. The number k_ϵ in equation (10) is an integer satisfying the inequalities

$$\sum_{j=1}^{k_\epsilon} p_j \leq \epsilon < \sum_{j=1}^{k_\epsilon+1} p_j$$

Equation (10) follows directly from the definition of AVaR⁸ under the assumption that the underlying distribution is discrete without the additional simplification that the outcomes are

equally probable. In the appendix to this entry, we demonstrate the connection between equation (10) and the definition of AVaR in equation (1).

The Monte Carlo Method

The Monte Carlo method assumes and estimates a multivariate statistical model for the asset return distribution. Then we sample from it, and we calculate scenarios for portfolio return. On the basis of these scenarios, we estimate portfolio AVaR using equation (6) in which r_1, \dots, r_n stands for the vector of generated scenarios.

Similar to the case of VaR, an artifact of the Monte Carlo method is the variability of the risk estimate. Since the estimate of portfolio AVaR is obtained from a generated sample of scenarios, by regenerating the sample, we will obtain a slightly different value.

Suppose that the portfolio daily return distribution is the standard normal law, $r_p \in N(0,1)$. By the closed-form expression in equation (4), we calculate that the AVaR of the portfolio at 1% tail probability equals

$$AVaR_{0.01}(r_p) = \frac{1}{0.01\sqrt{2\pi}} \exp\left(-\frac{2.326^2}{2}\right) = 2.665$$

In order to investigate how the fluctuations of the 99% AVaR change about the theoretical value, we generate samples of different sizes: 500, 1,000, 5,000, 10,000, 20,000, and 100,000 scenarios. The 99% AVaR is computed from these samples using equation 6 and the numbers are stored. We repeat the experiment 100 times. In the end, we have 100 AVaR numbers for each sample size. We expect that as the sample size increases, the AVaR values will fluctuate less about the theoretical value which is $AVaR_{0.01}(X) = 2.665, X \in N(0,1)$.

Panel A of Table 1 contains the result of the experiment. From the 100 AVaR numbers, we calculate the 95% confidence interval reported in the third column. The confidence intervals

Table 1 Confidence Intervals Calculated for AVaR and VaR

Number of Scenarios	AVaR at 99%	95% Confidence Interval
500	2.646	[2.2060, 2.9663]
1,000	2.771	[2.3810, 2.9644]
5,000	2.737	[2.5266, 2.7868]
10,000	2.740	[2.5698, 2.7651]
20,000	2.659	[2.5955, 2.7365]
50,000	2.678	[2.6208, 2.7116]
100,000	2.669	[2.6365, 2.6872]

Panel A: The 99% AVaR of the standard normal distribution computed from a sample of scenarios. The 95% confidence interval is calculated from 100 repetitions of the experiment. The true value is $AVaR_{0.01}(X) = 2.665$.

Number of Scenarios	99% VaR	95% Confidence Interval
500	2.067	[1.7515, 2.3825]
1,000	2.406	[2.1455, 2.6665]
5,000	2.286	[2.1875, 2.3845]
10,000	2.297	[2.2261, 2.3682]
20,000	2.282	[2.2305, 2.3335]
50,000	2.342	[2.3085, 2.3755]
100,000	2.314	[2.2925, 2.3355]

Panel B: The 99% VaR of the standard normal distribution computed from a sample of scenarios. The 95% confidence interval is calculated from 100 repetitions of the experiment. The true value is $VaR_{0.01}(X) = 2.326$.

cover the theoretical value 2.665 and also we notice that the length of the confidence interval decreases as the sample size increases. This effect is illustrated in Figure 4 with boxplot diagrams. A sample of 100,000 scenarios results in AVaR numbers, which are tightly packed around the true value while a sample of only 500 scenarios may give a very inaccurate estimate.

By comparing, Panel A of Table 1 to Panel B of the table, which shows the results for VaR, we notice that the length of the 95% confidence intervals for AVaR are larger than the corresponding confidence intervals for VaR. This result is not surprising. Given that both quantities are at the same tail probability of 1%, the AVaR has larger variability than the VaR for a fixed number of scenarios because the AVaR is the average of terms fluctuating more than the 1%

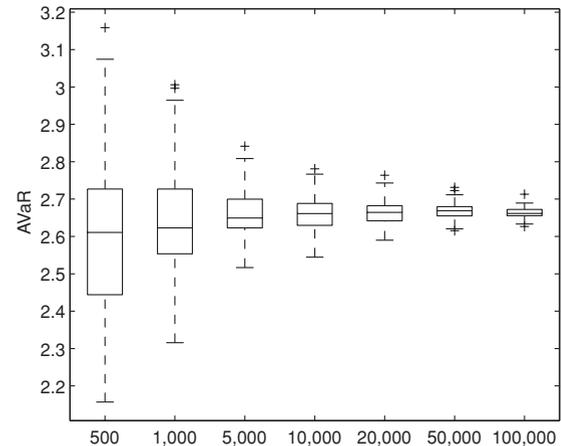


Figure 4 Boxplot Diagrams of the Fluctuation of the AVaR at 1% Tail Probability of the Standard Normal Distribution Based on Scenarios

Note: The horizontal axis shows the number of scenarios and the boxplots are computed from 100 independent samples.

VaR. This effect is more pronounced the more heavy-tailed the distribution is.

BACK-TESTING OF AVaR

Suppose that we have selected a method for calculating the daily AVaR of a portfolio. A reasonable question is how we can verify if the estimates of daily AVaR are realistic. This is done by back-testing. In the case of VaR back-testing consists of computing the portfolio VaR for each day back in time using the information available up to that day only. In this way, we have the VaR numbers back in time as if we had used exactly the same methodology in the past. On the basis of the VaR numbers and the realized portfolio returns, we can use statistical methods to assess whether the forecasted loss at the VaR tail probability is consistent with the observed losses. If there are too many observed losses larger than the forecasted VaR, then the model is too optimistic. Conversely, if there are too few losses larger than the forecasted VaR, then the model is too pessimistic.

Note that in the case of VaR back-testing, we are simply counting the cases in which there is an exceedance; that is, when the size of the observed loss is larger than the predicted VaR. The magnitude of the exceedance is immaterial for the statistical test.

Unlike VaR, back-testing of AVaR is not straightforward and is a much more challenging task. By definition, the AVaR at tail probability ϵ is the average of VaRs larger than the VaR at tail probability ϵ . Thus, the most direct approach to test AVaR would be to perform VaR back-tests at all tail probabilities smaller than ϵ . If all these VaRs are correctly modeled, then so is the corresponding AVaR.

One general issue with this approach is that it is impossible to perform in practice. Suppose that we consider the AVaR at tail probability of 1%, for example. Back-testing VaRs deeper in the tail of the distribution can be infeasible because the back-testing time window is too short. The lower the tail probability, the larger the time window we need in order for the VaR test to be conclusive. Another general issue is that this approach is too demanding. Even if the VaR back-testing fails at some tail probability ϵ_1 below ϵ , this does not necessarily mean that the AVaR is incorrectly modeled because the test failure may be due to purely statistical reasons and not to incorrect modeling.

These arguments illustrate why AVaR back-testing is a difficult problem—we need the information about the entire tail of the return distribution describing the losses larger than the VaR at tail probability ϵ and there may be too few observations from the tail upon which to base the analysis. For example, in one business year, there are typically 250 trading days. Therefore, a one-year back-testing results in 250 daily portfolio returns, which means that if $\epsilon = 1\%$, then there are only 2 observations available from the losses larger than the VaR at 1% tail probability.

As a result, in order to be able to back-test AVaR, we can assume a certain “structure” of

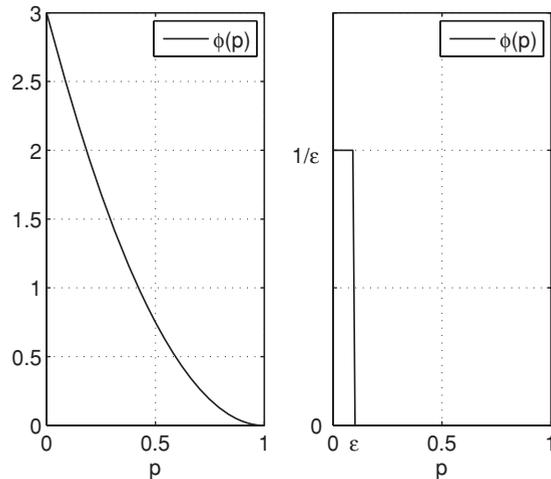


Figure 5 Examples of Risk-Aversion Functions
Note: The right plot shows the risk-aversion function yielding the AVaR at tail probability ϵ .

the tail of the return distribution that would compensate for the lack of observations. There are two general approaches:

1. Use the tails of the Lévy stable distributions as a proxy for the tail of the loss distribution and take advantage of the practical semi-analytic formula for the AVaR given in the appendix to this entry to construct a statistical test.
2. Make the weaker assumption that the loss distribution belongs to the domain of attraction of a max-stable distribution. Thus, the behavior of the large losses can be approximately described by the limit max-stable distribution and a statistical test can be based on it.

The rationale of the first approach is that, generally, the Lévy stable distribution provides a good fit to the stock returns data and, thus, the stable tail may turn out to be a reasonable approximation. Moreover, from the generalized central limit theorem we know that stable distributions have domains of attraction, which makes them an appealing candidate for an approximate model.

The second approach is based on a weaker assumption. The family of max-stable distributions arises as the limit distribution of properly scaled and centered maxima of IID random variables. If the random variable describes portfolio losses, then the limit max-stable distribution can be used as a model for the large losses (i.e., the ones in the tail). Unfortunately, as a result of the weaker assumption, estimators of poor quality have to be used to estimate the parameters of the limit max-stable distribution, such as the Hill estimator, for example. This represents the basic trade-off in this approach.

TECHNICAL APPENDIX

In this appendix, we start with a more general view that better describes the conditional loss distribution in terms of certain characteristics in which AVaR appears as a special case. We continue with the notion of higher-order AVaR, generating a family of coherent risk measures. Next, we provide an intuitive geometric interpretation of the minimization formula for the AVaR calculation. We also provide a semi-analytic expression for the AVaR of stable distributions and compare the expected tail loss measure to AVaR. Finally, we comment on the proper choice of a risk-aversion function in spectral risk measures, which does not result in an infinite risk measure.

Characteristics of Conditional Loss Distributions

In the entry, we defined AVaR as a risk measure and showed how it can be calculated in practice. While it is an intuitive and easy to use coherent risk measure, AVaR represents the average of the losses larger than the VaR at tail probability ϵ , which is only one characteristic of the distribution of extreme losses. We remarked that if the distribution function is continuous, then AVaR coincides with ETL, which is the mathematical expectation of the condi-

tional loss distribution. Besides the mathematical expectation, there are other important characteristics of the conditional loss distribution. For example, AVaR does not provide any information about how dispersed the conditional losses are around the AVaR value. In this section, we state a couple of families of useful characteristics in which AVaR appears as one example.

Consider the following tail moment of order n at tail probability ϵ ,

$$m_\epsilon^n(X) = \frac{1}{\epsilon} \int_0^\epsilon (F_X^{-1}(t))^n dt \quad (\text{A.1})$$

where $n = 1, 2, \dots$, $F_X^{-1}(t)$ is the inverse c.d.f. of the random variable X . If the distribution function of X is continuous, then the tail moment of order n can be represented through the following conditional expectation

$$m_\epsilon^n(X) = E(X^n | X < VaR_\epsilon(X)) \quad (\text{A.2})$$

where $n = 1, 2, \dots$. In the general case, if the c.d.f. has a jump at $VaR_\epsilon(X)$, a link exists between the conditional expectation and equation (A.1), which is similar to formula (A.12) later in this appendix for AVaR. In fact, AVaR appears as the negative of the tail moment of order one, $AVaR_\epsilon(X) = -m_\epsilon^1(X)$.

The higher-order tail moments provide additional information about the conditional distribution of the extreme losses. We can make a parallel with the way the moments of a random variable are used to describe certain properties of it. In our case, it is the conditional distribution that we are interested in.

In addition to the moments $m_\epsilon^n(X)$, we introduce the central tail moments of order n at tail probability ϵ ,

$$M_\epsilon^n(X) = \frac{1}{\epsilon} \int_0^\epsilon (F_X^{-1}(t) - m_\epsilon^1(X))^n dt \quad (\text{A.3})$$

where $m_\epsilon^1(X)$ is the tail moment of order one. If the distribution function is continuous, then the central moments can be expressed in terms of the conditional expectation

$$M_\epsilon^n(X) = E((X - m_\epsilon^1(X))^n | X < VaR_\epsilon(X))$$

The tail variance of the conditional distribution appears as $M_\epsilon^2(X)$ and the tail standard deviation equals

$$(M_\epsilon^2(X))^{1/2} = \left(\frac{1}{\epsilon} \int_0^\epsilon (F_X^{-1}(t) - m_\epsilon^1(X))^2 dt \right)^{1/2}$$

There is a formula expressing the tail variance in terms of the tail moments introduced in (A.2),

$$\begin{aligned} M_\epsilon^2(X) &= m_\epsilon^2(X) - (m_\epsilon^1(X))^2 \\ &= m_\epsilon^2(X) - (AVaR_\epsilon(X))^2 \end{aligned}$$

This formula is similar to the representation of variance in terms of the first two moments,

$$\sigma_X^2 = EX^2 - (EX)^2$$

The tail standard deviation can be used to describe the dispersion of conditional losses around AVaR as it satisfies the general properties of dispersion measures. It can be viewed as complementary to AVaR in the sense that if there are two portfolios with equal AVaRs of their return distributions but different tail standard deviations, the portfolio with the smaller standard deviation is preferable.

Another central tail moment that can be interpreted is $M_\epsilon^3(X)$. After proper normalization, it can be employed to measure the skewness of the conditional loss distribution. In fact, if the tail probability is sufficiently small, the tail skewness will be quite significant. In the same fashion, by normalizing the central tail moment of order 4, we obtain a measure of kurtosis of the conditional loss distribution.

In a similar way, we introduce the absolute central tail moments of order n at tail probability ϵ ,

$$\mu_\epsilon^n(X) = \frac{1}{\epsilon} \int_0^\epsilon |F_X^{-1}(t) - m_\epsilon^1(X)|^n dt \quad (A.4)$$

The tail moments $\mu_\epsilon^n(X)$ raised to the power of $1/n$, $(\mu_\epsilon^n(X))^{1/n}$, can be applied as measures of dispersion of the conditional loss distribution if the distribution is such that they are finite.

In the entry, we remarked that the tail of the random variable can be so heavy that AVaR becomes infinite. Even if it is theoretically finite, it can be hard to estimate because the heavy

tail will result in the AVaR estimator having a large variability. Thus, under certain conditions it may turn out to be practical to employ a robust estimator instead. The median tail loss (MTL), defined as the median of the conditional loss distribution, is a robust alternative to AVaR. It has the advantage of always being finite no matter the tail behavior of the random variable. Formally, it is defined as

$$MTL_\epsilon(X) = -F_X^{-1}(1/2|X < -VaR_\epsilon(X)) \quad (A.5)$$

where $F_X^{-1}(p|X < -VaR_\epsilon(X))$ stands for the inverse distribution function of the c.d.f. of the conditional loss distribution

$$\begin{aligned} F_X(x|X < -VaR_\epsilon(X)) &= P(X \leq x|X < -VaR_\epsilon(X)) \\ &= \begin{cases} P(X \leq x)/\epsilon, & x < -VaR_\epsilon(X) \\ 1, & x \geq -VaR_\epsilon(X) \end{cases} \end{aligned}$$

In effect, MTL, as well as any other quantile of the conditional loss distribution, can be directly calculated as a quantile of the distribution of X ,

$$\begin{aligned} MTL_\epsilon(X) &= -F_X^{-1}(\epsilon/2) \\ &= VaR_{\epsilon/2}(X) \end{aligned} \quad (A.6)$$

where $F_X^{-1}(p)$ is the inverse c.d.f. of X and ϵ is the tail probability of the corresponding VaR in equation (A.5). Thus, MTL shares the properties of VaR. Equation (A.6) shows that MTL is not a coherent risk measure even though it is a robust alternative to AVaR, which is a coherent risk measure.

In the universe of the three families of moments that we introduced, AVaR is one special case providing only limited information. It may be the only coherent risk measure among them but the other moments can be employed in addition to AVaR in order to gain more insight into the conditional loss distribution. Furthermore, it could appear that other reasonable risk measures can be based on some of the moments. Thus, we believe that they all should be considered in financial applications.

Higher-Order AVaR

By definition, AVaR is the average of VaRs larger than the VaR at tail probability ϵ . In the same fashion, we can pose the question of what happens if we average all AVaRs larger than the AVaR at tail probability ϵ . In fact, this quantity is an average of coherent risk measures and, therefore, is a coherent risk measure itself since it satisfies all defining properties of coherent risk measures. We call it *AVaR of order one* and denote it by $AVaR_\epsilon^{(1)}(X)$ because it is a derived quantity from AVaR. In this section, we consider similar derived quantities from AVaR which we call higher-order AVaRs.

Formally, the AVaR of order one is represented in the following way

$$AVaR_\epsilon^{(1)} = \frac{1}{\epsilon} \int_0^\epsilon AVaR_p(X) dp$$

where $AVaR_p(X)$ is the AVaR at tail probability p . Replacing AVaR by the definition given in equation (1), we obtain

$$\begin{aligned} AVaR_\epsilon^{(1)} &= -\frac{1}{\epsilon} \int_0^\epsilon \left(\int_0^1 F_X^{-1}(y) g_p(y) dy \right) dp \\ &= -\frac{1}{\epsilon} \int_0^1 F_X^{-1}(y) \left(\int_0^\epsilon g_p(y) dp \right) dy \end{aligned}$$

where

$$g_p(y) = \begin{cases} 1/p, & y \in [0, p] \\ 0, & y > p \end{cases}$$

and after certain algebraic manipulations, we get the expression

$$\begin{aligned} AVaR_\epsilon^{(1)}(X) &= -\frac{1}{\epsilon} \int_0^\epsilon F_X^{-1}(y) \log \frac{\epsilon}{y} dy \\ &= \int_0^\epsilon VaR_y(X) \phi_\epsilon(y) dy \quad (\text{A.7}) \end{aligned}$$

In effect, the AVaR of order one can be expressed as a weighted average of VaRs larger than the VaR at tail probability ϵ with a weighting function $\phi_\epsilon(y)$ equal to

$$\phi_\epsilon(y) = \begin{cases} \frac{1}{\epsilon} \log \frac{\epsilon}{y}, & 0 \leq y \leq \epsilon \\ 0, & \epsilon < y \leq 1 \end{cases}$$

The AVaR of order one can be viewed as a spectral risk measure with $\phi_\epsilon(y)$ being the risk aversion function.

Similarly, we define the higher-order AVaR through the recursive equation

$$AVaR_\epsilon^{(n)}(X) = \frac{1}{\epsilon} \int_0^\epsilon AVaR_p^{(n-1)}(X) dp \quad (\text{A.8})$$

where $AVaR_p^{(0)}(X) = AVaR_p(X)$ and $n = 1, 2, \dots$. Thus, the AVaR of order two equals the average of AVaRs of order one, which are larger than the AVaR of order one at tail probability ϵ . The AVaR of order n appears as an average of AVaRs of order $n - 1$.

The quantity $AVaR_\epsilon^{(n)}(X)$ is a coherent risk measure because it is an average of coherent risk measures. This is a consequence of the recursive definition in (A.8). It is possible to show that AVaR of order n admits the representation

$$AVaR_\epsilon^{(n)}(X) = \frac{1}{\epsilon} \int_0^\epsilon VaR_y(X) \frac{1}{n!} \left(\log \frac{\epsilon}{y} \right)^n dy \quad (\text{A.9})$$

and $AVaR_\epsilon^{(n)}(X)$ can be viewed as a spectral risk measure with a risk aversion function equal to

$$\phi_\epsilon^{(n)}(y) = \begin{cases} \frac{1}{\epsilon n!} \left(\log \frac{\epsilon}{y} \right)^n, & 0 \leq y \leq \epsilon \\ 0, & \epsilon < y \leq 1 \end{cases}$$

As a simple consequence of the definition, the sequence of higher-order AVaRs is monotonic,

$$AVaR_\epsilon(X) \leq AVaR_\epsilon^{(1)}(X) \leq \dots \leq AVaR_\epsilon^{(n)}(X) \leq \dots$$

In the entry, we remarked that if the random variable X has a finite mean, $E|X| < \infty$, then AVaR is also finite. This is not true for spectral risk measures and the higher-order AVaR in particular. In line with the general theory developed later in this appendix, $AVaR_\epsilon^{(n)}(X)$ is finite if all moments of X exist. For example, if

the random variable X has an exponential tail, then $AVaR_\epsilon^{(n)}(X) < \infty$ for any $n < \infty$.

The Minimization Formula for AVaR

In this section, we provide a geometric interpretation of the minimization formula (2) for AVaR. We restate equation (2) in the following equivalent form,

$$AVaR_\epsilon(X) = \frac{1}{\epsilon} \min_{\theta \in \mathbb{R}} (\epsilon\theta + E(-X - \theta)_+) \quad (\text{A.10})$$

where $(x)_+ = \max(x, 0)$. Note the similarity between equation (A.10) and the definition of AVaR in (A.2). Instead of the integral of the quantile function in the definition of AVaR, a minimization formula appears in (A.10). We interpreted the integral of the inverse c.d.f. as the shaded area in Figure 2. Similarly, we will find the area corresponding to the objective function in the minimization formula and we will demonstrate that as θ changes, there is a minimal area that coincides with the area corresponding to the shaded area in Figure 2. Moreover, the minimal area is attained for $\theta = VaR_\epsilon(X)$ when the c.d.f. of X is continuous at $VaR_\epsilon(X)$. In fact, all illustrations in this section are based on the assumption that X has a continuous distribution function.

Consider first the expectation in equation (A.10). Assuming that X has a continuous c.d.f., we obtain an expression for the expectation involving the inverse c.d.f.,

$$\begin{aligned} E(-X - \theta)_+ &= \int_{\mathbb{R}} \max(-x - \theta, 0) dF_X(x) \\ &= \int_0^1 \max(-F_X^{-1}(t) - \theta, 0) dt \\ &= - \int_0^1 \min(F_X^{-1}(t) + \theta, 0) dt \end{aligned}$$

This representation implies that the expectation $E(-X - \theta)_+$ equals the area closed between the graph of the inverse c.d.f. and a line parallel to the horizontal axis passing through the point $(0, -\theta)$. This is the shaded area on the right plot in Figure A.1. The same area can be represented

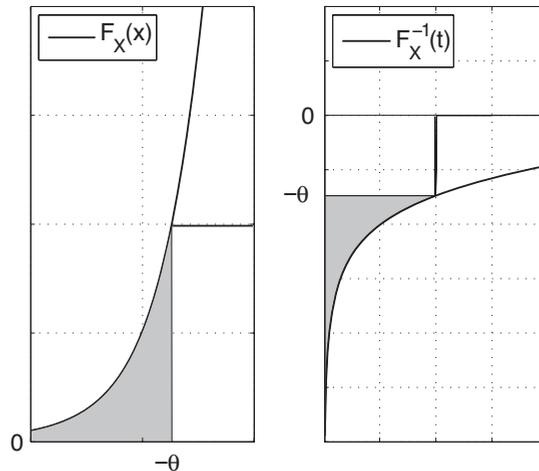


Figure A.1 Note: The shaded area is equal to the expectation $E(-X - \theta)_+$ in which X has a continuous distribution function.

in terms of the c.d.f. This is done on the left plot in Figure A.1.

Let us get back to equation (A.10). The tail probability ϵ is fixed. The product $\epsilon \times \theta$ equals the area of a rectangle with sides equal to ϵ and θ . This area is added to $E(-X - \theta)_+$. Figure A.2 shows the two areas together. The shaded areas on the top and the bottom plots equal $\epsilon \times AVaR_\epsilon(X)$. The top plot shows the case in which $-\theta < -VaR_\epsilon(X)$. Comparing the plot to Figure A.1, we find out that adding the marked area to the shaded area we obtain the total area corresponding to the objective in the minimization formula, $\epsilon\theta + E(-X - \theta)_+$. If $-\theta > -VaR_\epsilon(X)$, then we obtain a similar case shown on the bottom plot. Again, adding the marked area to the shaded area we obtain the the total area computed by the objective in the minimization formula. By varying θ , the total area changes but it always remains larger than the shaded area unless $\theta = VaR_\epsilon(X)$.

Thus, when $\theta = VaR_\epsilon(X)$ the minimum area is attained, which equals exactly $\epsilon \times AVaR_\epsilon(X)$. According to equation (A.10), we have to divide the minimal area by ϵ in order to obtain the AVaR. As a result, we have demonstrated that the minimization formula in equation (2) calculates the AVaR.

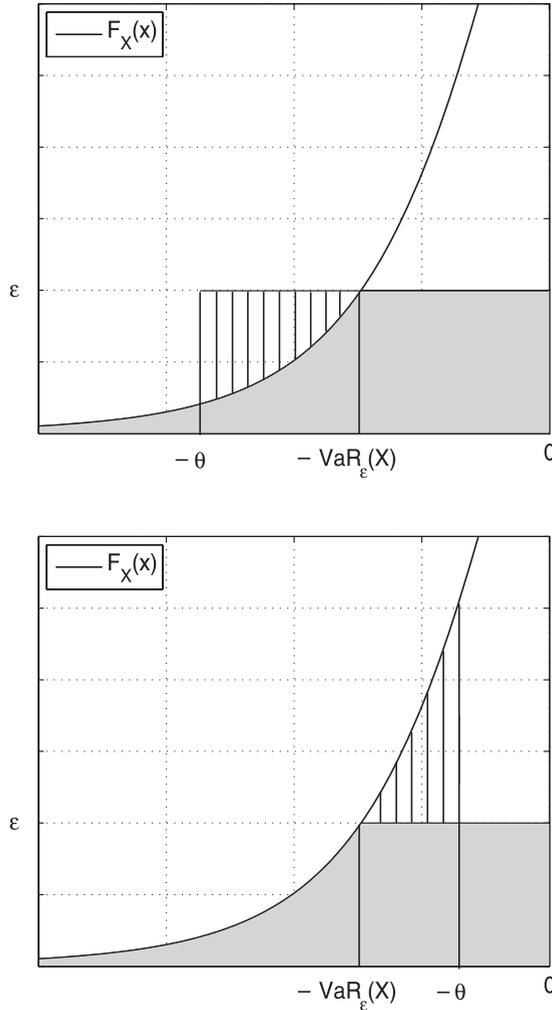


Figure A.2 Note: The marked area is in addition to the shaded one. The marked area is equal to zero if $\theta = \text{VaR}_\epsilon(X)$

AVaR for Stable Distributions

Working with the class of stable distributions in practice is difficult because there are no closed-form expressions for their densities and distribution functions. Thus, practical work relies on numerical methods.

Stoyanov et al. (2006) give an account of the approaches to estimating AVaR of stable distributions. It turns out that there is a formula that is not exactly a closed-form expression, such as the ones for the normal and Student's t AVaR stated in the entry, but is suit-

able for numerical work. It involves numerical integration but the integrand is nicely behaved and the integration range is a bounded interval. Numerical integration can be performed by standard toolboxes in many software packages, such as MATLAB, for example. Moreover, there are libraries freely available on the Internet. Therefore, numerical integration itself is not a severe restriction for applying a formula in practice. Since the formula involves numerical integration, we call it a semi-analytic expression.

Suppose that the random variable X has a stable distribution with tail exponent α , skewness parameter β , scale parameter σ , and location parameter μ , $X \in S_\alpha(\sigma, \beta, \mu)$. If $\alpha \leq 1$, then $\text{AVaR}_\epsilon(X) = \infty$. The reason is that stable distributions with $\alpha \leq 1$ have infinite mathematical expectation and the AVaR is unbounded.

If $\alpha > 1$ and $\text{VaR}_\epsilon(X) \neq 0$, then the AVaR can be represented as

$$\text{AVaR}_\epsilon(X) = \sigma A_{\epsilon, \alpha, \beta} - \mu$$

where the term $A_{\epsilon, \alpha, \beta}$ does not depend on the scale and the location parameters. In fact, this representation is a consequence of the positive homogeneity and the invariance property of AVaR. Concerning the term $A_{\epsilon, \alpha, \beta}$,

$$A_{\epsilon, \alpha, \beta} = \frac{\alpha}{1 - \alpha} \frac{|\text{VaR}_\epsilon(X)|}{\pi \epsilon} \int_{-\bar{\theta}_0}^{\pi/2} g(\theta) \exp(-|\text{VaR}_\epsilon(X)|^{\frac{\alpha}{\alpha-1}} v(\theta)) d\theta$$

where

$$g(\theta) = \frac{\sin(\alpha(\bar{\theta}_0 + \theta) - 2\theta)}{\sin \alpha(\bar{\theta}_0 + \theta)} - \frac{\alpha \cos^2 \theta}{\sin^2 \alpha(\bar{\theta}_0 + \theta)},$$

$$v(\theta) = (\cos \alpha \bar{\theta}_0)^{\frac{1}{\alpha-1}} \left(\frac{\cos \theta}{\sin \alpha(\bar{\theta}_0 + \theta)} \right)^{\frac{\alpha}{\alpha-1}} \frac{\cos(\alpha \bar{\theta}_0 + (\alpha - 1)\theta)}{\cos \theta},$$

in which $\bar{\theta}_0 = \frac{1}{\alpha} \arctan(\beta \tan \frac{\pi \alpha}{2})$, $\bar{\beta} = -\text{sign}(\text{VaR}_\epsilon(X))\beta$, and $\text{VaR}_\epsilon(X)$ is the VaR of the stable distribution at tail probability ϵ .

If $VaR_\epsilon(X) = 0$, then the AVaR admits a very simple expression,

$$AVaR_\epsilon(X) = \frac{2\Gamma\left(\frac{\alpha-1}{\alpha}\right) \cos \theta_0}{(\pi - 2\theta_0) (\cos \alpha\theta_0)^{1/\alpha}}$$

in which $\Gamma(x)$ is the gamma function and $\theta_0 = \frac{1}{\alpha} \arctan(\beta \tan \frac{\pi\alpha}{2})$.

ETL vs. AVaR

The expected tail loss and the average value-at-risk are two related concepts. In the entry, we remarked that ETL and AVaR coincide if the portfolio return distribution is continuous at the corresponding VaR level. However, if there is a discontinuity, or a point mass, then the two notions diverge. Still, the AVaR can be expressed through the ETL and the VaR at the same tail probability. In this section, we illustrate this relationship and show why the AVaR is more appealing. Moreover, it will throw light on why equation (6) should be used when considering a sample of observations.

The ETL at tail probability ϵ is defined as the average loss provided that the loss exceeds the VaR at tail probability ϵ ,

$$ETL_\epsilon(X) = -E(X|X < -VaR_\epsilon(X)) \quad (A.11)$$

As a consequence of the definition, the ETL can be expressed in terms of the c.d.f. and the inverse c.d.f. Suppose additionally, that the c.d.f. of X has a jump at $-VaR_\epsilon(X)$. In this case, the loss $VaR_\epsilon(X)$ occurs with probability equal to the size of the jump and, because of the strict inequality in (A.11), it will not be included in the average.

Figure A.3 shows the graphs of the c.d.f. and the inverse c.d.f. of a random variable with a point mass at $-VaR_\epsilon(X)$. If ϵ splits the jump of the c.d.f. as on the left plot in Figure A.3, then the ETL at tail probability ϵ equals

$$\begin{aligned} ETL_\epsilon(X) &= -E(X|X < -VaR_\epsilon(X)) \\ &= -E(X|X < -VaR_{\epsilon_0}(X)) \\ &= ETL_{\epsilon_0}(X) \end{aligned}$$

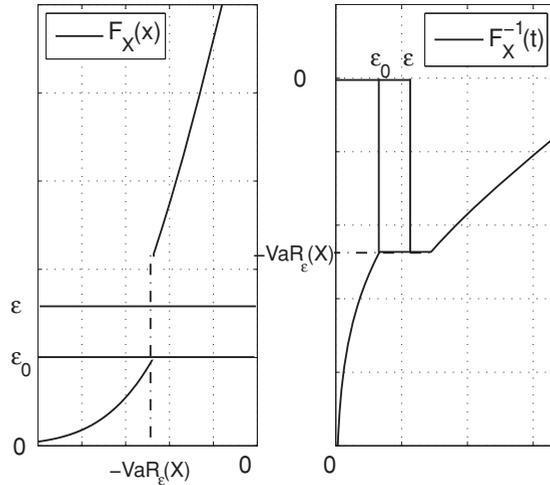


Figure A.3 The C.D.F. and the Inverse C.D.F. of a Random Variable X with a Point Mass at $-VaR_\epsilon(X)$
 Note: The tail probability ϵ splits the jump of the c.d.f.

In terms of the inverse c.d.f., the quantity $ETL_{\epsilon_0}(X)$ can be represented as

$$ETL_{\epsilon_0}(X) = -\frac{1}{\epsilon_0} \int_0^{\epsilon_0} F_X^{-1}(t) dt$$

The relationship between AVaR and ETL follows directly from the definition of AVaR.⁹ Suppose that the c.d.f. of the random variable X is as on the left plot in Figure A.3. Then,

$$\begin{aligned} AVaR_\epsilon(X) &= -\frac{1}{\epsilon} \int_0^\epsilon F_X^{-1}(t) dt \\ &= -\frac{1}{\epsilon} \left(\int_0^{\epsilon_0} F_X^{-1}(t) dt + \int_{\epsilon_0}^\epsilon F_X^{-1}(t) dt \right) \\ &= -\frac{1}{\epsilon} \int_0^{\epsilon_0} F_X^{-1}(t) dt + \frac{\epsilon - \epsilon_0}{\epsilon} VaR_\epsilon(X) \end{aligned}$$

where the last inequality holds because the inverse c.d.f. is flat in the interval $[\epsilon_0, \epsilon]$ and the integral is merely the surface of the rectangle shown on the right plot in Figure A.3. The integral in the first summand can be related to the ETL at tail probability ϵ and, finally, we arrive at the expression

$$AVaR_\epsilon(X) = \frac{\epsilon_0}{\epsilon} ETL_\epsilon(X) + \frac{\epsilon - \epsilon_0}{\epsilon} VaR_\epsilon(X) \quad (A.12)$$

Equation (A.12) shows that $AVaR_\epsilon(X)$ can be represented as a weighted average between the ETL and the VaR at the same tail probability as the coefficients in front of the two summands are positive and sum up to one. In the special case in which there is no jump, or if $\epsilon = \epsilon_1$, then AVaR equals ETL.

Why is equation (A.12) important if in all statistical models we assume that the random variables describing return or payoff distribution have densities? Under this assumption, not only are the corresponding c.d.f.s continuous but they are also smooth. Equation (A.12) is important because if the estimate of AVaR is based on the Monte Carlo method, then we use a sample of scenarios that approximate the nicely behaved hypothesized distribution. Even though we are approximating a smooth distribution function, the sample c.d.f. of the scenarios is completely discrete, with jumps at the scenarios the size of which equals the $1/n$, where n stands for the number of scenarios.

In fact, equation (6) given in the entry is actually equation (A.12) restated for a discrete random variable. The outcomes are the available scenarios, which are equally probable. Consider a sample of observations or scenarios r_1, \dots, r_n and denote by $r_{(1)} \leq r_{(2)} \leq \dots \leq r_{(n)}$ the ordered sample. The natural estimator of the ETL at tail probability ϵ is

$$\widehat{ETL}_\epsilon(r) = -\frac{1}{\lceil n\epsilon \rceil - 1} \sum_{k=1}^{\lceil n\epsilon \rceil - 1} r_{(k)} \quad (\text{A.13})$$

where $\lceil x \rceil$ is the smallest integer larger than x . Formula (A.13) means that we average $\lceil n\epsilon \rceil - 1$ of the $\lceil n\epsilon \rceil$ smallest observations, which is, in fact, the definition of the conditional expectation in (A.11) for a discrete distribution. The VaR at tail probability ϵ is equal to the negative of the empirical quantile,

$$\widehat{VaR}_\epsilon(r) = -r_{(\lceil n\epsilon \rceil)} \quad (\text{A.14})$$

It remains to determine the coefficients in (A.12). Having in mind that the observations in the sample are equally probable, we calculate

that

$$\epsilon_0 = \frac{\lceil n\epsilon \rceil - 1}{n}$$

Plugging ϵ_0 , (A.14), and (A.13) into equation (A.12), we obtain (6), which is the sample AVaR.

Similarly, equation (10) also arises from (A.12). The assumption is that the underlying random variable has a discrete distribution but the outcomes are not equally probable. Thus, the corresponding equation for the average loss on condition that the loss is larger than the VaR at tail probability ϵ is given by

$$\widehat{ETL}_\epsilon(r) = -\frac{1}{\epsilon_0} \sum_{j=1}^{k_\epsilon} p_j r_{(j)} \quad (\text{A.15})$$

where $\epsilon_0 = \sum_{j=1}^{k_\epsilon} p_j$ and k_ϵ is the integer satisfying the inequalities

$$\sum_{j=1}^{k_\epsilon} p_j \leq \epsilon < \sum_{j=1}^{k_\epsilon+1} p_j$$

The sum $\sum_{j=1}^{k_\epsilon} p_j$ stands for the cumulative probability of the losses larger than the the VaR at tail probability ϵ . Note that equation (A.15) turns into equation (A.13) when the outcomes are equally probable. With these remarks, we have demonstrated the connection between equations (6), (10), and (A.12).

The differences between ETL and AVaR are not without any practical importance. In fact, ETL is not a coherent risk measure. Furthermore, the sample ETL in (A.13) is not a smooth function of the tail probability while the sample AVaR is smooth. This is illustrated in Figure A.4. The top plot shows the graph of the sample ETL and AVaR with the tail probability varying between 1% and 10%. The sample contains 100 independent observations on a standard normal distribution, $X \in N(0,1)$. The bottom plot shows the same but the sample is larger. It contains 250 independent observations on a standard normal distribution.

Both plots demonstrate that the sample ETL is a step function of the tail probability while the AVaR is a smooth function of it. This is not

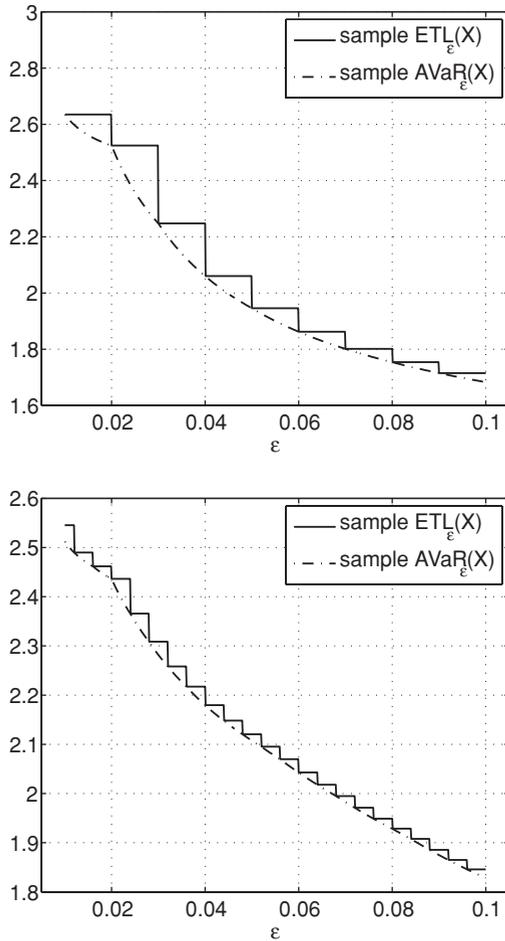


Figure A.4 The Graphs of the Sample ETL and AVaR with Tail Probability Varying between 1% and 10%

Note: The top plot is produced from a sample of 100 observations and the bottom plot from a sample of 250 observations. In both cases, $X \in N(0, 1)$.

surprising because, as ϵ increases, new observations appear in the sum in (A.13) producing the jumps in the graph of the sample ETL. In contrast, the AVaR changes gradually as it is a weighted average of the ETL and the VaR at the same tail probability. Note that, as the sample size increases, the jumps in the graph of the sample ETL diminish. In a sample of 5,000 scenarios, both quantities almost overlap. This is because the standard normal distribu-

tion has a smooth c.d.f. and the sample c.d.f. constructed from a larger sample better approximates the theoretical c.d.f. In this case, as the sample size approaches infinity, the AVaR becomes indistinguishable from the ETL at the same tail probability.¹⁰

KEY POINTS

- Although the value-at-risk (VaR) measure is a popular risk measure in the financial industry, it has a number of deficiencies. It is not a coherent risk measure because it does not satisfy the subadditivity property requirement of a coherent risk measure.
- In contrast to VaR, the average value-at-risk measure (AVaR)—also referred to as conditional value-at-risk and expected shortfall—is a coherent risk measure and has other advantages that results in its greater acceptance in risk modeling.
- There are convenient ways for computing and estimating AVaR that allow its application in optimal portfolio problems.
- A more general family of coherent risk measures is the spectral risk measure. The AVaR is a spectral risk measure with a specific risk-aversion function and is important for the proper selection of the risk-aversion function to avoid explosion of the risk measure.
- There is connection between the theory of probability metrics and risk measures. Basically, by choosing an appropriate probability metric one can guarantee that if two portfolio return distributions are close to each other, their risk profiles are also similar.

NOTES

1. In fact, $X = 0.05\sqrt{3}Z + 0.03$ where Z has Student's t distribution with 4 degrees of freedom and Y has a normal distribution with standard deviation equal to 0.1 and mathematical expectation equal to 0.01. The coefficient of Z is chosen so that the standard deviation of X is also equal to 0.1.

2. By comparing the c.d.f.s, we notice that the c.d.f. of X is “above” the c.d.f. of Y to the left of the crossing point, $F_X(x) \geq F_Y(x)$, $x \leq -0.15$.
3. This term is adopted in Rockafellar and Uryasev (2002).
4. Equation (2) was first studied by Pflug (2000). A proof that equation (1) is indeed the AVaR can be found in Rockafellar and Uryasev (2002).
5. As we remarked, $AVaR_\epsilon(X)$ can be infinite only if the mathematical expectation of X is infinite. Nevertheless, if this turns out to be an issue, one can use instead of AVaR the median of the loss distribution provided that the loss is larger than $VaR_\epsilon(X)$ as a robust version of AVaR. The median of the conditional loss is always finite and, therefore, the issue disappears but at the cost of violating the coherence axioms. The appendix to this entry provides more details.
6. This formula is a simple consequence of the definition of AVaR for discrete distributions; see the appendix to this entry. A detailed derivation is provided by Rockafellar and Uryasev (2002).
7. For example, $\lceil 3.1 \rceil = \lceil 3.8 \rceil = 4$.
8. A formal proof can be found in Rockafellar and Uryasev (2002). The reasoning in Rockafellar and Uryasev (2002) is based on the assumption that the random variable describes losses while in equation (10), the random variable describes the portfolio return or payoff.
9. Formal derivation of this relationship can be found, for example, in Rockafellar and Uryasev (2002).
10. In fact, this is a consequence of the celebrated Glivenko-Cantelli theorem claiming that the sample c.d.f. converges almost surely to the true c.d.f.

REFERENCES

- Acerbi, C. (2004). Coherent representation of subjective risk aversion. Chapter 10 in G. Szego (ed.), *Risk Measures for the 21st Century*. Chichester, UK: Wiley, pp. 147–208.
- Acerbi, C., and P. Simonetti (2002). Portfolio optimization with spectral measures of risk. Working paper. Abaxbank, Milan.
- Pflug, G. (2000). Some remarks on the value-at-risk and the conditional value-at-risk. In S. Uryasev (Ed.), *Probabilistic Constrained Optimization: Methodology and Applications*. Dordrecht: Kluwer Academic Publishers.
- Rockafellar, R. T., and S. Uryasev (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance* 26, 7: 1443–1471.
- Rudin, W. (1970). *Real and Complex Analysis*. New York: McGraw-Hill.
- Stoyanov, S. (2005). Optimal financial portfolios in highly volatile markets. PhD thesis, University of Karlsruhe.
- Stoyanov, S., Samorodnitsky, G., Rachev, S., and Ortobelli, S. (2006). Computing the portfolio conditional value-at-risk in the α -stable case. *Probability and Mathematical Statistics* 26, 1–22.
- Stoyanov, S., Rachev, S., and Fabozzi, F. (2008). Probability metrics with application in finance. *Journal of Statistical Theory and Practice*. 2, 2: 253–277.

Risk Measures and Portfolio Selection

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair-Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

CHRISTIAN MENN, Dr Rer Pol

Managing Partner, RIVACON

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: The standard assumption in financial models is that the distribution for the return on financial assets follows a normal (or Gaussian) distribution and therefore the standard deviation (or variance) is an appropriate measure of risk in the portfolio selection process. This is the risk measure that is used in the well-known Markowitz portfolio selection model (that is, mean-variance model) which is the foundation for modern portfolio theory. With mounting evidence since the early 1960s that return distributions do not follow a normal distribution, researchers have proposed alternative risk measures for portfolio selection. These risk measures fall into two disjointed categories: dispersion measures and safety-first measures. In addition, there has been considerable theoretical work in defining the features of a desirable risk measure.

Most of the concepts in theoretical and empirical finance that have been developed over the last 50 years rest upon the assumption that the return or price distribution for financial assets follow a normal distribution. Yet, with rare exception, studies that have investigated the validity of this assumption since the 1960s fail to find support for the normal distribution. Moreover, there is ample empirical evidence that many—if not most—financial return series are heavy-tailed and, possibly, skewed. The “tails” of the distribution are where the extreme values

occur. Empirical distributions for stock prices and returns have found that the extreme values are more likely than would be predicted by the normal distribution. This means that between periods where the market exhibits relatively modest changes in prices and returns, there will be periods where there are changes that are much higher (that is, crashes and booms) than predicted by the normal distribution. This is not only of concern to financial theorists, but also to practitioners seeking, for example, to produce probability estimates for financial

risk assessment. To more effectively implement portfolio selection, alternative risk measures are needed.

In this entry, we review alternative risk measures that can be employed in portfolio selection, which can accommodate non-normal return distributions. These risk measures are classified as dispersion measures and safety-first measures. We begin with a discussion of the desirable features of investment risk measures.

DESIRABLE FEATURES OF INVESTMENT RISK MEASURES

In portfolio theory, the variance of a portfolio's return has been historically the most commonly used measure of investment risk. However, different investors adopt different investment strategies in seeking to realize their investment objectives. Consequently, intuitively, it is difficult to believe that investors have come to accept only one definition of risk. Regulators of financial institutions and commentators to risk measures proposed by regulators have proffered alternative definitions of risk. As noted by Dowd (2002, p. 1):

The theory and practice of risk management—and, included with that, risk measurement—have developed enormously since the pioneering work of Harry Markowitz in the 1950s. The theory has developed to the point where risk management/measurement is now regarded as a distinct sub-field of the theory of finance. . . .

Szegö (2004, p. 1) categorizes risk measures as one of the three major revolutions in finance and places the start of that revolution in 1997. The other two major revolutions are mean-variance analysis (1952–1956) and continuous-time models (1969–1973). He notes that alternative risk measures have been accepted by practitioners but “rejected by the academic establishment and, so far discarded by regulators!” (Szegö, 2004, p. 4).

Basic Features of Investment Risk Measures

Balzer (2001) argues that a risk measure is investor specific and, therefore, there is “no single universally acceptable risk measure.” He suggests several features that an investment risk measure should capture. Here we describe the following three features:

- Relativity of risk
- Multidimensionality of risk
- Asymmetry of risk

The *relativity of risk* means that risk should be related to performing worse than some alternative investment or benchmark. Balzer (1994, 2001) and Sortino and Satchell (2001), among others, have proposed that investment risk might be measured by the probability of the investment return falling below a specified risk benchmark. The risk benchmark might itself be a random variable, such as a liability benchmark (e.g., an insurance product), the inflation rate or possibly inflation plus some safety margin, the risk-free rate of return, the bottom percentile of return, a sector index return, a budgeted return, or other alternative investments. Each benchmark can be justified in relation to the goal of the portfolio manager. Should performance fall below the benchmark, there could be major adverse consequences for the portfolio manager.

In addition, the same investor could have multiple objectives and, hence, multiple risk benchmarks. Thus, risk is also a *multidimensional* phenomenon. However, an appropriate choice of the benchmarks is necessary in order to avoid an incorrect evaluation of opportunities available to investors. For example, too often, little recognition is given to liability targets. This is the major factor contributing to the underfunding of U.S. corporate pension sponsors of defined benefit plans.¹

Intuition suggests that risk is an *asymmetric* concept related to the downside outcomes, and any realistic risk measure has to value and

consider upside and downside differently. The standard deviation considers the positive and the negative deviations from the mean as a potential risk. In this case overperformance relative to the mean is penalized just as much as underperformance.

Intertemporal Dependence and Correlation with Other Sources of Risk

The standard deviation is a measure of dispersion and it cannot always be used as a measure of risk. The preferred investment does not always present better returns than the other. It could happen that the worst investment presents the greatest return in some periods. Hence, time could influence the investor's choices.

Clearly, if the degree of uncertainty changes over time, the risk has to change during the time as well. In this case, the investment return process is not stationary; that is, we cannot assume that returns maintain their distribution unvaried in the course of time. In much of the research published, stationary and independent realizations are assumed. The latter assumption implies that history has no impact on the future. More concrete, the distribution of tomorrow's return is the same independent of whether the biggest stock market crash ever recorded took place yesterday or yesterday's return equaled 10%.

As a result, the oldest observations have the same weight in our decisions as the most recent ones. Is this assumption realistic? Recent studies on investment return processes have shown that historical realizations are not independent and present a clustering of the volatility effect (time-varying volatility). Those phenomena lead to the fundamental time-series model autoregressive conditional heteroscedasticity (ARCH) formulated by Engle (1981). In particular, the last observations have a greater impact in investment decisions than the oldest ones. Thus, any realistic measure of

risk changes and evolves over time taking into consideration the heteroscedastic (time-varying volatility) behavior of historical series. An examination of the returns of say the equity return indexes such as the S&P 500 over some time period would show a *propagation effect* on another equity market, say the DAX 30. When we observe the highest peaks in one return index series, for example, there is an analogous peak in the other return index series. This propagation effect is known as cointegration of the return series, introduced by the fundamental work of Granger (1981) and elaborated upon further by Engle and Granger (1987). The propagation effect in this case is a consequence of the globalization of financial markets—the risk of a country/sector is linked to the risk of the other countries/sectors. Therefore, it could be important to limit the propagation effect by diversifying the risk. As a matter of fact, it is largely proven that the diversification, opportunistically modeled, diminishes the probability of big losses. Hence, an adequate risk measure values and models correctly the correlation among different investments, sectors, and markets.

ALTERNATIVE RISK MEASURES FOR PORTFOLIO SELECTION

The goal of portfolio selection is the construction of portfolios that maximize expected returns consistent with individually acceptable levels of risk. Using both historical data and investor expectations of future returns, portfolio selection uses modeling techniques to quantify “expected portfolio returns” and “acceptable levels of portfolio risk,” and provides methods to select an optimal portfolio.

It would not be an overstatement to say that modern portfolio theory as developed by Harry Markowitz (1952, 1959) has revolutionized the world of investment management. Allowing managers to appreciate that the investment risk and expected return of a portfolio can be

quantified has provided the scientific and objective complement to the subjective art of investment management. More importantly, whereas previously the focus of portfolio management used to be the risk of individual assets, the theory of portfolio selection has shifted the focus to the risk of the entire portfolio. This theory shows that it is possible to combine risky assets and produce a portfolio whose expected return reflects its components, but with considerably lower risk. In other words, it is possible to construct a portfolio whose risk is smaller than the sum of all its individual parts.

Though practitioners realized that the risks of individual assets were related, prior to modern portfolio theory they were unable to formalize how combining them into a portfolio impacted the risk at the entire portfolio level or how the addition of a new asset would change the return/risk characteristics of the portfolio. This is because practitioners were unable to quantify the returns and risks of their investments. Furthermore, in the context of the entire portfolio, they were also unable to formalize the interaction of the returns and risks across asset classes and individual assets. The failure to quantify these important measures and formalize these important relationships made the goal of constructing an optimal portfolio highly subjective and provided no insight into the return investors could expect and the risk they were undertaking. The other drawback, before the advent of the theory of portfolio selection and asset pricing theory, was that there was no measurement tool available to investors for judging the performance of their investment managers.

The theory of portfolio selection set forth by Markowitz was based on the assumption that asset returns are normally distributed. As a result, Markowitz suggested that the appropriate risk measure is the variance of the portfolio's return and portfolio selection involved only two parameters of the asset return distribution: mean and variance. Hence, the approach to portfolio selection he proposed is popularly referred to as *mean-variance analysis*.

Markowitz recognized that an alternative to the variance is the *semivariance*.² The semivariance is similar to the variance except that, in the calculation, no consideration is given to returns above the expected return. Portfolio selection could be recast in terms of mean-semivariance. However, if the return distribution is symmetric, Markowitz (1959, p. 190) notes that "an analysis based on (expected return) and (standard deviation) would consider these . . . (assets) as equally desirable." He rejected the semivariance noting that the variance "is superior with respect to cost, convenience, and familiarity" and when the asset return distribution is symmetric, either measure "will produce the same set of efficient portfolios." (Markowitz 1959, pp. 193–194).

There is a heated debate on risk measures used for valuing and optimizing the investor's risk portfolio. In this section and the one to follow, we describe the various portfolio risk measures proposed in the literature and more carefully look at the properties of portfolio risk measures.

According to the literature on portfolio theory, two disjointed categories of risk measures can be defined: *dispersion measures* and *safety-first risk measures*. In the remainder of this entry, we review some of the most well-known dispersion measures and safety-first measures along with their properties.³

In the following, we consider a portfolio of N assets whose individual returns are given by r_1, \dots, r_N . The relative weights of the portfolio are denoted as x_1, \dots, x_n and, therefore, the portfolio return r_p can be expressed as

$$r_p = x_1 \cdot r_1 + \dots + x_N \cdot r_N = \sum_{i=1}^N x_i \cdot r_i$$

We also provide a sample version of the discussed risk measures. The sample version will be based on a sample of length T of independent and identically distributed observations $r_p^{(k)}$, $k = 1, \dots, T$ of the portfolio return r_p . These

observations can be obtained from a corresponding sample of the individual assets.

DISPERSION MEASURES

Several portfolio mean dispersion approaches have been proposed in the last few decades. The most significant ones are discussed below, and we provide for each measure an example to illustrate the calculation.

Mean Standard Deviation

In the *mean standard deviation* approach the dispersion measure is the standard deviation of the portfolio return r_p (see Markowitz, 1959, and Tobin, 1958):

$$\sigma(r_p) = \sqrt{E(r_p - E(r_p))^2} \quad (1)$$

The standard deviation is a special case of the mean absolute moment discussed below. The sample version can be obtained from the general case by setting $p = 2$.

Mean Absolute Deviation

In the *mean absolute deviation (MAD)* approach, the dispersion measure is based on the absolute deviations from the mean rather than the squared deviations as in the case of the standard deviation.⁴ The MAD is more robust with respect to outliers. The MAD for the portfolio return r_p is defined as

$$\text{MAD}(r_p) = E(|r_p - E(r_p)|) \quad (2)$$

Mean Absolute Moment

The *mean absolute moment (MAM)(q)* approach is the logical generalization of the MQ approach. Under this approach the dispersion measure is defined as

$$\text{MAM}(r_q, p) = (E(|r_p - E(r_p)|^q))^{1/q}, \quad q \geq 1 \quad (3)$$

Note that the mean absolute moment for $q = 2$ coincides with the standard deviation and for

$q = 1$ the mean absolute moment reduces to the mean absolute deviation. One possible sample version of (3) is given by

$$\text{MAM}(r_p, q) = \sqrt[q]{\frac{1}{T} \sum_{k=1}^T |r_p^{(k)} - \bar{r}_p|^q}$$

where

$$\bar{r}_p = \frac{1}{T} \sum_{k=1}^T r_p^{(k)}$$

denotes the sample mean of the portfolio return.

Gini Index of Dissimilarity

The *index of dissimilarity* is based on the measure introduced by Gini (1912, 1921).⁵ Gini objected to the use of the variance or the MAD because they measure deviations of individuals from the individual observations of the mean or location of a distribution. Consequently, these measures linked location with variability, two properties that Gini argued were distinct and do not depend on each other. He then proposed the pairwise deviations between all observations as a measure of dispersion, which is now referred to as the *Gini measure*.

While this measure has been used for the past 80 years as a measure of social and economic conditions, its interest as a measure of risk in the theory of portfolio selection is relatively recent. Interest in a Gini-type risk measure has been fostered by Rachev (1991) and Rachev and Gamrowski (1995). Mathematically, the Gini risk measure for the random portfolio return r_p is defined as

$$\text{GM}(r_p, r_b) = \text{Min}\{E|r_p - r_b|\} \quad (4)$$

where the minimum is taken over all joint distributions of (r_p, r_b) with fixed marginal distribution functions F and G :

$$\begin{aligned} F(x) &= P(r_p \leq x) \text{ and} \\ G(x) &= P(r_b \leq x), \quad x \text{ real} \end{aligned}$$

Here r_b is the benchmark return, say, the return of a market index, or just the risk-free rate (U.S. Treasury rate or LIBOR, for example).

Expression (4) can be represented as the mean absolute deviation between the two distribution functions F and G :

$$GM(r_p, r_b) = \int_{-\infty}^{+\infty} |F(x) - G(x)| dx$$

Given a sample or a distributional assumption for the benchmark return r_b , the latter expression can be used for estimating the Gini index by calculating the area between the graphs of the empirical distribution function of r_p and the (empirical) distribution function of r_b .

Mean Entropy

In the *mean entropy* (M-entropy) approach, the dispersion measure is the exponential entropy. Exponential entropy is a dispersion measure only for portfolios with continuous return distribution because the definition of entropy for discrete random variables is formally different and does not satisfy the properties of the dispersion measures (positive and positively homogeneous). The concept of entropy was introduced in the last century in the classical theory of thermodynamics. Roughly speaking, it represents the average uncertainty in a random variable.

Probably its most important application in finance is to derive the probability density function of the asset underlying an option on the basis of the information that some option prices provide.⁶ Entropy was used also in portfolio theory by Philippatos and Wilson (1972) and Philippatos and Gressis (1975) and is defined as

$$\text{Entropy} = -E(\log f(r_p))$$

where f is the density of the portfolio return. Thus, the exponential entropy is given by

$$EE(r_p) = e^{-E(\log f(r_p))} \quad (5)$$

The valuation of entropy can be done either by considering the empirical density of a portfolio or assuming that portfolio returns belong to a given family of continuous distributions and estimate their unknown parameters.

Mean Colog

In the *mean colog* (M-colog) approach, the dispersion measure is the covariance between the random variable and its logarithm.⁷ That is, the colog of a portfolio return is defined as

$$\begin{aligned} \text{Colog}(1 + r_p) &= E(r_p \log(1 + r_p)) \\ &\quad - E(r_p)E(\log(1 + r_p)) \quad (6) \end{aligned}$$

Colog can easily be estimated based on a sample of the portfolio return distribution by:

$$\begin{aligned} \text{Colog}(1 + r_p) &\approx \frac{1}{T} \sum_{k=1}^T (r_p^{(k)} - \bar{r}_p) \cdot (\log(1 + r_p^{(k)})) \\ &\quad - \overline{\log(1 + r_p)} \end{aligned}$$

where

$$\overline{\log(1 + r_p)} = \frac{1}{T} \sum_{k=1}^T \log(1 + r_p^{(k)})$$

denotes the sample mean of the logarithm of one plus the portfolio return.

SAFETY-FIRST RISK MEASURES

Many researchers have suggested the safety-first rules as a criterion for decision making under uncertainty.⁸ In these models, a subsistence, a benchmark, or a disaster level of returns is identified. The objective is the maximization of the probability that the returns are above the benchmark. Thus, most of the *safety-first risk measures* proposed in the literature are linked to the benchmark-based approach.

Even if there are not apparent connections between the expected utility approach and a more appealing benchmark-based approach, Castagnoli and LiCalzi (1996) have proven that the expected utility can be reinterpreted in terms of the probability that the return is above a given benchmark. Hence, when it is assumed that investors maximize their expected utility, it is implicitly assumed that investors minimize

the probability of the investment return falling below a specified risk benchmark.

Although it is not always simple to identify the underlying benchmark, expected utility theory partially justifies the using of the benchmark-based approach. Moreover, it is possible to prove that the two approaches are in many cases equivalent even if the economic reasons and justifications are different.⁹

Some of the most well-known safety-first risk measures proposed in the literature are described in the next section.

Classical Safety First

In the *classical safety-first* portfolio choice problem the risk measure is the probability of loss or, more generally, the probability $P_\lambda = P(r_p \leq \lambda)$ of portfolio return less than λ . Generally, safety-first investors have to solve a complex, mixed integer linear programming problem to find the optimal portfolios. However, when short sales are allowed and return distributions are elliptical, depending on a dispersion matrix Q and a vector mean μ , then there exists a closed-form solution to the investor's portfolio selection problem:

$$\begin{aligned} & \text{Minimize: } P(r_p \leq \lambda) \\ & \text{Subject to: } \sum_{i=1}^N x_i = 1, \quad x_i \geq 0 \end{aligned}$$

The interesting property of this optimization problem is that we are able to express the set of optimal portfolios explicitly as a function of the shortfall barrier λ , the mean vector μ , and the dispersion matrix Q . The mean m and the dispersion σ^2 of these optimal portfolios can again be expressed as a function of the threshold λ , the mean vector μ , and the dispersion matrix Q . In the case where the elliptical family has finite variance (as, for example, the normal distribution), then the dispersion σ^2 corresponds to the variance.

As the risk measure consists of the probability that the return falls below a given barrier λ , we can estimate the risk measure by the ratio between the number of observations being

smaller than λ and the total number of observations in the sample.

Value at Risk

Value at risk ($\text{VaR}_{1-\alpha}$) is a closely related possible safety-first measure of risk defined by the following equality:

$$\text{VaR}_{1-\alpha}(r_p) = -\min\{z | (P(r_p \leq z) > \alpha)\} \quad (7)$$

Here, $1 - \alpha$ is denoted as the confidence level and α usually takes values like 1% or 5%. Theoretically, the VaR figure defined by equation (7) can admit negative values. In reality, however, it is likely and often implicitly assumed that the VaR is positive, and it can be interpreted as the level at which the losses will not exceed with a probability of $1 - \alpha\%$. Sometimes VaR is, therefore, defined as the maximum of zero and the expression defined in equation (7) to guarantee a positive value for VaR.

VaR can be used as a risk measure to determine reward-risk optimal portfolios. Moreover, this simple risk measure can also be used by financial institutions to evaluate the market risk exposure of their trading portfolios. The main characteristic of VaR is that of synthesizing in a single value the possible losses that could occur with a given probability in a given temporal horizon. This feature, together with the (very intuitive) concept of maximum probable loss, allows the nonexpert investor to figure out how risky the position is and the correcting strategies to adopt. Based on a sample of return observations, VaR estimates coincide with the empirical alphaquantile. VaR and sophisticated methodologies for estimating VaR are explained in Chapter 14 of Rachev, Menn, and Fabozzi (2005).

Conditional Value at Risk/Expected Tail Loss

The *conditional value at risk* ($\text{CVaR}_{1-\alpha}$) or *expected tail loss* (ETL) is defined as:

$$\begin{aligned} \text{CVaR}_{1-\alpha}(r_p) &= E(\max(-r_p, 0) | -r_p) \\ &\geq \text{VaR}_{1-\alpha}(r_p) \end{aligned} \quad (8)$$

where $\text{VaR}_{1-\alpha}(X)$ is defined in equation (7) and we assume that portfolio return distribution is continuous.¹⁰ From this definition we observe that the CVaR can be seen as the expected shortfall assuming the $\text{VaR}_{1-\alpha}(X)$ as the benchmark.

A sophisticated estimation of CVaR depends strongly on the estimation of VaR. An explanation and illustration of the calculation of CVaR is provided in Rachev, Menn, and Fabozzi (2005). Based on a large sample of observations, a natural estimate for CVaR can be obtained by averaging all observations in the sample that are smaller than the corresponding VaR estimate.

MiniMax

An alternative way to derive some safety-first optimal portfolios is minimizing the *MiniMax* (MM) risk measure (see Young, 1998). The MiniMax of a portfolio return is given by:

$$MM(r_p) = -\sup\{c | P(r_p \leq c) = 0\} \quad (9)$$

This risk measure can be seen as an extreme case of CVaR.

Lower Partial Moment

A natural extension of semivariance is the *lower partial moment* risk measure (see Bawa, 1976, and Fishburn, 1977) also called *downside risk* or *probability-weighted function of deviations below a specified target return*. This risk measure depends on two parameters:

1. A power index that is a proxy for the investor's degree of risk aversion.
2. The target rate of return that is the minimum return that must be earned to accomplish the goal of funding the plan within a cost constraint.

The lower partial moment of a portfolio r_p bounded from below is given by

$$\text{LPM}(r_p, q) = \sqrt[q]{E(\max(t - r_p, 0)^q)} \quad (10)$$

where q is the power index and t is the target rare of return.

Given a sample of return observations, we can approximate equation (10) as follows:

$$\text{LPM}(r_p, q) = \sqrt[q]{\frac{1}{T} \sum_{k=1}^T \max(r_p^{(k)} - \bar{r}_p, 0)^q}$$

where as before

$$\bar{r} = \frac{1}{T} \sum_{k=1}^T r_p^{(k)}$$

denotes the sample mean of the portfolio return.

Power Conditional Value at Risk

The *power conditional value at risk* measure, introduced in Rachev, Jasic, Biglova, and Fabozzi (2005), is the CVaR of the lower partial moment of the return. It depends on a power index that varies with respect to an investor's degree of risk aversion. Power CVaR generalizes the concept of CVaR and is defined as

$$\begin{aligned} \text{CVaR}_{q,1-\alpha}(r_p) &= E(\max(-r_p, 0)^q | -r_p \\ &\geq \text{VaR}_{1-\alpha}(r_p)) \end{aligned} \quad (11)$$

A sample version of power CVaR can be obtained in the same way as sample version for the regular CVaR, that is, one calculates the q -th sample moment of all observations in the sample that are smaller than the corresponding VaR estimate.

KEY POINTS

- While the underpinning of financial theory is that the distribution of the return on financial assets is normally distributed, little evidence supports this assumption. Consequently, the justification for the use of the standard deviation or variance as a measure of risk in financial applications such as portfolio selection is difficult to justify.
- Alternative risk measures that can accommodate the properties of asset returns that have been observed in financial markets have been proposed.

- Alternative risk measures include dispersion measures and safety-first risk measures.
- Dispersion measures include mean standard deviation, mean absolute deviation, mean absolute moment, index of dissimilarity, mean entropy, and mean colog.
- Safety-first risk measures include classical safety first, value at risk, conditional value at risk, expected tail loss, MiniMax, lower partial moment, downside risk, probability-weighted function of deviations below a specified target return, and power conditional value at risk.

NOTES

1. See Ryan and Fabozzi (2002).
2. The mean semivariance approach was revisited by Stefani and Szegö (1976).
3. For more details, see Rachev, Menn, and Fabozzi (2006).
4. See Konno and Yamazaki (1991), Zenios and Kang (1993), Speranza (1993), and Ogryczak and Ruszczyński (2001).
5. For a further discussion of this index, see Rachev (1991).
6. See Buchen and Kelly (1996) and Avellaneda (1998).
7. See Giacometti and Ortobelli (2001).
8. See, among others, Roy (1952), Tesler (1955/6), and Bawa (1976, 1978).
9. See Castagnoli and LiCalzi (1996, 1999), Bordley and LiCalzi (2000), Ortobelli and Rachev (2001), Rachev and Mittnik (2000, pp. 424–464), and Rachev, Ortobelli, and Schwartz (2004).
10. See Bawa (1978), Uryasev (2000), and Martin, Rachev, and Siboulet (2003).

REFERENCES

- Artzner, P., Delbaen, F., Eber, J-M, and Heath, D. (2000). Coherent measures of risk. *Mathematical Finance* 9: 203–228.
- Artzner, P., Delbaen, F., Eber, J-M, Heath, D., and Ku, H. (2003). Coherent multiperiod measures of risk. Unpublished paper.
- Avellaneda, M. (1998). Minimum entropy calibration of asset pricing models. *International Journal of Theoretical and Applied Finance* 1: 447–472.
- Balzer, L. A. (2001). Investment Risk: A unified approach to upside and downside returns. In F. A. Sortino and S. E. Satchell (eds.), *Managing Downside Risk in Financial Markets: Theory Practice and Implementation* (pp. 103–155). Oxford: Butterworth-Heinemann.
- Bawa, V. S. (1976). Admissible portfolio for all individuals. *Journal of Finance* 31: 1169–1183.
- Bawa, V. S. (1978). Safety-first stochastic dominance and optimal portfolio choice. *Journal of Financial and Quantitative Analysis* 13: 255–271.
- Bordley, R., and LiCalzi, M. (2000). Decision analysis using targets instead of utility functions. *Decision in Economics and Finance* 23: 53–74.
- Buchen, P. W., and Kelly, M. (1996). The maximum entropy distribution of an asset inferred from option prices. *Journal of Financial and Quantitative Analysis* 31: 143–159.
- Castagnoli, E., and LiCalzi, M. (1996). Expected utility without utility. *Theory and Decision* 41: 281–301.
- Castagnoli, E., and LiCalzi, M. (1999). Non-expected utility theories and benchmarking under risk. *SZIGMA* 29: 199–211.
- Dowd, K. (2002). *Measuring Market Risk*. Chichester: John Wiley & Sons.
- Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica* 50: 987–1008.
- Engle, R. F., and Granger, C. W. J. (1987). Cointegration and error correction: Representation, estimation, and testing. *Econometrica* 55: 251–276.
- Fishburn, P. C. (1977). Mean-risk analysis with risk associated with below-target returns. *American Economic Review* 67: 116–126.
- Frittelli, M., and Rosazza Gianin, E. (2004). Dynamic convex risk measures. In G. Szegö (ed.), *Risk Measures for the 21st Century* (pp. 227–249). Chichester, UK: John Wiley & Sons.
- Giacometti, R., and Ortobelli, S. (2004). Risk measures for asset allocation models. In G. Szegö (ed.), *Risk Measures for the 21st Century* (pp. 69–87). Chichester, UK: John Wiley & Sons.
- Gini, C. (1921). Measurement of inequality of incomes. *The Economic Journal* 31: 124–126.
- Gini, C. (1965). La dissomiglianza. *Metron* 24: 309–331.
- Granger, C. W. J. (1981). Some properties of time series and their use in econometric model specification. *Journal of Econometrics* 16: 121–130.

- Konno, H., and Yamazaki, H. (1991). Mean-absolute deviation portfolio optimization model and its application to Tokyo Stock Market. *Management Science* 37: 519–531.
- Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance* 7: 77–91.
- Markowitz, H. M. (1959). *Portfolio Selection: Efficient Diversification of Investment*. New York: John Wiley & Sons.
- Martin, D., Rachev, S. T., and Siboulet, F. (2003). Phi-Alpha optimal portfolios and extreme risk management. *Willmot Magazine of Finance*, November: 70–83.
- Ogryczak, W., and Ruszczyński, A. (2001). On consistency of stochastic dominance and mean-semideviation models. *Mathematical Programming* 89: 217–232.
- Olsen, R. A. (1997). Investment risk: The experts' perspective. *Financial Analysts Journal*, March/April: 62–66.
- Ortobelli, S. (2001). The classification of parametric choices under uncertainty: Analysis of the portfolio choice problem. *Theory and Decision* 51: 297–327.
- Ortobelli, S., and Rachev, S. T. (2001). Safety first analysis and stable Paretian approach. *Mathematical and Computer Modelling* 34: 1037–1072.
- Ortobelli, S., Huber, I., and Schwartz, E. (2002). Portfolio selection with stable distributed returns. *Mathematical Methods of Operations Research* 55: 265–300.
- Ortobelli, S., Huber, I., Rachev, S. T., and Schwartz, E. (2003). Portfolio choice theory with non-Gaussian distributed returns. In S.T. Rachev (ed.), *Handbook of Heavy Tailed Distributions in Finance* (pp. 547–594). Amsterdam: North Holland Handbooks of Finance.
- Philippatos, G. C., and Wilson, C. J. (1972). Entropy, market risk and the selection of efficient portfolio. *Applied Economics* 4: 209–220.
- Philippatos, G. C., and Gressis, N. (1975). Conditions of equivalence among E-V, SSD and E-H portfolio selection criteria: The case for uniform, normal and lognormal distributions. *Management Science* 21: 617–625.
- Rachev, S. T. (1991). *Probability Metrics and the Stability of Stochastic Models*. New York: John Wiley & Sons.
- Rachev, S. T., and Gamrowski, S. (1995). Financial models using stable laws. In Y. V. Prohorov (ed.), *Probability Theory and Its Application in Applied and Industrial Mathematics*, Vol. 2 (pp. 556–604), New York: Springer Verlag.
- Rachev, S. T., Jasic, T., Biglova, A., and Fabozzi, F. J. (2005). Risk and return in momentum strategies: Profitability from portfolios based on risk-adjusted stock ranking criteria. Technical report, Chair of Econometrics, Statistics and Mathematical Finance, School of Economics, University of Karlsruhe, Postfach 6980, D-76128, Karlsruhe, Germany and Technical Report, Department of Statistics and Applied Probability, UCSB, CA 93106, USA.
- Rachev, S. T., Menn, C., and Fabozzi, F. J. (2005). *Fat-Tailed and Skewed Asset Return Distributions: Implications for Risk Management, Portfolio Selection, and Option Pricing*. Hoboken, NJ: John Wiley & Sons.
- Rachev, S. T., and Mittnik, S. (2000). *Stable Paretian Model in Finance*. Chichester: John Wiley & Sons.
- Rachev, S., Ortobelli, S., and Schwartz, S. (2004). The problem of optimal asset allocation with stable distributed returns. In A. Krinik and R. J. Swift (eds.), *Stochastic Processes and Functional Analysis: A Dekker Series of Lecture Notes in Pure and Applied Mathematics* (pp. 295–361), New York: Dekker.
- Roy, A. D. (1952). Safety-first and the holding of assets. *Econometrica* 20: 431–449.
- Ryan, R., and Fabozzi, F. J. (2002). Rethinking pension liabilities and asset allocation. *Journal of Portfolio Management* 28, 4: 7–15.
- Shalit, H., and Yitzhaki, S. (1984). Mean-Gini, portfolio theory, and the pricing of risky assets. *Journal of Finance* 39: 1449–1468.
- Sortino, F. A., and Satchell, S. E. (eds.) (2001). *Managing Downside Risk in Financial Markets: Theory Practice and Implementation*. Oxford: Butterworth-Heinemann.
- Speranza, M. G. (1993). Linear programming models for portfolio optimization. *Finance* 14: 107–123.
- Stefani, S., and Szegö, G. (1976). Formulazione analitica della funzione utilità dipendente da media e semivarianza mediante il principio dell'utilità attesa. *Bollettino UML*, 13A: 157–162.
- Stoyanov, S. V., Rachev, S. T., and Fabozzi, F. J. (2008). Optimal financial portfolios. Forthcoming in *Applied Mathematical Finance*.
- Szegö, G. (2004). On the (non)acceptance of innovations. In G. Szegö (ed.), *Risk Measures for the 21st Century* (pp. 1–10). Chichester, UK: John Wiley & Sons.

- Tesler, L. G. (1955/6). Safety first and hedging. *Review of Economic Studies* 23: 1–16.
- Tobin, J. (1958). Liquidity preference as behavior toward risk. *Review of Economic Studies* 25: 65–86.
- Uryasev, S. P. (2000). *Probabilistic Constrained Optimization Methodology and Applications*. Dordrecht: Kluwer Academic Publishers.
- Von Neumann, J., and Morgenstern, O. (1953). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.
- Young, M. R. (1998). A MiniMax portfolio selection rule with linear programming solution. *Management Science* 44: 673–683.
- Zenios, S. A., and Kang, P. (1993). Mean absolute deviation portfolio optimization for mortgage-backed securities. *Annals of Operations Research* 45: 433–450.

Back-Testing Market Risk Models

KEVIN DOWD, PhD

Partner, Cobden Partners, London

Abstract: Back-testing is the quantitative evaluation of a model, and back-testing a risk or probability density forecasting model involves a comparison of the model's density forecasts against subsequently realized outcomes of the random variable whose density is forecast. One purpose of back-testing is to determine whether the forecasts are sufficiently close to realized outcomes to enable us to conclude that the forecasts are statistically compatible with those outcomes. Back-tests conducted for this purpose involve statistical hypothesis tests to determine if a model's forecasts are acceptable. Hypothesis tests can be applied to observations involving a loss that exceeds the value-at-risk at a given confidence interval, or they can be applied to forecasts of VaRs at multiple confidence intervals. A second purpose of back-testing is to assist risk managers to diagnose problems with their risk models and so help improve them. A third purpose of back-testing is to rank the performance of a set of alternative risk models to determine which model gives the "best" density forecast evaluation performance.

To back-test a model is to evaluate it in quantitative terms, and *back-testing* a risk (or probability density forecasting) model involves a comparison of the model's density forecasts against subsequently realized outcomes of the underlying random variable whose density is forecast. The importance of back-testing is self-evident: If risk managers are to have confidence in their risk models, then those models need to be properly back-tested and to have performed well under those back-tests.

Back-tests can be used for three complementary purposes. The first is to assess whether a model's density forecasts are statistically compatible with the realized values of the underlying random variable. The second purpose is diagnostic: to generate feedback about the model's potential weaknesses to assist the

model builder and help him/her to "correct" the model. The third purpose is to rank alternative models. A good risk model should fare well by all three criteria: It should pass its statistical tests, should not generate any worrying diagnostics, and should rank well in comparison to alternative models.

The archetypal *market risk model* is a model that forecasts the *value at risk* (VaR) of a portfolio over one or more confidence levels, for a specified horizon. We will assume for the most part that the horizon is a trading day.

To back-test such a model, we need a dataset that consists of the model's forecasts, on the one hand, and the daily profits or losses (P/L) generated by the portfolio, on the other. The first task in back-testing is therefore to assemble such a dataset. For most market risk managers,

the forecasts themselves should be readily available. However, obtaining suitable profit and loss data is a more difficult problem than it might initially appear to be. The reason is that we do not need data on the profits or losses actually generated by a portfolio, but data on the profits or losses attributable to the market risks taken: We want P/L data that reflect underlying market volatility rather than accounting prudence. We also need to clean our P/L data to get rid of components that are not directly related to current or recent market risk-taking. Such components include fee income, hidden and unrealized P/L, earnings attributable to nonmarket risks, such as yields on corporate bonds, and the impact of intraday trading on P/L.

Having obtained our dataset, the next stage is to carry out a preliminary data analysis. We should plot a back-testing chart—a plot of the realized P/L over time with the VaR forecasts superimposed on it—and look for any odd or outstanding features. It is also good practice to supplement back-testing charts with P/L histograms, which sometimes give a clearer indication of the empirical P/L distribution, and quantile-quantile (QQ) charts, which plot the quantiles of an empirical P/L distribution against those of a forecasted P/L distribution. It is also a good idea to examine summary P/L statistics, including the obvious statistics of mean, variance, skewness, kurtosis, range, and so on and the number and size of extreme observations. A preliminary data analysis can be very helpful in enabling practitioners to get to know their data and get a feel for any problems they might encounter.

STATISTICAL BACK-TESTING

The first type of back-tests are statistical tests based on a hypothesis-testing paradigm. We first specify the null hypothesis that we wish to test—typically the null hypothesis is that the model is adequate—and select an alternative

hypothesis to be accepted if the null is rejected. We then select a significance level and estimate the probability associated with the null hypothesis being “true.” We would accept the null hypothesis if the estimated value of this probability, the estimated prob-value, exceeds the chosen significance level, and we would reject it otherwise. The higher the significance level, the more likely we are to accept the null hypothesis, and the less likely we are to incorrectly reject a true model (that is, to make a Type I error). Unfortunately, it also means that we are more likely to incorrectly accept a false model (that is, to make a Type II error). Any test therefore involves a trade-off between these two types of possible error. Ideally, we should select a significance level that takes account of the likelihoods and costs of these errors and strikes an appropriate balance between them. However, in practice, it is common to select some arbitrary significance level such as 5% and apply that level in all our tests. A significance level of this magnitude gives the model a certain benefit of the doubt, and implies that we would reject the model only if the evidence against it is reasonably strong.

EXCEEDANCE-BASED STATISTICAL APPROACHES

Suppose that we have a sample of n daily VaR forecasts VaR_t and a corresponding sample of n realized loss outcomes L_t , where t goes from 1 to n . L_t is denominated in units in which realized losses are positive and realized profits are negative.

Some common approaches to back-testing involve exceedance observations, where an exceedance observation (also called a tail loss) is a loss that exceeds the VaR. These exceedance observations h_t are obtained by putting our sample observations through the following transformation:

$$h_t = \begin{cases} 1 \\ 0 \end{cases} \quad \text{if} \quad \begin{cases} L_t > VaR_t \\ L_t \leq VaR_t \end{cases} \quad (1)$$

This transformation gives a unit value to all observations where there is a loss exceeding VaR and a zero value to all other observations.

Binomial (Kupiec) Approach

We can now apply the basic frequency (or binomial) test suggested by Kupiec (1995): We test whether the observed frequency of exceedances is consistent with the frequency predicted by the model. In particular, under the null hypothesis that the model is “good,” the number of exceedances x follows a *binomial distribution* with probability p , where p is the tail probability or 1 minus the confidence level. The probability of x exceedances given n observations is therefore:

$$\text{Prob}(x|n, p) = \binom{n}{x} p^x (1-p)^{n-x} \quad (2)$$

Equation (2) also tells us that the only information required to implement a binomial test is information about the values of n , p , and x . This probability is then calculated using a suitable calculation engine (e.g., using the “binomdist” function in Excel).

To illustrate, suppose $n = 1,000$ and we take the confidence level α to be 0.95. Our model therefore predicts that $p = 1 - \alpha = 0.05$ and the null hypothesis is $H_0: p = 0.05$. We then expect $np = 50$ exceedances under the null. Now suppose that the number of exceedances, x , is 60. This corresponds to an empirical frequency, \hat{p} , equal to 0.060. Since \hat{p} exceeds 0.05; we might specify a one-sided alternative hypothesis $H_1: p > 0.05$. The prob-value of the test is the probability under the null that $x \geq 60$. This is most easily calculated as $1 - \text{Pr}[x \leq 59]$, which equals 0.0867 given the values of n and p . At a conventional significance level such as 5%, we would then “pass” the model as acceptable. It is also clear that as x gets larger and moves further away from its predicted value of 50, then the probability of observing x exceedances will fall. Values of x with prob-values lower than our significance level would lead to rejections of

the null hypothesis and a “fail” result for the model. In fact, if we work with a 5% significance level, it is straightforward to show that we would accept the null if $x \leq 62$ and reject it if $x \geq 63$.

We can also apply binomial tests using a two-sided alternative hypothesis $H_1: p \neq 0.05$. We could do so by estimating a confidence interval for the number of exceedances and checking whether x lies within this interval. For example, if we want to test using a 5% significance level, we would estimate a 95% confidence interval for x , the bounds of which would delineate the lower and upper 5% tails of x 's density function. With $n = 1,000$ and $p = 0.05$, the 95% confidence interval for x is [36, 66]. We would then accept the null if x falls within this range and otherwise reject it.

A Normal Approximation

Testing can be simplified further if we work with a normal approximation to the binomial. Provided n is sufficiently large—and n would be sufficiently large with the sample sizes that risk managers typically work with—then the distribution of x is approximately normal with mean np and variance $np(1-p)$. This implies, in turn, that the variable $z = (x - np)/\sqrt{np(1-p)}$ is distributed as standard normal, and we can test whether the observed value of z is compatible with this distribution. For instance, if we wished to carry out a two-sided test, we know that the 95% confidence interval for a standard normal is [−1.96, +1.96], so we would accept the null if (and only if) z falls in this range.

Tests of Independence

Besides predicting that x should be binomial or approximately normal with large samples, the null hypothesis of model adequacy often leads to the prediction that x should be independent. “Independence” means that there should be no temporal pattern in the x series that is, the probability of the next observation being an

exceedance should be independent of whether any previous observation was an exceedance or not. Where this prediction arises, it is important that it be tested too: A bad model might pass the earlier tests, but still be inadequate because it produces predictable exceedances or clusters of exceedances that ought not to arise. Evidence of exceedance clustering would suggest that the model is misspecified, even if the model has the correct exceedance frequency.

One of the simplest independence tests is a *runs test*, in which we test whether the number of runs in a time series is consistent with what we would expect under independence. We can apply a runs test to any data that are time-ordered and expressed in binary form, as is the case with observations in our x series that either take the value 0 or the value 1. A run is then a sequence of consecutive identical numbers, and the number of runs R is equal to the number of sign changes plus 1. If u is the number of observations taking one value and v the number taking the other value, then under the independence null the mean and variance of the number of runs are, respectively:

$$\mu_R = 1 + \frac{2uv}{u+v} \quad (3)$$

$$\sigma_R^2 = \frac{2uv(2uv - u - v)}{(u+v)^2(u+v-1)} \quad (4)$$

If the total number of observations is large, then R is approximately normal and $z = (R - \mu_R)/\sigma_R$ approximately standard normal, and we can test accordingly.

A more sophisticated version of the same idea is suggested by Engle and Manganelli (2004): They propose estimating a binary regression model—that is, they regress h_i against possible explanatory variables, such as lagged returns or lagged squared returns—and then test for the joint insignificance of the explanatory variables. A binary regression approach is more powerful than a basic runs test because it can take account of the impact of other possible variables, which a runs test does not.

Conditional Testing (Christoffersen) Approach

We can also carry out tests of the distribution and independence of x within the same testing framework, and this takes us to the *conditional back-testing approach* of Christoffersen (1998). His idea is to separate out the particular predictions being tested and then test each prediction separately. We begin by rephrasing the earlier frequency or unconditional coverage test in likelihood ratio (LR) form.

Given that the observed frequency of exceedances is x/n , then under the hypothesis/prediction of correct unconditional coverage, the test statistic

$$LR_{uc} = -2 \ln[(1-p)^{n-x} p^x] + 2 \ln[(1-x/n)^{n-x} (x/n)^x] \quad (5)$$

is distributed as a $\chi^2(1)$, a chi-squared with 1 degree of freedom. As we can see from equation (5), this boils down to a test of whether the empirical frequency x/n is “close” to the predicted frequency p .

Turning to the independence prediction, let n_{ij} be the number of days that state j occurred after state i occurred the previous day, where the states refer to the occurrence or not of an exceedance, and let π_{ij} be the probability of state j in any given day, given that the previous day’s state was i . Under the hypothesis of independence, the test statistic

$$LR_{ind} = -2 \ln \left[(1 - \hat{\pi}_2)^{n_{00} + n_{11}} \hat{\pi}_2^{n_{01} + n_{11}} \right] + 2 \ln \left[(1 - \hat{\pi}_{01})^{n_{00}} \hat{\pi}_{01}^{n_{01}} (1 - \hat{\pi}_{11})^{n_{10}} \pi_{11}^{n_{11}} \right] \quad (6)$$

is also distributed as a $\chi^2(1)$, and note that we can recover estimates of the probabilities from

$$\begin{aligned} \hat{\pi}_{01} &= \frac{n_{01}}{n_{00} + n_{01}} \\ \hat{\pi}_{11} &= \frac{n_{11}}{n_{10} + n_{11}} \\ \hat{\pi}_2 &= \frac{n_{01} + n_{11}}{n_{00} + n_{10} + n_{01} + n_{11}} \end{aligned} \quad (7)$$

It follows that under the combined hypothesis of correct coverage and independence the test statistic

$$LR_{cc} = LR_{uc} + LR_{ind} \quad (8)$$

is distributed as $\chi^2(2)$. The *Christoffersen approach* enables us to test both coverage and independence hypotheses at the same time. Moreover, if the model fails such a test, this approach enables us to test each hypothesis separately, and so establish whether the model fails because of incorrect coverage or because of lack of independence.

Strengths and Limitations of Exceedance-Based Approaches

These exceedance tests have the advantages that they have a simple intuition, are easy to apply, and do not require a great deal of information. However, they often lack power (that is, the ability to identify bad models) except with very large sample sizes, because they throw potentially valuable information away: Focusing on tests of exceedances over VaR at a given confidence level is equivalent to throwing away information about the model's forecasts of VaRs at other confidence levels, and this discarded information often includes useful information about the sizes of tail losses predicted by a risk model (or information about VaRs at higher confidence levels). This can mean that a "bad" risk model will pass an exceedance-based test if it generates an acceptably accurate frequency of exceedances, even if its forecasts of losses larger than VaR are very poor.

STATISTICAL BACK-TESTING OF VaRs AT MULTIPLE CONFIDENCE LEVELS

This line of reasoning suggests that we should consider back-testing the performance of a model's VaR forecasts over multiple confidence

levels. Indeed, pushed to the limit, it suggests that we consider back-testing a model's VaR forecasts over all confidence levels at the same time. We would proceed by applying the following transformation:

$$p_t = F_t(X_t) \quad (9)$$

where $F_t(\cdot)$ is the (typically time-dependent) *probability-integral transformation* (PIT) that maps the realized one-day loss or profit, X_t , to its cumulative density value, where the forecast is made the previous day. So, for example, if our model specifies that losses are standard normal, then a value $X_t = 1.645$ would give us $p_t = F_t(1.645) = 0.95$, and so forth.

We can now deduce that p_t is stationary and distributed as standard uniform under the hypothesis that the VaR model is adequate. p_t is also independent because consecutive values of p_t have no common factors. Hence p_t is predicted to be independent and identically distributed (IID) $U(0,1)$ under the null hypothesis.

As an aside, it is worth noting at this point that the independence assumption does not arise in cases where we have a multi-step-ahead as opposed to a one-step-ahead VaR model: An example of the latter is a VaR model that produces daily VaR forecasts over a daily forecast horizon; an example of the former is a VaR model that produces daily VaR forecasts over a multi-day horizon. The forecast horizon is equal to one day in the one case, and equal to more than one day in the other. The p_t are predicted to be independent for one-day-ahead VaR forecasts because consecutive observations are not affected by common shocks; however, for multi-day forecasts, there is no independence prediction because consecutive p_t observations are subject to at least one common random factor. For example, the two-day return over Monday and Tuesday and the two-day return over Tuesday and Wednesday are both affected by the Tuesday daily return. This means that they have a common random factor and are therefore not independent. We will ignore multistep-ahead models in the rest of our discussion, but the

reader should keep in mind that we cannot assume independence for multi-step-ahead models or regard independence tests applied to such models as tests of model adequacy.

Testing Uniformity

Returning to the one-step-ahead case, we can now test our model by applying conventional uniformity tests. One of the best known of these is the Kolmogorov-Smirnov (KS) test. The KS test statistic D is then the maximum distance between the predicted cumulative density $F(x)$, which is a 45-degree line, and the empirical cumulative density $\hat{F}(x)$, evaluated over each data point X_t :

$$D = \max_t |F(X_t) - \hat{F}(X_t)| \quad (10)$$

The test value of the KS statistic is then compared to the relevant critical value and the null is accepted or rejected accordingly. This test is easy to implement because the test statistic is straightforward to calculate and its critical values are easily obtained using Monte Carlo simulation. However, the KS test tends to be more sensitive to the distributional differences near the center of the distribution, and is less sensitive at the tails. This is obviously a drawback when back-testing VaR models, where we are usually much more interested in the tail than in the central mass of a distribution.

A way around this latter problem is to replace the KS test with a Kuiper test. The Kuiper test statistic D^* is the sum of the maximum amount by which each distribution exceeds the other:

$$D^* = \max_t |F(X_t) - \hat{F}(X_t)| + \max_t |\hat{F}(X_t) - F(X_t)| \quad (11)$$

The Kuiper test can be implemented in much the same way as the KS test: Its test statistic is straightforward to calculate and its critical values can be obtained by Monte Carlo simulation. The Kuiper test has the advantage over the KS test that it is more sensitive to deviations in the tail regions. It is also believed to be more robust

to transformations in the data, and to be good at detecting cyclical and other features in the data. However, there is also evidence that it is very data intensive and needs large datasets to get reliable results.

We can also test uniformity by applying a textbook χ^2 test to binned (or classified) data). We divide the data into k classes and then compute the test statistic:

$$\sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (12)$$

where O_i is the observed frequency of data in bin i , and E_i is the expected frequency of data in bin i . Under the null hypothesis, this test statistic is distributed as $\chi^2(k - c)$, where c is the number of estimated parameters in the VaR model. The main disadvantage of the χ^2 test is that results are dependent on the way in which the data are binned and binning is (largely) arbitrary. In using it, we should be careful to check the sensitivity of results to alternative ways of binning the data.

Applying the Berkowitz Transformation and Testing for Standard Normality

It is often more convenient to put the p_t through a second (or *Berkowitz*) transformation to make them standard normal under the null of model adequacy; that is, we work with the transformed variable:

$$z_t = \Phi^{-1}(p_t) \quad (13)$$

where $\Phi(\cdot)$ is the standard normal distribution function (see Berkowitz, 2001). This second transformation is helpful because *testing for standard normality* is more convenient than testing for standard uniformity, and because a normal variable is more convenient when dealing with temporal dependence. Under the null, z_t will be distributed as IID standard normal [denoted by IID $N(0,1)$].

Testing model adequacy now boils down to testing whether z_t is distributed as IID $N(0,1)$. There are two distinct tasks here:

1. We need to test whether z_t is $N(0,1)$, taking as given that z_t is IID, and there are various tests we might apply. If z_t is standard normal, then it should have a zero mean, a variance of 1, a zero skew, and a kurtosis of 3. Assuming IID, we can test the mean prediction using a z -test or t -test, we can test the variance prediction using a variance ratio test, and we can test the skewness and kurtosis predictions using a Jarque-Bera test, which can also be regarded as a test of normality itself. All these tests are conventional textbook tests and are easy to apply.
2. We need to test whether z_t is IID, and there are many tests of the IID prediction. These include runs and binary regression tests, which we have already discussed above. We can also estimate the autocorrelation structure of our z_t observations or fit an autoregressive moving average (ARMA) process to them. All the parameters in an autocorrelation function or an ARMA process should be insignificant, and we can test for their significance using standard tests such as a Box-Pierce Q test. Another possibility, if we have enough data, is to test independence using a BDS test (Brock et al., 1987): a BDS test is very powerful, but also data-intensive.

Since the hypothesis of model adequacy predicts both $N(0,1)$ and IID, it is important to note that the model must “pass” both types of test if it is to “pass” overall.

Tests Applied to Truncated Distributions

There are also situations where we are only interested in part of the P/L distribution: For example, we might be interested only in the distribution of losses in excess of VaR. If we are working to a confidence level α , we can take our earlier p_t series and delete all nontail obser-

vations from it. We then end up with a series that is IID uniformly distributed over the interval $[0, 1 - \alpha]$, and this implies that $p_t/(1 - \alpha)$ is IID uniformly distributed over the interval $[0, 1]$. We can test this prediction using one of the uniformity tests discussed earlier. If we wish to, we can apply the Berkowitz transformation to $p_t/(1 - \alpha)$ to obtain the series $z_t = \Phi^{-1}(p_t/(1 - \alpha))$, which is distributed as IID $N(0,1)$ under the null. We can then apply the tests just discussed.

USING BACK-TESTS FOR DIAGNOSTIC PURPOSES

We can also modify many of these back-test procedures to help diagnose problems with our VaR model. Model diagnosis is a key ingredient to successful model building, and requires the modeler to be on the lookout for evidence of possible problems. So, to use an earlier example, if we have 60 exceedances out of a sample of 1,000 and we are operating to a VaR at the 95% confidence level, then we know that this is associated with a prob-value of 0.0867. Were we carrying out a formal back-test of model adequacy at a conventional significance level such as 5%, we would dismiss this result as statistically insignificant because the significance level gives the model the benefit of the doubt. However, for diagnostic purposes we do not wish to give the model the benefit of the doubt: Instead, we are looking for evidence “against” the model, even if that evidence is statistically “weak.” In these circumstances, a result like this would lead us to suspect whether the model has a tendency to underestimate the VaR. A wise risk manager would then start to ask whether other evidence could be found that would confirm or refute this suspicion. And, to put the same point a little differently, the last thing a risk manager should do in the face of such evidence is to wait and do nothing till the evidence has become overwhelming: The risk manager should act in a timely manner on the basis of any reasonable evidence available.

Independence tests can also be useful diagnostic tools. If we apply an independence test and the test result gives us some (not necessarily strong) reason to suspect that the model does not satisfy a valid independence prediction, then we can interpret this evidence as suggesting that there might be some dynamic misspecification in our model: Even if the broad coverage is about right, there might still be something wrong with the updating of our VaR forecasts from one day to the next. So, for example, if we have a parametric VaR model, then we might suspect that a key parameter in the model was not being updated efficiently, and the obvious suspects would be volatility or correlation parameters. Again, the evidence might be statistically “weak,” but even weak evidence can be useful in pointing to areas of weakness in the model.

Another useful *diagnostic* is provided by empirical moments of the Berkowitz-transformed series (see equation (13) above), which we saw earlier are predicted to be standard normal under the null of model adequacy. Some very useful diagnostic information can then be obtained by estimating their sample moments and considering any departures from their predicted values:

- If the sample mean is different from zero, we might suspect whether the model’s forecasts are biased in one direction or the other.
- If sample variance is less than 1, we might suspect that the model’s predicted dispersion is too low, in which case the model might overestimate risk; and if the sample variance is greater than 1, we might suspect that the predicted dispersion is too high and the model underestimates risk.
- If the sample skew is positive or negative, we might suspect that the forecasts are skewed in one direction or the other.
- If the sample kurtosis is less than 3 or (as is more likely in risk management contexts) bigger than 3, we might ask ourselves if the model is overestimating or underestimating its tails.

In each of these cases, we should also check the strength of the evidence and we can do so by applying the relevant tests and checking out their prob-values: The lower the prob-value, the stronger the evidence against the model. However, since we are especially concerned in risk management with the possibility that the model might underestimate risks, then a sample variance that considerably exceeds 1 or a sample kurtosis that considerably exceeds 3 is potentially important evidence that might warrant further scrutiny.

RANKING ALTERNATIVE MODELS

It is often the case that we are interested in how different models compare to each other. We can compare models using *forecast evaluation* methods that give each model a score in terms of some loss function; we then use the loss scores to rank the models—the lower the loss, the better the model. These approaches are not statistical tests of model adequacy and this means that they do not suffer from the low power of tests such as frequency tests: This makes them attractive for back-testing with the datasets typically available in real-world applications. In addition, they also allow us to tailor the loss function to take account of particular concerns: For example, we might be more concerned about higher losses than lower losses, and might therefore wish to give higher losses a greater weight in our loss function.

The ranking process has four key ingredients for each model:

1. A set of n paired observations—paired observations of losses or profits for each period and their associated VaR forecasts.
2. A loss function that gives each observation a score C_t depending on how the observed loss or profit compares to the VaR forecasted for that period.
3. A benchmark, which gives us an idea of the score we could expect from a “good” model.

4. A score function, which takes as its inputs our loss-function and benchmark values.

We need to specify the loss function, and a number of different loss functions have been proposed. Perhaps the most straightforward is the binary loss function proposed by Lopez (1998, p. 121), which gives exceedance observations a value of 1 and other observations a value of 0. C_t is then as follows:

$$C_t = \begin{cases} 1 & \text{if } L_t > VaR_t \\ 0 & \text{if } L_t \leq VaR_t \end{cases} \quad (14)$$

This loss function is intended for the user who is (exclusively) concerned with the frequency of exceedances. The natural benchmark for this loss function is p , the exceedance probability or expected value of $E(C_t)$. If we take our benchmark to be the expected value of C_t under the null hypothesis that the model is “good,” then Lopez (1998) suggests that a good choice of score function is the following quadratic probability score (QPS) function:

$$QPS = \frac{2}{n} \sum_{t=1}^n (C_t - p)^2 \quad (15)$$

The QPS takes a value in the range $[0,2]$, and the closer the QPS-value to zero, the better the model. We can therefore use the QPS (or some similar score function) to rank our models, with the better models having the lower scores. In addition, the QPS criterion has the attractive property that it (usually) encourages truth-telling by VaR modelers: If VaR modelers wish to minimize their QPS score, they will (usually) report their VaRs “truthfully.” This is a useful property in situations where the back-tester and the VaR modeler are different, and where the back-tester might be concerned about the VaR modeler reporting false VaR forecasts to alter the results of the back-test.

A drawback of this loss function is that it ignores the magnitude of tail losses. If we wish to remedy this defect, Lopez suggests a second,

size-adjusted, loss function:

$$C_t = \begin{cases} 1 + (L_t - VaR_t)^2 & \text{if } L_t > VaR \\ 0 & \text{if } L_t \leq VaR_t \end{cases} \quad (16)$$

This loss function allows for the sizes of tail losses in a way that (15) does not: A model that generates higher tail losses would generate higher values of (16) than one that generates lower tail losses, other things being equal. However, with this loss function, there is no longer a straightforward condition for the benchmark, so we need to estimate the benchmark by some other means (e.g., Monte Carlo simulation). The size-adjusted loss function (17) also has the drawback that it loses some of its intuition, because squared monetary returns have no ready monetary interpretation.

A way around this last problem is suggested by Blanco and Ihle (1998), who suggest the following loss function:

$$C_t = \begin{cases} (L_t - VaR_t)/VaR_t & \text{if } L_t > VaR_t \\ 0 & \text{if } L_t \leq VaR_t \end{cases} \quad (17)$$

This loss function gives each tail-loss observation a weight equal to the tail loss divided by the VaR. This has a nice intuition and ensures that higher tail losses get awarded higher C_t values without the impaired intuition introduced by squaring the tail loss. The benchmark for this forecast evaluation procedure is also easy to derive: The benchmark is equal to the difference between the Expected Shortfall (ES) and the VaR, divided by the VaR. However, the Blanco-Ihle loss function also has a problem of its own: Because (17) has the VaR as its denominator, it is not defined if the VaR is zero, and can give awkward answers if VaR gets “close” to zero or becomes negative. We should therefore only use it if we can be confident of the VaR being sufficiently large and positive.

We therefore seek a size-based loss function that avoids the squared term in the second Lopez loss function, but also avoids denominators that might be zero-valued. A promising

candidate is the tail loss itself:

$$C_t = \begin{cases} L_t & \text{if } L_t > VaR \\ 0 & \text{if } L_t \leq VaR_t \end{cases} \quad (18)$$

The expected value of the tail loss is of course the ES, so we can choose the ES as our benchmark and use a quadratic score function such as:

$$QS = \frac{2}{n} \sum_{t=1}^n (C_t - ES_t)^2 \quad (19)$$

This approach penalizes deviations of tail losses from their expected value, which makes intuitive sense. Moreover, because it is quadratic, it gives very high tail losses much greater weight than more common tail losses, and thereby comes down hard on large losses.

KEY POINTS

- In general, back-testing is the quantitative evaluation of a model. When back-testing is applied to a risk or probability density forecasting model, it involves a comparison of the model's density forecasts against subsequently realized outcomes of the random variable whose density is forecast.
- The main purposes of back-testing market risk models are to test model adequacy, to diagnose potential model problems, and to compare or rank alternative models. A good risk model should fare well by all three criteria: It should pass its statistical tests, should not generate any worrying diagnostics, and should rank well in comparison to alternative models.
- Because the typical market risk model is a model that forecasts the value-at-risk of a portfolio over one or more confidence levels for a specified horizon, back-testing of market risk models involves some comparison of VaR forecasts against subsequently realized values of profit or loss.
- Formal tests of market risk model adequacy can be applied to the frequency and inde-

pendence of exceedance observations, but can also be applied to forecasts of VaR at multiple confidence levels.

- Comparable approaches can be used for model diagnostic purposes, where the main concern is not to test model adequacy in a formal way, as such, but instead to gather evidence of possible model misspecification.
- Simple loss-scoring approaches can be used to rank the forecast performance of alternative models.

REFERENCES

- Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics* 19, 4: 465–474.
- Blanco, C., and Ihle, G. (1999). How good is your VaR? Using back-testing to assess system performance. *Financial Engineering News* 11, August: 1–2.
- Campbell, S. (2007). A review of back-testing and back-testing procedures. *Journal of Risk* 9, 2: 1–17.
- Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review* 39, 4: 841–862.
- Crnkovic, C., and Drachman, J. (1996). Quality control. *Risk* 9, 9: 139–143.
- Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review* 39, 4: 863–883.
- Dowd, K. (2004). A modified Berkowitz back-test. *Risk* 17, 4: 86.
- Dowd, K. (2005). *Measuring Market Risk*, 2nd edition. Chichester: John Wiley & Sons.
- Engle, R. F., and Manganelli, S. (2004). CAViaR: Conditional autoregressive value-at-risk by regression quantiles. *Journal of Business and Economic Statistics* 22, 4: 367–381.
- Hendricks, D. (1996). Evaluation of value-at-risk models using historical data. *Federal Reserve Bank of New York Economic Policy Review* 2 (April): 39–70.
- Kupiec, P. (1995). Techniques for verifying the accuracy of risk management models. *Journal of Derivatives* 4, 3: 73–84.
- Lopez, J. A. (1998). Regulatory evaluation of value-at-risk models. *Federal Reserve Bank of New York Economic Policy Review* 4, 3: 119–124.

Estimating Liquidity Risks

KEVIN DOWD, PhD

Partner, Cobden Partners, London

Abstract: The measurement of liquidity risk risks is an underdeveloped area of market risk measurement and management. Liquidity issues affect the estimation of conventional market risk measures, but the measurement of liquidity risks is an important subject in its own right. Liquidity issues also figure prominently in periods of market crisis. There are various easily implementable and often complementary approaches to the estimation of liquidity-adjusted Value-at-Risk: These involve modeling the bid-ask spread or the liquidity discount incurred when liquidating a position. There are also approaches to the modelling of Liquidity-at-Risk, which deals with the riskiness of cash flows, in both noncrisis and crisis situations.

Market practitioners often assume that markets are liquid—that is, that we can liquidate or unwind positions at going market prices, usually taken to be the mean of bid and ask prices, without too much difficulty or cost. This assumption is very convenient and provides a justification for the practice of marking positions to market prices. However, it is often empirically questionable and the failure to allow for it can seriously undermine market risk measurement. In any case, liquidity risk is a major risk factor in its own right, and we will often want to measure it too.

This entry looks at liquidity issues and how they affect the estimation of market and *liquidity risk* measures. *Liquidity* issues affect market risk measurement through their impact on standard measures of market risk. In addition, because effective market risk management involves an ability to estimate and manage liquidity risk itself, we also need to be able to estimate liquidity risk—or liquidity-at-risk. Finally, since liquidity

problems are particularly prominent in market crises, we also need to address how to estimate crisis-related market risks and liquidity risks. Accordingly, the main themes of this entry are:

- The nature of market liquidity and illiquidity, and their associated costs and risks.
- The estimation of *value-at-risk (VaR)* in illiquid or partially liquid markets—*liquidity-adjusted VaR (or LVaR)*.
- Estimating *liquidity-at-risk (LaR)*.
- Estimating *crisis-related liquidity risks*.

For convenience, and to be faithful to the literature, we focus on the (discredited, but computationally convenient) VaR risk measure, but we should note that any of the approaches suggested here can be adapted to estimate superior risk measures such as coherent risk measures (see, e.g., Artzner et al., 1999) or any other quantile-based risk measures. For example, estimates of these alternative risk measures can be obtained using the “average tail VaR”

approach set out in Dowd (2005, Chapter 3): This is based on the idea that, since the VaR is a quantile, any of these other quantile-based risk measurements can be estimated as a weighted average of VaRs predicated on a suitable range of confidence levels.

LIQUIDITY AND LIQUIDITY RISKS

The notion of liquidity refers to the ability of a trader to execute a trade or liquidate a position with little or no cost, risk, or inconvenience. Liquidity is a function of the market, and depends on such factors as the number of traders in the market, the frequency and size of trades, the time it takes to carry out a trade, and the cost (and sometimes risk) of transacting. It also depends on the commodity or instrument traded, and more standardized instruments (e.g., such as FX or equities) tend to have more liquid markets than nonstandardized or tailor-made instruments (e.g., such as over-the-counter [OTC] derivatives). Markets vary greatly in their liquidity: Markets such as the FX market and the big stock markets are (generally) highly liquid; but other markets are less so, particularly those for many OTC instruments and instruments that are usually held to maturity and, hence, are rarely traded once initially bought. However, even the “big” standardized markets are not perfectly liquid—their liquidity fluctuates over time and can fall dramatically in a crisis—so we cannot take their liquidity for granted.

Imperfect liquidity also implies that there is no such thing as the going market price. Instead, there are two going market prices—an ask price, which is the price at which a trader sells, and a (lower) bid price, which is the price at which a trader buys. The “market” price often quoted is just an average of the bid and ask prices, and this price is fictional because no one actually trades at this price. The difference between the bid and ask prices is a cost of liquidity, and in principle we should allow for this cost in estimating market risk measures.

The bid-ask spread also has an associated risk, because the spread itself is a random variable. This means there is some risk associated with the price we can obtain, even if the fictional mid-spread price is given. Other things being equal, if the spread rises, the costs of closing out our position will rise, so the risk that the spread will rise should be factored into our risk measures along with the usual market price risk.

We should also take account of a further distinction. If our position is small relative to the size of the market (e.g., because we are a very small player in a very large market), then our trading should have a negligible impact on the market price. In such circumstances we can regard the bid-ask spread as exogenous to us, and we can assume that the spread is determined by the market beyond our control. However, if our position is large relative to the market, our activities will have a noticeable effect on the market itself and can affect both the market price and the bid-ask spread. For example, if we suddenly unload a large position, we should expect the market price to fall and the bid-ask spread to widen. This is partly because there is a limited market, and prices must move to induce other traders to buy. A second reason is a little more subtle: Large trades often reveal information, and the perception that they do will cause other traders to revise their views. Consequently, a large sale may encourage other traders to revise downward their assessment of the prospects for the instrument concerned, and this will further depress the price. In these circumstances the market price and the bid-ask spread are to some extent endogenous (i.e., responsive to our trading activities) and we should take account of how the market might react to us when estimating liquidity costs and risks. Other things again being equal, the bigger our trade, the bigger the impact we should expect it to have on market prices.

In sum, we are concerned with both liquidity costs and liquidity risks, and we need to take account of the difference between exogenous and endogenous liquidity. We now consider some of

the approaches available to adjust our estimates of VaR to take account of these factors.

ESTIMATING LIQUIDITY-ADJUSTED VaR

There are many ways we could estimate liquidity-adjusted VaR. These vary in their degrees of sophistication and in their ease (or otherwise) of implementation, and there is no single “best” method. However, sophisticated approaches are not necessarily more useful than more basic ones, and the best method, even if we could establish what it is, is not necessarily better than a collection of inferior ones. Instead, what we really seek are simple-to-implement (i.e., spreadsheet-executable) approaches that are transparent in terms of their underlying assumptions; in effect, we are looking for liquidity “add-ons” that allow us to modify original VaR estimates that were obtained without any consideration for liquidity. We can then easily assess the impact of our assumptions on our estimates of VaR. Moreover, there is a premium on compatibility, because different methods look at different aspects of illiquidity, and it can be helpful to combine them to get some sense of an overall liquidity adjustment. Because of this, a really good method might not always be as useful as two inferior methods that actually work well together.

Whichever models we used, we also need to check their sensitivities—how does the liquidity adjustment change as we change the confidence level, holding period, or any other parameters? A priori, we should have some idea of what these should be (e.g., that the liquidity adjustment should fall as the holding period rises, etc.), and we need to satisfy ourselves that the models we use have sensitivities of the right sign and approximate magnitude. Going further, we should also try to ensure that models are calibrated against real data (e.g., bid-ask spread parameters should be empirically plausible, etc.) and be properly stress-tested

and back-tested. In addition, we should keep in mind that different approaches are often suited to different problems, and we should not seek a best approach to the exclusion of any others. In the final analysis, liquidity issues are much more subtle than they look, and there is no established consensus on how we should deal with them. So perhaps the best advice is for risk measurers to hedge their bets, and use different approaches to highlight different liquidity concerns.

The Constant Spread Approach

Ideally, if we had actual transaction prices, we could infer the actual returns obtained by traders, in which case conventional VaR methods would take account of spread liquidity factors without the need for any further adjustment. In such cases, we would model actual returns (taking account of how they depend on market volume, etc.), infer a relevant conditional distribution (e.g., a t), and plug in the values of the parameters concerned into an appropriate parametric VaR equation. For more on how this might be done, see Giot and Grammig (2003).

However, practitioners often lack such data and have to work with market prices that are averages of bid and ask prices. They might then attempt to take account of liquidity factors by working with the bid-ask spread, and the simplest way to incorporate liquidity risk into a VaR calculation is in terms of a spread that is assumed to be constant. If we make this assumption, the liquidity cost is then equal to half the spread times the size of the position liquidated. Using obvious notation, this means that we add the following liquidity cost (LC) to a “standard” VaR:

$$LC = \frac{1}{2} \text{spread} * P \quad (1)$$

where *spread* is expressed as actual spread divided by the midpoint. For the sake of comparison and using obvious notation, let us compare

this to a benchmark conventional lognormal VaR with no adjustment for liquidity risk:

$$VaR = P [1 - \exp(\mu_R - \sigma_R z_\alpha)] \quad (2)$$

where the returns have been calculated using prices that are the midpoints of the bid-ask spread. The liquidity-adjusted VaR, $LVaR$, is then given by:

$$LVaR = VaR + LC = P[1 - \exp(\mu_R - \sigma_R z_\alpha) + \frac{1}{2}spread] \quad (3)$$

Setting $\mu_R = 0$ to clarify matters, the ratio of $LVaR$ to VaR is then

$$\frac{LVaR}{VaR} = 1 + \frac{spread}{2[1 - \exp(-\sigma_R z_\alpha)]} \quad (4)$$

It is easy to show that the liquidity adjustment (a) rises in proportion with the assumed spread, (b) falls as the confidence level increases, and (c) falls as the holding periods each increase. The first and third of these are obviously correct, but the second implication is one that may or may not be compatible with one's prior expectations.

This approach is easy to implement and requires only minimal information, but the assumption of a constant spread is highly implausible, and it takes no account of any other liquidity factors.

The Exogenous Spread Approach

A superior alternative is to assume that traders face random spreads. If our position is sufficiently small relative to the market, we can also regard our spread risk as exogenous to us (i.e., independent of our own trading), for any given holding period. We could assume any process for the spread that we believe to be empirically plausible. For example, we might believe that the spread is normally distributed:

$$spread \sim N(\mu_{spread}, \sigma_{spread}^2) \quad (5)$$

where μ_{spread} is the mean spread and σ_{spread} is the spread volatility. Alternatively, we might use some heavy-tailed distribution to accommodate excess kurtosis in the spread.

We could now estimate the $LVaR$ using Monte Carlo simulation: We could simulate both P and the spread, incorporate the spread into P to get liquidity-adjusted prices, and then infer the liquidity-adjusted VaR from the distribution of simulated liquidity-adjusted prices.

However, in practice, we might take a shortcut suggested by Bangia et al. (1999). They suggest that we specify the liquidity cost (LC) as:

$$LC = \frac{P}{2}(\mu_{spread} + k\sigma_{spread}) \quad (6)$$

where k is some parameter whose value is to be determined. The value of k could be determined by a suitably calibrated Monte Carlo exercise, but they suggest that a particular value ($k = 3$) is plausible (e.g., because it reflects the empirical facts that spreads appear to have excess kurtosis and are negatively correlated with returns, etc.). The liquidity-adjusted VaR, $LVaR$, is then equal to the conventional VaR plus the liquidity adjustment (6):

$$LVaR = VaR + LC = P[1 - \exp(\mu_R - \sigma_R z_\alpha) + \frac{P}{2}(\mu_{spread} + 3\sigma_{spread})] \quad (7)$$

Observe that this $LVaR$ incorporates (3) as a special case when $\sigma_{spread} = 0$. It therefore retains many of the properties of (3), but generalizes from (3) in allowing for the spread volatility as well. The ratio of $LVaR$ to VaR is then:

$$\frac{LVaR}{VaR} = 1 + \frac{LC}{VaR} = 1 + \frac{1}{2} \frac{(\mu_{spread} + 3\sigma_{spread})}{[1 - \exp(-\sigma_R z_\alpha)]} \quad (8)$$

This immediately tells us that the spread volatility σ_{spread} serves to increase the liquidity adjustment relative to the earlier case. The Bangia et al. framework was also further developed by Erwan (2002), who presented empirical results that are similar to the illustrative ones presented here in suggesting that the liquidity adjustment can make a big difference to our VaR estimates.

Endogenous-Price Approaches

The previous approaches assume that prices are exogenous and therefore ignore the possibility of the market price responding to our trading. However, we have also noted that this is often unreasonable, and we may wish to make a liquidity adjustment that reflects the response of the market to our own trading. If we sell, and the act of selling reduces the price, then this market-price response creates an additional loss relative to the case where the market price is exogenous, and we need to add this extra loss to our VaR. The liquidity-adjustment will also depend on the responsiveness of market prices to our trade: The more responsive the market price, the bigger the loss.

We can estimate this extra loss in various ways, but the simplest is to make use of some elementary economic theory. We begin with the notion of the price elasticity of demand, η , defined as the ratio of the proportional change in price divided by the proportional change in quantity demanded:

$$\eta = \frac{\Delta P/P}{\Delta N/N} < 0; \quad \Delta N/N > 0 \quad (9)$$

where in this context N is the size of the market and ΔN is the size of our trade. $\Delta N/N$ is therefore the size of our trade relative to the size of the market. The impact of the trade on the price is therefore

$$\frac{\Delta P}{P} = \eta \frac{\Delta N}{N} \quad (10)$$

We can therefore estimate $\Delta P/P$ on the basis of information about η and $\Delta N/N$, and both of these can be readily guessed at using a combination of economic and market judgement. The LVaR is then:

$$LVaR = VaR \left(1 - \frac{\Delta P}{P}\right) = VaR \left(1 - \eta \frac{\Delta N}{N}\right) \quad (11)$$

bearing in mind that the change in price is negative. The ratio of LVaR to VaR is therefore:

$$\frac{LVaR}{VaR} = 1 - \eta \frac{\Delta N}{N} \quad (12)$$

This gives us a very simple liquidity adjustment that depends on only two easily calibrated parameters. It is even independent of the VaR itself: The adjustment is the same regardless of whether the VaR is normal, lognormal, etc.

The ratio of LVaR to VaR thus depends entirely on the elasticity of demand η and the size of our trade relative to the size of the market ($\Delta N/N$).

This type of approach is easy to implement, and it is of considerable use in situations where we are concerned about the impact on VaR of endogenous market responses to our trading activity, as might be the case where we have large portfolios in thin markets. However, it is also narrow in focus and entirely ignores bid-ask spreads and transactions costs.

On the other hand, the fact that this approach focuses only on endogenous liquidity and the earlier ones focus on exogenous liquidity means that this last approach can easily be combined with one of the others; in effect, we can add one adjustment to the other. Thus, two very simple approaches can be added to produce an adjustment that addresses both exogenous and endogenous liquidity risk. This combined adjustment is given by

$$\frac{LVaR}{VaR} \Big|_{combined} = \frac{LVaR}{VaR} \Big|_{exogenous} + \frac{LVaR}{VaR} \Big|_{endogenous} \quad (13)$$

The Liquidity Discount Approach

A more sophisticated approach is suggested by Jarrow and Subramanian (1997). They consider a trader who faces an optimal liquidation problem—the trader must liquidate his or her position within a certain period of time to maximize expected utility, and seeks the best way to do so. Their approach is impressive, as it encompasses exogenous and endogenous market liquidity, spread cost, spread risk, an endogenous holding period, and an optimal liquidation policy.

Their analysis suggests that we should modify the traditional VaR in three ways. First, instead of using some arbitrary holding period,

we should use an optimal holding period determined by the solution to the trader's expected-utility optimization problem, which takes into account liquidity considerations and the possible impact of the trader's own trading strategy on the prices obtained. We should also add the average liquidity discount to the trader's losses (or subtract it from our prices) to take account of the expected loss from the selling process. Finally, their analysis also suggests that the volatility term should take account of the volatility of the time to liquidation and the volatility of the liquidity discount factor, as well as the volatility of the underlying market price.

To spell out their approach more formally, assume that prices between trades follow a geometric Brownian motion with parameters μ and σ . The current time is 0 and the price at time t is $p(t)$, so that geometric returns $\log(p(t)/p(0))$ are normally distributed. However, the prices actually obtained from trading are discounted from $p(t)$; more specifically, the prices obtained are $p(t)c(s)$, where $c(s)$ is a random quantity-dependent proportional discount factor, s is the amount traded, $0 \leq c(s) \leq 1$ and, other things being equal, $c(s)$ falls as s rises. Any order placed at time t will be also be subject to a random execution lag $\Delta(s)$, and therefore take place at time $t + \Delta(s)$. Other things again being equal, the execution lag $\Delta(s)$ rises with s : Bigger orders usually take longer to carry out. Our trader has S shares and wishes to maximize the present value of his or her current position, assuming that it is liquidated by the end of some horizon t , taking account of all relevant factors, including both the quantity discount $c(s)$ and the execution lag $\Delta(s)$. After solving for this problem, they produce the following expression for the liquidity-adjusted VaR:

$$\begin{aligned} LVaR &= P \left\{ E[\ln(p(\Delta(S))c(S)/p(0))] \right. \\ &\quad \left. + \text{std}[\ln(p(\Delta(S))c(S)/p(0))]z_\alpha \right\} \quad (14) \\ &= P \left\{ \left(\mu - \frac{\sigma^2}{2} \right) \mu_{\Delta(S)} + \mu_{\ln c(S)} \right. \\ &\quad \left. + \left[\sigma \sqrt{\mu_{\Delta(S)}} + \left(\mu - \frac{\sigma^2}{2} \right) \sigma_{\Delta(S)} + \sigma_{\ln c(S)} \right] z_\alpha \right\} \end{aligned}$$

where all parameters have the obvious interpretations. This expression differs from the conventional VaR in three ways. First, the liquidation horizon t in the conventional VaR is replaced by the expected execution lag $\mu_{\Delta(S)}$ in selling S shares. Clearly, the bigger is S , the longer the expected execution lag. Second, the LVaR takes account of the expected discount $\mu_{\ln c(s)}$ on the shares to be sold. And, third, the volatility σ in the conventional VaR is supplemented by additional terms related to $\sigma_{\Delta(s)}$ and $\sigma_{\ln c(s)}$, which reflect the volatilities of the execution time and the quantity discount. Note, too, that if our liquidity imperfections disappear, then $\mu_{\Delta(S)} = t$, $\sigma_{\Delta(S)} = 0$, and $c(S) = 1$ (which in turn implies $\mu_{\ln c(s)} = \sigma_{\ln c(s)} = 0$) and our LVaR (14) collapses to a conventional VaR as a special case—which is exactly as it should be.

To use this LVaR expression requires estimates of the usual Brownian motion parameters μ and σ , as well as estimates of the liquidity parameters $\mu_{\Delta(S)}$, $\sigma_{\Delta(S)}$, $\mu_{\ln c(s)}$ and $\sigma_{\ln c(s)}$, all of which are fairly easily obtained. The approach is therefore not too difficult to implement. All we have to do is then plug these parameters into (14) to obtain our LVaR.

ESTIMATING LIQUIDITY-AT-RISK (LAR)

We turn now to liquidity-at-risk (LaR), sometimes also known as cash flow-at-risk (CFaR). LaR (or CFaR) relates to the risk attached to prospective cash flows over a defined horizon period, and can be defined in terms analogous to the VaR. Thus, the LaR is the maximum likely cash outflow over the horizon period at a specified confidence level: for example, the 1-day LaR at the 95% confidence level is the maximum likely cash outflow over the next day, at the 95% confidence level, and so forth. A positive LaR means that the likely worst outcome, from a cash flow perspective, is an outflow of cash; and a negative LaR means that the likely worst outcome is an inflow of cash. The LaR is the cash flow equivalent to the VaR, but whereas

VaR deals with the risk of losses (or profits), LaR deals with the risk of cash outflows (or inflows).

These cash flow risks are quite different from the risks of liquidity-related losses. Nonetheless, they are closely related to these latter risks, and we might use LaR analysis as an input to evaluate them. Indeed, the use of LaR for such purposes is an important liquidity management tool.

An important point to appreciate about LaR is that the amounts involved can be very different from the amounts involved with VaR. Suppose for the sake of illustration that we have a large market-risk position that we hedge with a futures hedge of much the same amount. If the hedge is a good one, the basis or net risk remaining should be fairly small, and our VaR estimates should reflect that low basis risk and be relatively small themselves. However, the futures hedge leaves us exposed to the possibility of margin calls, and our exposure to margin calls will be related to the size of the futures position, which corresponds to the gross size of our original position. Thus, the VaR depends largely on the netted or hedged position, whilst the LaR depends on the larger gross position. If the hedge is a good one, the basis risk (or the VaR) will be low relative to the gross risk of the hedge position (or the LaR), and so the LaR can easily be an order of magnitude greater than the VaR. On the other hand, there are also many market risk positions that have positive VaR, but little or no cash flow risk (e.g., a portfolio of long European option positions, which generates no cash flows until the position is sold or the options expire), and in such cases the VaR will dwarf the LaR. So the LaR can be much greater than the VaR or much less than it, depending on the circumstances.

As we might expect, the LaR is potentially sensitive to any factors or activities, risky or otherwise, that might affect future cash flows. These include:

- Borrowing or lending, the impact of which on future cash flows is obvious.

- Margin requirements on market risk positions that are subject to daily marking-to-market.
- Collateral obligations, such as those on swaps, which can generate inflows or outflows of cash depending on the way the market moves. Collateral obligations can also change when counterparties like brokers alter them in response to changes in volatility, and collateral requirements on credit-sensitive positions (e.g., such as default-risky debt or credit derivatives) can change in response to credit events such as credit-downgrades.
- Unexpected cash flows can be triggered by the exercise of options, including the exercise of convertibility features on convertible debt and call features on callable debt.
- Changes in risk management policy; for instance, a switch from a futures hedge to an options hedge can have a major impact on cash flow risks, because the futures position is subject to margin requirements and marking to market whilst a (long) option position is not.

Two other points are also worth emphasizing here. The first is that obligations to make cash payments often come at bad times for the firms concerned, because they are often triggered by bad events. The standard example is where a firm suffers a credit downgrade that leads to an increase in its funding costs, and yet this same event also triggers a higher collateral requirement on some existing (e.g., swap) position and so generates an obligation to make a cash payment. It is axiomatic in many markets that firms get hit when they are most vulnerable. The second point is that positions that might be similar from a market risk perspective (e.g., such as a futures hedge and an options hedge) might have very different cash flow risks. The difference in cash flow risks arises, not so much because of differences in market risk characteristics, but because the positions have different *credit* risk characteristics, and it is the measures taken to manage the credit risk—the margin

and collateral requirements, and so on—that generate the differences in cash flow risks.

We can estimate LaR using many of the same methods used to estimate VaR and other measures of market risk. One approach, suggested by Singer (1997), is to use our existing VaR estimation tools to estimate the VaRs of marginable securities only (i.e., those where P/L translates directly into cash flows), thus allowing us to infer an LaR directly from the VaR. We could then combine this LaR estimate with comparable figures from other sources of liquidity risk within the organization (e.g., such as estimates of LaR arising from the corporate treasury) to produce an integrated measure of firm-wide liquidity risk. The beauty of this strategy is that it makes the best of the risk measurement capabilities that already exist within the firm, and effectively tweaks them to estimate liquidity risks.

However, this approach is also fairly rough and ready, and cannot be relied upon when the firm faces particularly complex liquidity risks. In such circumstances, it is often better to build a liquidity-risk measurement model from scratch, and we can start by setting out the basic types of cash flow to be considered. These might include:

- Known certain (or near certain) cash flows (e.g., income from government bonds, etc.). These are very easy to handle because we know them in advance.
- Unconditional uncertain cash flows (e.g., income from default-risky bonds, etc.). These are uncertain cash flows, which we model in terms of the probability density functions (pdfs) (i.e., we choose appropriate distributions, assign parameter values, etc.).
- Conditional uncertain cash flows. These are uncertain cash flows that depend on other variables (e.g., a cash flow might depend on whether we proceeded with a certain investment, and so we would model the cash flow in terms of a pdf, conditional on that investment); other conditioning variables that might trigger cash flows could be interest

rates, exchange rates, decisions about major projects, and so forth.

Once we specify these factors, we can then construct an appropriate engine to carry out our estimations. The choice of engine would depend on the types of cash flow risks we have to deal with. For instance, if we had fairly uncomplicated cash flows we might use an historical simulation or variance-covariance approach, or some specially designed term-structure model; however, since some cash flows are likely to be dependent on other factors such as discrete random variables (e.g., such as downgrades or defaults), it might not be easy tweaking such methods to estimate LaRs with sufficient accuracy. In such circumstances, it might be better to resort to simulation methods, which are much better suited to handling discrete variables and the potential complexities of cash flows in larger firms.

Another alternative is to use scenario analysis. We can specify liquidity scenarios, such as those arising from large changes in interest rates, default by counterparties, the redemption of puttable debt, calls for collateral on repos and derivatives, margin calls on swaps or futures positions, and so forth. We would then (as best we could) work through the likely/possible ramifications of each scenario, and so get an idea of the liquidity consequences associated with each scenario. Such exercises can be very useful, but, as with all scenario analyses, they might give us an indication of what could happen if the scenario occurs, but don't as such tell us anything about the probabilities associated with those scenarios or the LaR itself.

ESTIMATING LIQUIDITY IN CRISES

We now consider liquidity in crisis situations. As we all know, financial markets occasionally experience major crises—these include, for example, the stock market crash of 1987, the ERM crisis of 1992, the Russian default crisis of the

summer of 1998, and, of course, the many liquidity problems experienced since the onset of the financial crisis in August 2007. Typically, some event occurs that leads to a large price fall. This event triggers a huge number of sell orders, traders become reluctant to buy, and the bid-ask spread rises dramatically. At the same time, the flood of sell orders can overwhelm the market and drastically slow down the time it takes to get orders executed. Selling orders that would take minutes to execute in normal times instead take hours, and the prices eventually obtained are often much lower than sellers had anticipated. Market liquidity dries up, and does so at the very time market operators need it most. Assumptions about the market—and in particular, about market liquidity—that hold in “normal” market conditions can thus break down when markets experience crises. This means that estimating crisis liquidity is more than just a process of extrapolation from LaR under more normal market conditions: We need to estimate crisis-liquidity risks using methods that take into account the distinctive features of a crisis—large losses, high bid-ask spreads, and so forth.

One way to carry out such an exercise is by applying “crashmetrics” (Wilmott, 2000, Chapter 58). To give a simple example, we might have a position in a single derivatives instrument, and the profit/loss Π on this instrument is given by a delta-gamma approximation:

$$\Pi = \delta \Delta S + \frac{\gamma}{2} (\Delta S)^2 \quad (15)$$

where ΔS is the change in the stock price, and so forth. The maximum loss occurs when $dS = -\delta/\gamma$ and is equal to:

$$L^{\max} = -\Pi^{\min} = \frac{\delta^2}{2\gamma} \quad (16)$$

The worst-case cash outflow is therefore $m\delta^2/(2\gamma)$, where m is the margin or collateral requirement. This approach can also be extended to handle the other Greek parameters (the vegas, thetas, rhos, etc.), multi-

option portfolios, counterparty risk, and so forth. The basic idea—of identifying worst-case outcomes and then evaluating their liquidity consequences—can also be implemented in other ways as well. For example, we might identify the worst-case outcome as the expected outcome at a chosen confidence level, and we could estimate this (e.g., using extreme-value methods) as the ES at that confidence level. The cash outflow would then be m times this ES.

There are also other ways we can estimate crisis-LaR. Instead of focusing only on the high losses associated with crises, we can also take account of the high-bid ask spreads and/or the high bid-ask spread risks associated with crises. We can do so, for example, by estimating these spreads (or spread risks), and inputting these estimates into the relevant liquidity-adjusted VaR models discussed earlier.

However, these suggestions (i.e., Greek- and ES-based) are still rather simplistic, and with complicated risk factors—such as often arise with credit-related risks—we might want a more sophisticated model that was able to take account of the complications involved, such as:

- The discreteness of credit events.
- The interdependency of credit events.
- The interaction of credit and market risk factors (e.g., the ways in which credit events depend, in part, on market risk factors).
- Complications arising from the use of credit-enhancement methods such as netting arrangements, periodic settlement, credit derivatives, credit guarantees, and credit triggers (see, e.g., Wakeman, 1998).

These complicating factors are best handled using simulation methods tailor-made for the problems concerned.

The obvious alternative to probabilistic approaches to the estimation of crisis-liquidity is to use crisis-scenario analyses. We would imagine a big liquidity event—a major market crash, the default of a major financial institution or government, the outbreak of a war, or whatever—and work through the ramifications

for the liquidity of the institution concerned. One attraction of scenario analysis in this context is that we can work through scenarios in as much detail as we wish, and so take proper account of complicated interactions such as those mentioned in the last paragraph. This is harder to do using probabilistic approaches, which are by definition unable to focus on any specific scenarios. However, as with all scenario analyses, the results of these exercises are highly subjective, and the value of the results is critically dependent on the quality of the assumptions made.

KEY POINTS

- Liquidity refers to the ability to execute a trade or liquidate a position with little or no cost or inconvenience.
- Liquidity is a function of the market and depends on the type of position traded and sometimes the size and trading strategy of an individual trader.
- Liquidity risks are those associated with the prospect of imperfect or imperfect market liquidity, and can relate to risk of loss or risk to cash flows.
- There are two main aspects to liquidity risk measurement: the measurement of liquidity-adjusted measures of market risk (e.g., liquidity-adjusted value-at-risk, LVaR) and the measurement of liquidity risks per se (e.g., liquidity-at-risk, LaR).
- There are a number of easily implementable and often complementary approaches to the estimation of liquidity-adjusted measures of market risk: the *constant spread*, *exogenous spread*, and *endogenous price approaches*, and the liquidity discount approach.

- These approaches can produce risk estimates that differ substantially from the risk estimates obtained if liquidity is ignored.
- There are a number of approaches to the estimation of liquidity risks in noncrisis situations. These include both LaR approaches and scenario analyses.
- The LaR can be much greater than the VaR or much less than the VaR, depending on the circumstances.
- Crisis-related liquidity risks can be estimated using “crashmetrics” or scenario analyses hypothesized on crisis events such as a dry-up in market liquidity.

REFERENCES

- Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance* 9 (November): 203–228.
- Bangia, A., Diebold, F., Schuermann, T., and Stroughair, J. (1999). Liquidity on the outside. *Risk* 12 (June): 68–73.
- Dowd, K. (2005). *Measuring Market Risk*, 2nd ed. Chichester and New York: John Wiley.
- Giot, P., and Grammig, J. (2003). How large is liquidity risk in an automated auction market? *Mimeo*. University of Namur and University of Tübingen.
- Jarrow, R. A., and Subramanian, A. (1997). Mopping up liquidity. *Risk* 10 (December): 170–173.
- Singer, R. (1997). To VaR, a sister. *Risk* 10 (August): 86–87.
- Wakeman, L. (1998). Credit enhancement. Pp. 255–275 in C. Alexander (ed.), *Risk Analysis and Management. Volume 1: Measuring and Modelling Financial Risk*. Chichester and New York: John Wiley and Sons.
- Wilmott, P. (2000). *Paul Wilmott on Quantitative Finance. Volumes 1 and 2*. Chichester and New York: John Wiley and Sons.

Estimate of Downside Risk with Fat-Tailed and Skewed Models

JAMES X. XIONG, PhD, CFA

Senior Research Consultant, Ibbotson Associates, A Morningstar Company

Abstract: Asset returns are often not normally distributed and exhibit several stylized empirical facts: fat tails, skewness, finite variance, time scaling, and volatility clustering. Modeling the tail distribution of asset returns plays an essential role in downside risk management. The “left tail” of the distribution is where market crashes or crises occur. Downside risk can be measured in terms of conditional value-at-risk and estimated by fat-tailed and skewed models such as Lévy stable, truncated Lévy flight, skewed Student’s t , mixture of normal distributions, and GARCH models. These fat-tailed and skewed models have different characteristics in describing the tail distribution of asset returns. The objective is to select appropriate ones that can accurately model the downside risk.

The financial crisis of 2008 has led many practitioners and academics to reassess the adequacy of the return distribution models, in particular, the left tail. This entry focuses on modeling the left fat tails since they reflect market crashes or crises and play an essential role in downside risk management.

The most common model of asset returns is assumed to be normally or Gaussian distributed (see Bachelier, 1900). In other words, the returns follow a random walk or Brownian motion. This model is natural if one assumes the return over a time interval to be the result of many small independent shocks, which leads to a Gaussian distribution by the central limit theorem. However, empirical studies have observed that the return distributions

are more leptokurtic or fat-tailed than Gaussian distributions.

A normal distribution model assumes that an asset return that is three standard deviations below its arithmetic mean (popularly referred to as a “three-sigma event”) has a probability of only approximately 0.13%; that is, once every 1,000 times. For example, from January 1926 to March 2010, the S&P 500 total return index had a monthly mean return of 0.93% and a monthly standard deviation of 5.54%. A negative three-sigma event would be a return lower than -15.69% . During this time period of 1,010 months, there were 10 monthly returns worse than -15.69% as shown in Table 1 (the three-sigma event), with the most recent loss of -16.79% in October 2008 being ranked at ninth.

Table 1 The Worst 10 Monthly Returns for the S&P 500 (from 1/1926 to 3/2010)

	S&P 500 (%)
Sep 1931	-29.73
Mar 1938	-24.87
May 1940	-22.89
May 1932	-21.96
Oct 1987	-21.52
Apr 1932	-19.97
Oct 1929	-19.73
Feb 1933	-17.72
Oct 2008	-16.79
Jun 1930	-16.25

Source: Morningstar Encorr.

This implies the probability of a three-sigma event is about 1% rather than 0.13%, or eight times greater than we would expect under a normal distribution. Hence, a normal distribution fails to describe the “fat” or “heavy” tails of the stock market.

Many statistical models have been put forth to account for the heavy tails. We discuss several standard and popular fat-tailed models, such as Mandelbrot’s *Lévy stable* hypothesis (see Mandelbrot, 1963), the *Student’s t*-distribution (see Blattberg and Gonedes, 1974), the *mixture of normal distributions* (see Clark, 1973), and *GARCH* (see Bollerslev, 1986) models. There are many other fat-tailed candidates, and this entry does not aim at being exhaustive. Instead, we select representative models and illustrate them through examples so that practitioners may have some intuition about these practically implementable models.

Along the way, we introduce a relatively new fat-tailed and skewed model: the *truncated Lévy flight* (TLF). Another name for the TLF is the tempered stable distribution. The TLF model has a few interesting properties that we will illustrate later, such as possessing *fat tails*, *skewness*, finite moments, and time scaling. Of course, these quantitative models are not the only tool, and they need to be integrated with judgmental analyses and other estimates, but

they represent a good starting point for the management of *downside risk*.

DOWNSIDE RISK MEASURE

Before we dive into the discussions of fat-tailed models, we need to specify an appropriate downside risk measure. A popular downside risk measure is value-at-risk (VaR), which is an estimate of the loss that we expect to be exceeded with a given level of probability (e.g., 5%) over a specified time period. VaR has been recommended as a way of measuring risk by regulators and various financial industry advisory committees.

Conditional value-at-risk (CVaR), a closely related measure to VaR, is derived by taking a weighted average between the VaR and losses exceeding the VaR. Other terms for CVaR include mean shortfall, tail VaR, and expected tail loss. Studies such as Rockafellar and Uryasev (2000), for example, have shown that CVaR has more attractive properties than VaR. Specifically, CVaR is a coherent measure of risk as proved by Pflug (2000) in the sense of Artzner et al. (1999). One of the coherent measures is subadditivity; that is, the risk of a combination of investments is at most as large as the sum of the individual risks. VaR is not always subadditive, which means that the VaR of a portfolio with two instruments may be greater than the sum of individual VaRs of these two instruments. In contrast, CVaR is subadditive. Therefore, CVaR is a more appropriate measure of downside risk.

LÉVY STABLE DISTRIBUTION

Lévy distributions are stable; that is, the sum of two independent random variables, characterized by the same Lévy distribution of tail index α , is itself characterized by a Lévy distribution of the same index. In other words, the functional form of the distribution is maintained, if

we sum up independent, identically distributed Lévy stable random variables. The characteristic function of the Lévy stable distribution is (Lévy, 1925):

$$\begin{aligned}\ln \varphi(q) &= i\delta q - \gamma |q|^\alpha \left[1 - i\beta \frac{q}{|q|} \tan\left(\frac{\pi}{2}\alpha\right) \right] \quad \text{for } \alpha \neq 1 \\ &= i\delta q - \gamma |q| \left[1 - i\beta \frac{q}{|q|} \frac{2}{\pi} \ln |q| \right] \quad \text{for } \alpha = 1\end{aligned}$$

The probability density function is obtained by performing the inverse Fourier transform on the characteristic function. The four parameters associated with the Lévy stable distribution are: α determines the tail weight or the distribution's kurtosis with $0 < \alpha \leq 2$; β determines the distribution's skewness; γ is a scale parameter; and δ is a location parameter. One can generate univariate stable distributed returns through a numerical software package, for example, written by John Nolan (2009).¹ (In his software, the function "stablernd()" takes four parameters, α , β , γ , and δ , and generates random returns that follow a Lévy stable distribution. For empirical analyses, these four parameters can be estimated by the software's function "stablefit().")

In 1963, Mandelbrot modeled cotton prices with a Lévy stable process (Mandelbrot, 1963). Mandelbrot observed that in addition to being fat-tailed, the returns show another interest-

ing property: time scaling. This means that the distributions of returns have similar functional forms for different time intervals, ranging from one day to one month. The time scaling property is very appealing as it allows the sum of two independent Lévy stable distributed variables to be stable distributed, with the same stability index α . The normal distribution is a special case of the Lévy stable distribution, and it is scaled in the same way that the sum of two normally distributed variables is also normally distributed.

Figure 1 shows the time scaling of the S&P 500 index returns at time intervals of 1, 2, 3, and 5 days. The scaling variable for a Lévy stable process of index α is $\tilde{Z} = \frac{Z}{(\Delta t)^{1/\alpha}}$. The best fit gives $\alpha = 1.5$, and a good data collapse can be observed in Figure 1.

Mandelbrot's finding was later supported by Fama's study on stocks (Fama, 1965). A Lévy stable distribution model has fat tails and obeys scaling properties, but it has an infinite variance, which conflicts with empirical observations that the return variance is finite. For example, extensive analyses on high-frequency data (ranging from 1 minute to 1 day) for the 1,000 largest companies provided evidence that the returns have finite variance (Gopikrishnan et al., 1998). Infinite variance complicates the

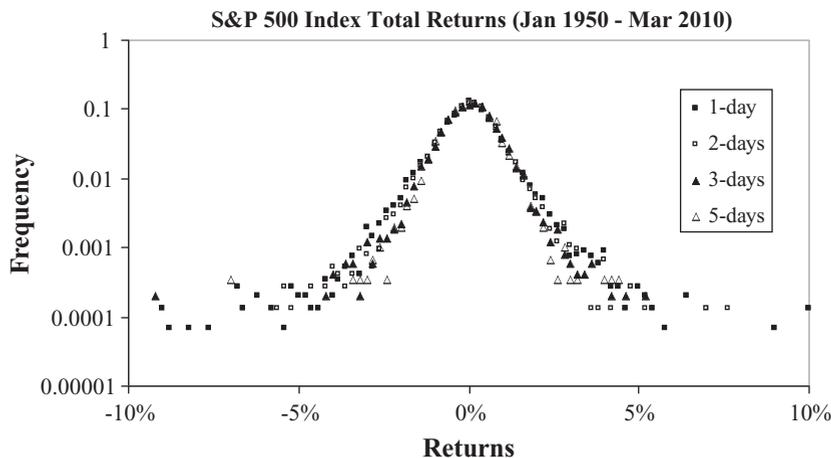


Figure 1 The Time Scaling of the S&P 500 Index with a Stability Index $\alpha = 1.5$

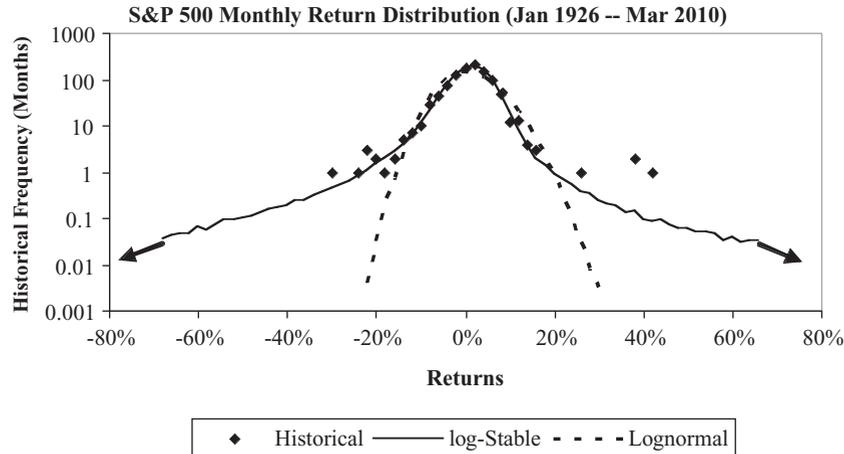


Figure 2 The Distributions of S&P 500 Monthly Returns Fitted by the Log-Stable and Lognormal Models

task of risk estimation, as well as the application of mean-variance portfolio construction.

Figure 2 illustrates the log-stable and log-normal distributions in fitting the distribution of monthly S&P 500 returns (also see Martin, Rachev, and Siboulet, 2003). Log-stable distribution applies the stable distribution to log-returns. The vertical axis of Figure 2 is in log scale with a base of 10, and this helps to view the tails of the distribution more clearly. It is clear that the lognormal distribution fails to fit the return distribution below -15% (the above-mentioned three-sigma events). The log-stable distribution fits the tail well, but it extends far beyond the historical maximum loss or gain with nonnegligible probabilities, which eventually results in an infinite variance. In other words, the tail for the log-stable distribution is perhaps too fat.

The infinite variance associated with the stable distribution induces a challenging problem in risk estimation. In practice, what is needed is a model with a distribution falling between the normal and stable distributions so that its tail is appropriately fat, but finite. By truncating the extreme tails of the stable distribution, a model named the *truncated Lévy flight* has such properties.

Truncated Lévy Flight

The TLF model was first introduced by Mantegna and Stanley (1994) in the physics literature, and it has drawn widespread attention since then. Koponen (1995) modified it in such a way as to allow an analytical calculation of the characteristic function and determination of the complete probability density distribution. Another name for the TLF is the tempered stable distribution—introduced and extended by Boyarchenko and Levendorskii (2000), Carr et al. (2002), Rosinski (2007), and Kim et al. (2008, 2010). Another application is the so-called smoothly truncated stable distribution introduced by Menn and Rachev (2009).

In this entry, we focus on the simplest TLF model by Mantegna and Stanley (1994). The probability density function (PDF) of a simple TLF process is defined as:

$$\begin{aligned} P(x) &= 0, & x < -l; \\ P(x) &= P_{Levy}(x), & -l \leq x \leq l; \\ P(x) &= 0, & x > l \end{aligned}$$

where $P_{Levy}(x)$ is the PDF of return x for a Lévy stable distribution and l is the cutoff length for the truncation. It can be seen that the truncation is abrupt. Alternative TLF models are similar

Table 2 Parameter Estimates with the Log-TLF Model for Monthly S&P 500, Weekly MSCI EM, and Weekly MSCI EAFE Returns

Log-TLF	α	β	γ	δ	Cutoff Length
S&P 500 Monthly	1.42	-0.12	0.024	0.010	6.8
MSCI EM Weekly	1.58	-0.40	0.015	0.0054	8.0
MSCI EAFE Weekly	1.79	-0.52	0.014	0.0033	10.0

and have in general smoother truncations in the form of exponential tails.

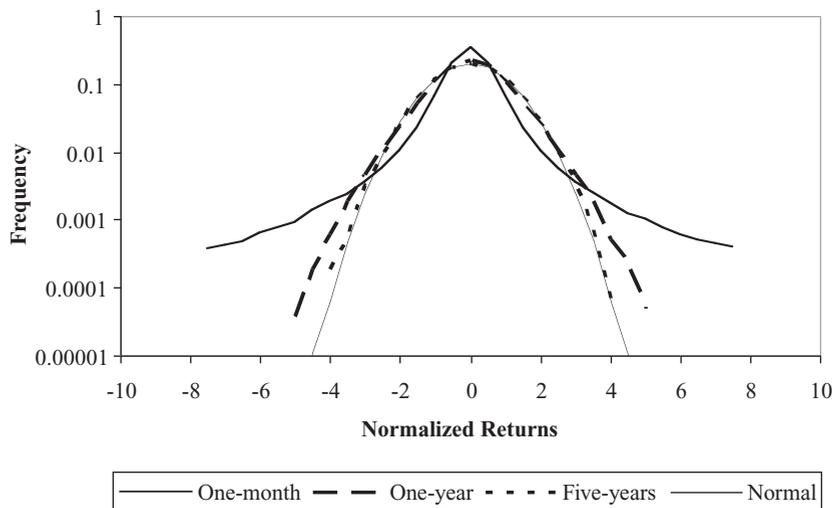
To simulate a TLF process from a Lévy stable process, we apply a truncation method on the Lévy stable distributed returns generated in the previous section so that the return series follows a TLF model. These truncated returns are then used in the distribution analyses and CVaR estimates, as well as the Monte Carlo simulations.

The truncation is simply implemented, for example, by truncating returns that are beyond 8-sigma for the MSCI Emerging Market index weekly returns or 6.8-sigma for S&P 500 monthly returns. The estimates of the five parameters are shown in Table 2. In the table, we choose the cutoff length in such a way that it is slightly larger than the historical maximum loss (in terms of standard deviation) over the entire historical period. The cutoff length shown in the table is normalized. One can think of a normalized cutoff length of 6 as a six-sigma event.

The other four parameters are estimated by the maximum likelihood method.

An interesting feature of the TLF model is its time scaling behavior. Mantegna and Stanley (1994, 1999) show that for a small time interval (e.g., a minute), the TLF distribution approximates a Lévy stable distribution with Lévy stable scaling; while for a significantly large but finite time interval (e.g., a year), the TLF distribution slowly converges to a Gaussian distribution. In other words, the TLF undergoes a crossover from a Lévy stable distribution to a Gaussian distribution as the time interval increases. This crossover is consistent with an independent empirical study of the distribution of daily, weekly and monthly returns for which a progressive convergence to a Gaussian process is deemed to be observed (Akgiray and Booth, 1988).

Figure 3 shows the convergence of the TLF from the Lévy stable distribution at a small time

**Figure 3** Time Scaling of the TLF process

interval to the Gaussian distribution at a large time interval. It shows that as the time interval increases from one month to one year and finally to five years, the normalized return distribution converges from the approximate Lévy stable distribution (one-month interval) to the normal distribution (five-year interval).

The truncation is able to mathematically solve the infinite variance problem inherent in the stable distribution. In fact, the truncation leads to the advantage that all four moments are finite. An interesting question is whether there are economic rationales for the truncation, even though the empirical evidence of finite variance is convincing. The truncation implies an upside or downside boundary for the returns. For the left tail, it is easy to see that the return is bounded by -100% due to limited liability for shareholders for unleveraged indexes or portfolios. However, the existence of the boundary for the upside tail is debatable and it may require extensive separate research. Factors that can limit an infinite positive gain for a large market index such as the S&P 500 may include competitive industries, business cycles, government intervention such as antitrust law and increasing interest rates, contrarian strategies that lead to mean reversion of returns, and so on. Fundamental "intrinsic valuation" indicates that the asset prices should be commensurate with the overall economic growth, which is limited by population growth, labor resources, productivity, and so on.

On the drawback side, like the normal or Lévy stable distribution model, the TLF model assumes an independent and identically distributed process and therefore it cannot describe the time-dependent volatility or volatility clustering observed in market data. Volatility clustering means that a period of high volatility tends to be followed by high volatility and a period of low volatility is likely followed by low volatility.

An attempt to address this drawback is to assume TLF innovations instead of Gaussian innovations in GARCH models. A few stud-

ies have investigated the option pricing problem with GARCH dynamics and non-Gaussian innovations. For example, Menn and Rachev (2009) considered smoothly truncated stable innovations in order to provide a practical framework to extend option pricing theory to the Lévy stable model. Kim et al. (2010) studied parametric models based on tempered stable innovations, and they showed that the GARCH model with tempered stable innovations explains both asset price behavior and European option prices better than the normal GARCH model.

STUDENT'S *t*-DISTRIBUTION

The Student's *t*-distribution is well documented in the literature. Its probability density function is given by:

$$P(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

where ν is the degrees of freedom. The Student's *t*-distribution coincides with the Cauchy distribution for $\nu = 1$, and approaches Gaussian for $\nu \rightarrow \infty$. Finite variance only exists for $\nu > 2$.

Blattberg and Gonedes (1974) proposed that the returns are distributed with a Student's *t*-distribution. Markowitz and Usmen (1996) found that the daily log-return data of the S&P 500 index can be fitted by the Student's *t*-distribution with about 4.5 degrees of freedom. Hurst and Platen (1997) reached a similar conclusion. Platen and Sidorowicz (2007) investigated the log-returns of a variety of diversified world stock indexes in different currency denominations by applying the maximum likelihood ratio test to the large class of generalized hyperbolic distributions, and showed that the Student's *t*-distribution with about four degrees of freedom was the best fit among the models they tested.

The Student's t -distribution is symmetric, thus it cannot model skewness. In order to model negative skewness, Hansen (1994) introduced the skewed Student's t -distribution, which is able to model skewness, but it requires one more parameter to be estimated.

The Student's t -distribution has fat tails but does not obey time scaling, which indicates that the sum of two independent Student's t -distributed variables is not a Student's t -variable with the same degrees of freedom. It cannot model volatility clustering.

The kurtosis of the Student's t distribution is given by $\frac{6}{4-\nu}$, and it is only defined for $\nu > 4$. In other words, the kurtosis is infinite when ν is less than or equal to 4, and the skewness tends to be unstable for $\nu \leq 4$. In order to avoid an infinite kurtosis, we set the minimum ν as 4.1 when the maximum likelihood estimate gives a value of ν less than 4 (shown as MLE- ν in Table 3). Our numerical simulations show that the CVaR estimate is not sensitive to this small change of ν .

For the symmetric Student's t -distribution, ν is the only parameter that needs to be estimated for normalized returns. For the skewed Student's t -distribution, we need to add a parameter, λ , to capture the skewness (see Hansen, 1994). These estimated parameters are shown in Table 3.

Table 3 Parameter Estimates with the Log Student's t and Log Skewed Student's t Distributions for Monthly S&P 500, Weekly MSCI EM, and Weekly MSCI EAFE Returns

Log Student's t		
	ν	MLE- ν
S&P 500 Monthly	4.1	3.6
MSCI EM Weekly	4.1	4.0
MSCI EAFE Weekly	4.4	4.4
Log Skewed t		
	ν	λ
S&P 500 Monthly	4.1	-0.13
MSCI EM Weekly	4.1	-0.25
MSCI EAFE Weekly	4.4	-0.09

MIXTURE OF NORMAL DISTRIBUTIONS

In the mixture of normal distributions model, the fat tails are obtained through subordination. The model considered for the log-returns is:

$$d \log S(t) = \mu dt + \sigma g(t) dW$$

where μ and σ are associated with the normal process of an individual trade. W is a standard Brownian motion. This model becomes the standard geometric Brownian motion when $g(t)$ is constant. $g(t)$ is a subordinator and positive increasing random process that characterizes the market trading activity time.

If $g(t)$ is assumed to be lognormally distributed with mean μ_s and standard deviation σ_s , this mixture process is also referred to as the normal-lognormal mixture. The probability density function for the normal-lognormal mixture is given in Clark (1973).

Other kinds of mixtures exist in the literature, such as a normal-gamma mixture, also referred to as a variance gamma process (Madan and Seneta, 1990). In this entry, we only illustrate the normal-lognormal mixture, one of the simplest mixture models. The estimated parameters for the normal-lognormal mixture are shown in Table 4.

The mixture of normal distributions utilizes the concept of a subordinated process. Clark (1973) assumes that trading volume is a plausible measure of the evolution of price dynamics. Indeed, a sizeable literature has demonstrated a strong positive contemporaneous correlation between trading volume and return volatility (see, for example, Andersen, 1996). More specifically, the distribution of log-returns occurring

Table 4 Parameter Estimates with the Mixture Distribution for Monthly S&P 500, Weekly MSCI EM, and Weekly MSCI EAFE Returns

	μ	σ	μ_s	σ_s
S&P 500 Monthly	0.0075	0.0382	0.0006	1.193
MSCI EM Weekly	0.0019	0.0206	0.0002	1.241
MSCI EAFE Weekly	0.0013	0.0152	0.0003	1.280

from a given level of trading volume is subordinate to the distribution of an individual trade and directed by the distribution of the trading volume. By assuming the normal distribution for the individual trade and finite moments for the distribution of the trading volume, Clark (1973) proves that the mixed distribution has fat tails with all moments finite.

The mixture of normal distributions is intuitively appealing because it is directly linked to market microstructure such as information flow, trading volume, and number of transactions. The subordinated process premise has also evolved into stochastic volatility that now receives vigorous attention in the finance literature (see Andersen, 1996). In general, mixture of normal distributions has fat tails but does not obey time scaling. A generalized mixture of normal distributions, however, can describe volatility clustering.

GARCH Models

General autoregressive conditional heteroscedasticity (GARCH) models, first introduced by Bollerslev (1986), are now widely employed in financial time-series analyses. In particular, they are used to predict short horizon volatilities (ranging from one day to one month).

The return generating process is based on geometric Brownian motion but with the variance being a time-dependent GARCH(1,1) process, which is defined by the relation:

$$\sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

where α_0 , α_1 , and β_1 are the control parameters of the GARCH(1,1) stochastic process. r_t is a random variable with zero mean and variance σ_t^2 , and is characterized by a conditional probability density function $f_t(x)$, which is arbitrary but is often chosen to be Gaussian. In this entry, the innovation σ_t^2 is assumed to be Gaussian. These three control parameters are estimated by the maximum likelihood method and shown in Table 5.

Table 5 Parameter Estimates with the GARCH(1,1) Model for Monthly S&P 500, Weekly MSCI EM, and Weekly MSCI EAFE Returns

	α_0	α_1	β_1
S&P 500 Monthly	0.00006	0.1291	0.8474
MSCI EM Weekly	0.00002	0.1431	0.8309
MSCI EAFE Weekly	0.00002	0.0897	0.8815

GARCH models assume that volatility changes with time and with past information. Because of the time-dependent volatility, the unconditional distribution of returns exhibit fat tails. GARCH models allow for volatility clustering or autocorrelation in the volatility.

The most popular GARCH model is GARCH(1,1). The scaling properties of GARCH(1,1) are not clear from the theory; however, numerical simulations of GARCH(1,1) with Gaussian innovations show that it fails to describe the scaling properties of high-frequency data (see Mantegna and Stanley, 1999).

GARCH(1,1) processes are unconditionally stationary with finite variance if $1 - \alpha_1 - \beta_1 > 0$, and have finite kurtosis if $1 - \beta_1^2 - 2\alpha_1\beta_1 - 3\alpha_1^2 > 0$.

MODELING RETURN DISTRIBUTIONS FOR MAJOR INDEXES

Applications of the Lévy stable, Student's t , and mixture of normal distribution models in modeling market indexes are well documented (see, for example, Mandelbrot [1963], Clark [1973], Blattberg and Gonedes [1974], Markowitz and Usmen [1996], Hurst and Platen [1997], Martin, Rachev and Siboulet [2003], Platen and Sidorowicz [2007], etc.). The literature offered detailed methodology on how the model parameters are estimated. In some cases, they performed comparisons for these models.

Mantegna and Stanley (1999) studied the TLF model and GARCH(1,1) with Gaussian innovations processes. They found that the TLF model well describes the time scaling, while it is not

able to properly describe the volatility clustering. The GARCH(1,1) model seems to be complementary to the TLF: It is able to describe the volatility clustering, but it fails to describe the time scaling. As mentioned earlier, however, the GARCH model with TLF innovations might offer a better solution to the TLF model or GARCH with Gaussian innovations.

Many previous studies have focused on high-frequency data such as daily return data. Here, we are interested in weekly or monthly data because investors typically have a relatively long investment horizon and portfolios are often rebalanced monthly. We apply these fat-tailed models to some well-known weekly or monthly returns of equity indexes. Our test assets include the monthly S&P 500 total return index, the weekly MSCI Emerging Market index, and the weekly MSCI EAFE index. One reason to use weekly data is to have more data points in the tails given that the MSCI indexes

have relatively short histories. A few other equity and fixed income indexes, such as the MSCI UK, U.S. Long-Term Government Bond, Muni bonds, and some individual stocks were tested with the same methodologies and the results are similar, so they are not reported (e.g., Xiong, 2010).

We apply the maximum likelihood method to calibrate model parameters as previous studies did. The estimated parameters for the TLF, Student's t , normal-lognormal mixture, and GARCH(1,1) are shown in Tables 2, 3, 4, and 5, respectively. Since we are more interested in modeling downside risk, our goal is to fit the model's tail distribution to the empirical tail distribution in terms of CVaR through Monte Carlo simulations.

Table 6 focuses on nonstable distribution models and presents the empirical statistics as well as the Monte Carlo simulation results for the six models. The statistics for each model

Table 6 Statistics Summary for Historical Returns, as Well as Simulated Returns for Lognormal, Log-TLF, Log Student's t , Log Skewed Student's t , Normal-Lognormal Mixture, and GARCH(1,1) Models

S&P 500 Monthly					
	Mean	Std Dev	Skewness	Kurtosis	CVaR
Empirical	0.93%	5.54%	0.35	12.45	-12.20%
Lognormal	0.93%	5.54%	0.16	3.05	-9.96%
log-TLF	0.93%	5.54%	0.59	12.90	-12.20%
log-Student t	0.93%	5.54%	1.35	47.93	-10.91%
log-Skewed t	0.93%	5.54%	0.69	50.70	-11.91%
Mixture	0.93%	5.54%	1.02	18.85	-11.34%
GARCH(1,1)	0.93%	5.54%	0.46	9.50	-10.77%
MSCI EM Weekly					
Empirical	0.25%	3.04%	-0.52	8.38	-7.45%
Lognormal	0.25%	3.04%	0.09	3.02	-5.88%
log-TLF	0.25%	3.04%	-0.38	12.29	-7.45%
log-Student t	0.25%	3.04%	0.71	22.90	-6.49%
log-Skewed t	0.25%	3.04%	-0.81	14.23	-7.45%
Mixture	0.25%	3.04%	0.62	16.10	-6.76%
GARCH(1,1)	0.25%	3.04%	0.58	23.91	-6.47%
MSCI EAFE Weekly					
Empirical	0.16%	2.29%	-0.76	10.25	-5.27%
Lognormal	0.16%	2.29%	0.07	3.01	-4.47%
log-TLF	0.16%	2.29%	-0.47	9.25	-5.27%
log-Student t	0.16%	2.29%	0.46	16.71	-4.93%
log-Skewed t	0.16%	2.29%	-0.18	13.03	-5.27%
Mixture	0.16%	2.29%	0.52	16.71	-5.17%
GARCH(1,1)	0.16%	2.29%	0.10	4.20	-4.73%

are based on 1,000,000 simulated random returns that follow the corresponding distribution models. It can be seen that the lognormal model underestimates the *monthly* CVaR by 2.04% for the S&P 500, the *weekly* CVaR by 1.57% for the MSCI EM, and the *weekly* CVaR by 0.8% for the MSCI EAFE, respectively. The log Student's *t*-distribution, normal-lognormal mixture, and GARCH(1,1) have similar CVaR estimates, and all of them are better than the lognormal model but appear to underestimate the tail risk. On the other hand, both the log-TLF model and the log skewed Student's *t*-model provide a good fit for CVaR for all three indexes: S&P 500, MCSI EM, and MSCI EAFE.

Note that the log Student's *t*, normal-lognormal mixture, and GARCH(1,1) are positively skewed by design in a way similar to the lognormal distribution because we are working with the log-returns. The positive skewness resulted from taking the exponential function on the log-returns. None of these three models can account for negative skewness without modifications.

Therefore there are two reasons why the log-TLF and the log skewed Student's *t*-models do well in fitting the CVaR. First, their tails are appropriately fat, and second, both of them are able to capture negative skewness. For the TLF model, the fatness of the tail is controlled by α and the cutoff length and the skewness is controlled by β as shown in Table 2. For the skewed Student's *t*-distribution, the fatness of the tail is controlled by the degrees of freedom ν and the skewness is controlled by λ as shown in Table 3.

Figures 4, 5, and 6 compare the log-TLF model with other models in fitting the historical return distributions for monthly S&P 500 returns, weekly MSCI EM returns, and weekly MSCI EAFE returns, respectively. The figures confirm the results shown in Table 6. It can be seen that the log-TLF provides a good fit for the three indexes. The log skewed Student's *t* is almost as effective as the log-TLF model in fitting CVaRs. Compared to the log skewed Student's *t*-distribution, the log-TLF has a fatter but shorter tail because of the truncation. On the other hand, the normal-lognormal mixture

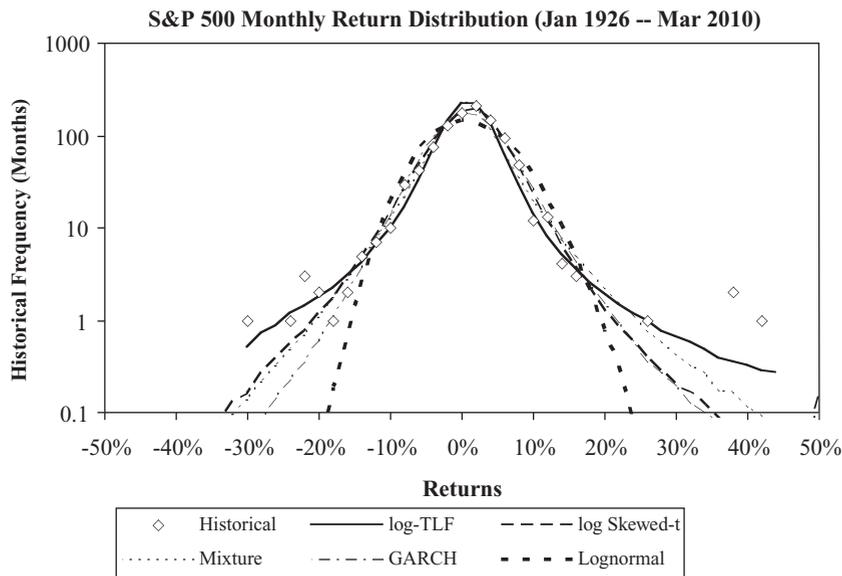


Figure 4 The Historical Distributions of S&P 500 Monthly Returns Fitted by the Log-TLF, Log Skewed Student's *t*, Normal-Lognormal Mixture, GARCH(1,1), and Lognormal Models

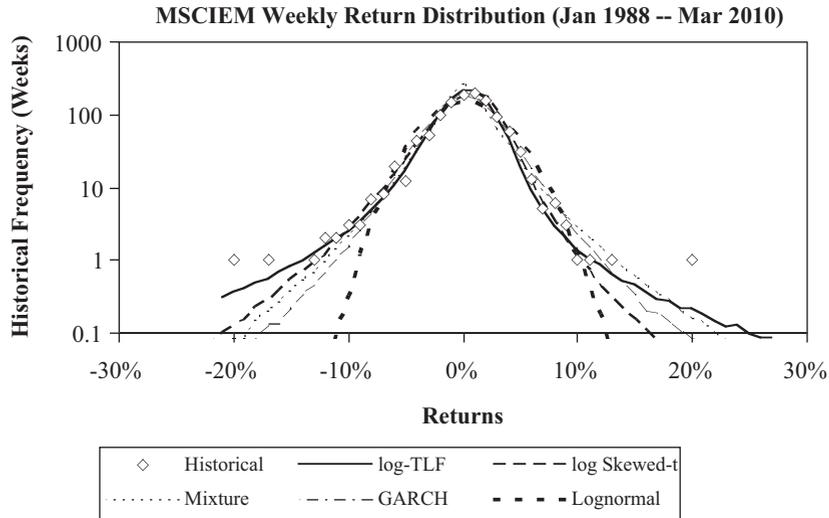


Figure 5 The Historical Distributions of MSCI EM Weekly Returns Fitted by the Log-TLF, Log Skewed Student's t , Normal-Lognormal Mixture, GARCH(1,1), and Lognormal Models

and GARCH(1,1) model have CVaRs that fall between those of the log-TLF and lognormal models. The finding for the log symmetric Student's t -distribution, not plotted due to space limitations, is similar to the normal-lognormal mixture and GARCH(1,1) model.

Table 7 summarizes the underestimated CVaRs for the six models that have been applied to the three indexes. The underestimated

tails are reported on a relative basis based on CVaR estimates shown in Table 6. For example, the lognormal model underestimates the monthly CVaR by a relative percentage of 18% ($= \frac{12.2 - 9.96}{12.2}$) for the S&P 500 index.

Averaging over the three indexes, the lognormal model underestimates the CVaR by about 18% on a relative basis. The normal-lognormal mixture, the log Student's t -distribution, and

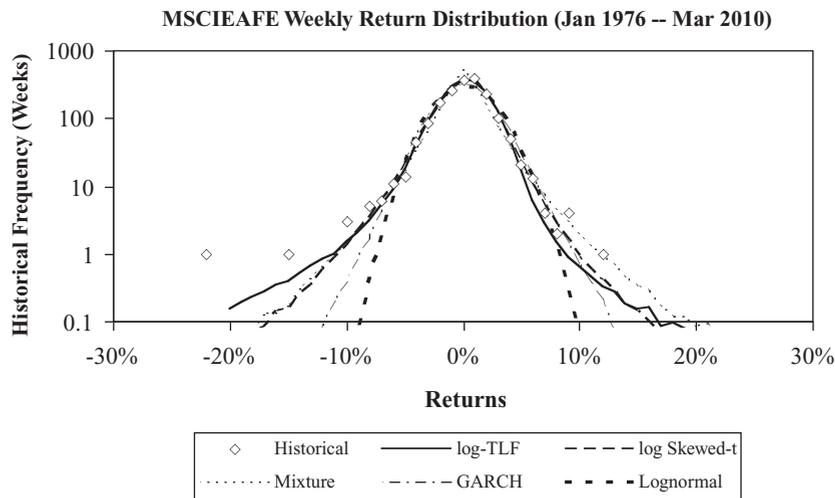


Figure 6 The Historical Distributions of MSCI EAFE Weekly Returns Fitted by the Log-TLF, Log Skewed Student's t , Normal-Lognormal Mixture, GARCH(1,1), and Lognormal Models

Table 7 Underestimated CVaRs in Relative Percentage for the Six Models

Index	S&P 500	MSCI EM	MSCI EAFE
Data Range	1926.1–2010.3	1988.1–2010.4	1976.1–2010.4
Number of Periods	1011 Monthly	1164 Weekly	1792 Weekly
Lognormal	18%	21%	15%
Log-TLF	0%	0%	0%
Log Student's t	11%	13%	6%
Log Student's Skewed t	2%	0%	0%
Normal-Lognormal Mixture	7%	9%	2%
GARCH(1,1)	12%	13%	10%

GARCH(1,1) with Gaussian innovations perform better than the lognormal model but appear to underestimate the CVaR by about 6%, 10%, and 12%, respectively. In contrast, both the log-TLF and log skewed t -distribution did a better job in modeling the CVaR.

KEY POINTS

- It is well known that asset returns often exhibit fat tails, negative skewness, time scaling, and volatility clustering. Fat-tailed and skewed models can be used to estimate the downside risk of assets. It is important that the selected models are able to capture fat tails and skewness, among others.
- The lognormal distribution is the fundamental assumption of many important financial models, but it has thin tails and thus can significantly underestimate the downside risk. On the other side, the Lévy stable distribution exhibits time scaling and fat tails, but it tends to overestimate the downside risk due to its infinite variance.
- The Student's t -distribution can model fat tails but not negative skewness. A modification results in the skewed Student's t -distribution, which can model both fat tails and negative skewness. However, both of them do not possess time scaling properties and cannot model volatility clustering.
- The normal-lognormal mixture is intuitive as it is directly linked to market microstructure such as information flow and trading volume. It has fat tails but cannot model negative

skewness. In general, it does not possess time scaling.

- The truncated Lévy flight model can describe the asymptotic return distributions measured at all frequencies and the scaling properties (self-similarities). More specifically, for a small time interval (e.g., a minute), this distribution approximates a Lévy stable distribution with Lévy stable scaling; while for a significantly large but finite time interval (e.g., a year), the truncated Lévy flight distribution slowly converges to a Gaussian distribution. It has finite four moments and can model both fat tails and negative skewness.
- The truncated Lévy flight or tempered stable distribution model cannot describe volatility clustering. In contrast, GARCH with Gaussian innovations can model volatility clustering but it is often found that the tail is not fat enough. Recent studies show that a GARCH with truncated Lévy flight innovations appears to be able to describe most of the stylized empirical facts: fat tails, skewness, and volatility clustering.

NOTE

1. For details, see <http://academic2.american.edu/~jpnolan/stable/stable.html/>

REFERENCES

- Akgiray, V., and Booth, G. G. (1988). The stable-law model of stock returns. *Journal of Business & Economic Statistics* 6: 51–57.
- Anderson, T. (1996). Return volatility and trading volume: An information flow interpretation

- of stochastic volatility. *Journal of Finance* 51, 169–204.
- Artzner, P., Delbaen, F., Eber, J. M., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance* 9, 3: 203–228.
- Bachelier, L. (1900). Théorie de la spéculation. Doctoral dissertation. *Annales Scientifiques de l'École Normale Supérieure* (ii) 17, 21–86. Trans. P. H. Cootner, ed. (1964). *The Random Character of Stock Market Prices*. Cambridge, MA: MIT Press.
- Blattberg, R. C., and Gonedes, N. J. (1974). A comparison of the stable and Student distributions as statistical models for stock prices. *Journal of Business* 47, 244–280.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 3: 307–327.
- Boyarchenko, S. I., and Levendorskii, S. Z. (2000). Option pricing for truncated Lévy processes. *International Journal of Theoretical and Applied Finance* 3.
- Carr, P., Geman, H., Madan, D., and Yor, M. (2002). The fine structure of asset returns: An empirical investigation. *The Journal of Business* 75, 2: 305–332.
- Clark, P. K. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica* 41, 135–155.
- Fama, E. F. (1965). The behavior of stock-market prices. *The Journal of Business* 38, 1: 34–105.
- Gopikrishnan, P., Meyer, M., Amaral, L.A.N., and Stanley, H. E. (1998). Inverse cubic law for the distribution of stock price variations. *The European Physical Journal B*, 3, 2: 139–140.
- Hansen, B. E. (1994). Autoregressive conditional density estimation. *International Economic Review* 35, 705–730.
- Hurst, S. R., and Platen, E. (1997). The Marginal Distributions of Returns and Volatility. Lecture Notes—Monograph Series. Vol. 31, *L₁-Statistical Procedures and Related Topics*, pp. 301–314. Institute of Mathematical Statistics.
- Kim, Y., Rachev, S., Bianchi, M., and Fabozzi, F. (2008). A new tempered stable distribution and its application to finance. In G. Bol, S. T. Rachev, and R. Würth, (Eds.), *Risk Assessment: Decisions in Banking and Finance*, pp. 51–84. Physika Verlag, Springer.
- Kim, Y., Rachev, S., Bianchi, M., and Fabozzi, F. (2010). Tempered stable and tempered infinitely divisible GARCH models. *Journal of Banking and Finance*.
- Koponen, I. (1995). Analytic approach to the problem of convergence of truncated Lévy flights towards the Gaussian stochastic process. *Physical Review E*, 52.
- Lévy, P. (1925). *Calcul des probabilités*. Paris: Gauthier-Villars.
- Madan, D. B., and Seneta E. (1990). The variance gamma (v.g.) model for share market returns. *Journal of Business* 63, 511–524.
- Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business* 36, 392–417.
- Mantegna, R. N., and Stanley, H. E. (1994). Stochastic process with ultraslow convergence to a Gaussian: The truncated Lévy flight. *Physical Review Letters* 73, 2946–2949.
- Mantegna, R. N., and Stanley, H. E. (1999). *An Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge: Cambridge University Press.
- Markowitz, H. M., and Usmen, N. (1996). The likelihood of various stock market return distributions, Part 2: Empirical results. *Journal of Risk and Uncertainty* 13, 3: 221–247.
- Martin, R. D., Rachev S., and Siboulet, F. (2003). Phi-alpha optimal portfolios & extreme risk management. *Wilmott Magazine of Finance* November: 70–83.
- Menn, C., and Rachev, S. (2009). Smoothly truncated stable distributions, GARCH-models, and option pricing. *Mathematical Methods of Operations Research* 63, 3: 411–438.
- Nolan, J. (2009). Software Stable 5.1 for MATLAB.
- Platen, E., and Sidorowicz, R. (2007). Empirical evidence on Student-*t* log-returns of diversified world stock indices. Research Paper 194. University of Technology, Sydney. School of Finance and Economics. Quantitative Finance Research Centre.
- Pflug, G. Ch. (2000). Some remarks on the value-at-risk and the conditional value-at-risk. In *Probabilistic Constrained Optimization: Methodology and Applications*, S. Uryasev, (Ed.). New York: Springer.
- Rockafellar, R. T., and Uryasev, S. (2000). Optimization of conditional value-at-risk. *The Journal of Risk* 2, 3: 21–41.
- Rosinski, J. (2007). Tempering stable processes. *Stochastic Processes and Their Applications* 117, 6: 677–707.
- Xiong, J. X. (2010). Using truncated Lévy flight to estimate downside risk. *Journal of Risk Management in Financial Institutions* 3, 3: 231–242.

Moving Average Models for Volatility and Correlation, and Covariance Matrices

CAROL ALEXANDER, PhD

Professor of Finance, University of Sussex

Abstract: The volatilities and correlations of the returns on a set of assets, risk factors, or interest rates are summarized in a covariance matrix. This matrix lies at the heart of risk and return analysis. It contains all the information necessary to estimate the volatility of a portfolio, to simulate correlated values for its risk factors, to diversify investments, and to obtain efficient portfolios that have the optimal trade-off between risk and return. Both risk managers and asset managers require covariance matrices that may include very many assets or risk factors. For instance, in a global risk management system of a large international bank all the major yield curves, equity indexes, foreign exchange rates, and commodity prices will be encompassed in one very large dimensional covariance matrix.

Variances and *covariances* are parameters of the joint distribution of asset (or risk factor) returns. It is important to understand that they are unobservable. They can only be estimated or forecast within the context of a model. Continuous-time models, used for option pricing, are often based on stochastic processes for the variance and covariance. Discrete-time models, used for measuring portfolio risk, are based on time series models for variance and covariance. In each case, we can only ever estimate or forecast variance and covariance within the context of an assumed model.

It must be emphasized that there is no absolute “true” variance or covariance. What is “true” depends only on the statistical model.

Even if we knew for certain that our model was a correct representation of the data generation process, we could never measure the true variance and covariance parameters exactly because pure variance and covariance are not traded in the market. An exception to this is the futures on *volatility* indexes such as the Chicago Board Options Exchange Volatility Index (VIX). Hence, some risk-neutral volatility is observed. However, this entry deals with covariance matrices in the physical measure.

Estimating a variance according to the formulas given by a model, using historical data, gives an observed variance that is “realized” by the process assumed in our model. But this “realized variance” is still only ever an

estimate. Sample estimates are always subject to sampling error, which means that their value depends on the sample data used.

In summary, different statistical models can give different estimates of variance and covariance for two reasons:

- A true variance (or covariance) is different between models. As a result, there is a considerable degree of model risk inherent in the construction of a covariance or *correlation matrix*. That is, very different results can be obtained using two different statistical models even when they are based on exactly the same data.
- The estimates of the true variances (and covariances) are subject to sampling error. That is, even when we use the same model to estimate a variance, our estimates will differ depending on the data used. Both changing the sample period and changing the frequency of the observations will affect the covariance matrix estimate.

This entry covers moving average discrete-time series models for variance and covariance, focusing on the practical implementation of the approach and providing an explanation for their advantages and limitations. Other statistical tools are described in Alexander (2008a, Chapter 9).

BASIC PROPERTIES OF COVARIANCE AND CORRELATION MATRICES

The covariance matrix is a square, symmetric matrix of variance and covariances of a set of m returns on assets, or on risk factors, given by:

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \dots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \dots & \dots & \sigma_{2m} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \dots & \sigma_{3m} \\ \dots & \dots & \dots & \dots & \dots \\ \sigma_{m1} & \dots & \dots & \dots & \sigma_m^2 \end{pmatrix} \quad (1)$$

Since

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \dots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \dots & \dots & \sigma_{2m} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \dots & \sigma_{3m} \\ \dots & \dots & \dots & \dots & \dots \\ \sigma_{m1} & \dots & \dots & \dots & \sigma_m^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \varrho_{12}\sigma_1\sigma_2 & \dots & \dots & \varrho_{1m}\sigma_1\sigma_m \\ \varrho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \dots & \dots & \varrho_{2m}\sigma_2\sigma_m \\ \varrho_{31}\sigma_3\sigma_1 & \varrho_{32}\sigma_3\sigma_2 & \sigma_3^2 & \dots & \varrho_{3m}\sigma_3\sigma_m \\ \dots & \dots & \dots & \dots & \dots \\ \varrho_{m1}\sigma_m\sigma_1 & \dots & \dots & \dots & \sigma_m^2 \end{pmatrix}$$

a covariance matrix can also be expressed as

$$\mathbf{V} = \mathbf{D}\mathbf{C}\mathbf{D} \quad (2)$$

where \mathbf{D} is a diagonal matrix with elements equal to the standard deviations of the returns and \mathbf{C} is the correlation matrix of the returns. That is:

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \dots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \dots & \dots & \sigma_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \sigma_{1m} & \sigma_{2m} & \dots & \dots & \sigma_m^2 \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 & \dots & \dots & 0 \\ 0 & \sigma_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & 0 & \sigma_m \end{pmatrix} \times \begin{pmatrix} 1 & \varrho_{12} & \dots & \dots & \varrho_{1m} \\ \varrho_{12} & 1 & \dots & \dots & \varrho_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \varrho_{1m} & \varrho_{2m} & \dots & \dots & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & \dots & \dots & 0 \\ 0 & \sigma_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & 0 & \sigma_m \end{pmatrix}$$

Hence, the covariance matrix is simply a mathematically convenient way to express the asset volatilities and their *correlations*.

To illustrate how to estimate an annual covariance matrix and a 10-day covariance matrix, assume three assets that have the following volatilities and correlations:

Asset 1 volatility	20%	Asset 1–Asset 2 correlation	0.8
Asset 2 volatility	10%	Asset 1–Asset 3 correlation	0.5
Asset 3 volatility	15%	Asset 3–Asset 2 correlation	0.3

Then,

$$\mathbf{D} = \begin{pmatrix} 0.2 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.15 \end{pmatrix} \quad \mathbf{C} = \begin{pmatrix} 1 & 0.8 & 0.5 \\ 0.8 & 1 & 0.3 \\ 0.5 & 0.3 & 1 \end{pmatrix}$$

So the annual covariance matrix \mathbf{DCD} is:

$$\begin{pmatrix} 0.2 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.15 \end{pmatrix} \begin{pmatrix} 1 & 0.8 & 0.5 \\ 0.8 & 1 & 0.3 \\ 0.5 & 0.3 & 1 \end{pmatrix} \begin{pmatrix} 0.2 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 0.15 \end{pmatrix} \\ = \begin{pmatrix} 0.04 & 0.016 & 0.015 \\ 0.016 & 0.01 & 0.0045 \\ 0.015 & 0.0045 & 0.0225 \end{pmatrix}$$

To find a 10-day covariance matrix in this simple case, one is forced to assume the returns are independent and identically distributed in order to use the square root of time rule: that is, that the h -day covariance matrix is h times the 1 day covariance matrix. Put another way, the 10-day covariance matrix is obtained from the annual matrix by dividing each element by 25, assuming there are 250 trading days per year.

Alternatively, we can obtain the 10-day matrix using the 10-day volatilities in \mathbf{D} . Note that under the independent and identically distributed returns assumption \mathbf{C} should not be affected by the holding period. That is,

$$\mathbf{D} = \begin{pmatrix} 0.04 & 0 & 0 \\ 0 & 0.02 & 0 \\ 0 & 0 & 0.03 \end{pmatrix} \mathbf{C} = \begin{pmatrix} 1 & 0.8 & 0.5 \\ 0.8 & 1 & 0.3 \\ 0.5 & 0.3 & 1 \end{pmatrix}$$

because each volatility is divided by 5 (that is, the square root of 25). Then we get the same result as above, that is

$$\begin{pmatrix} 0.04 & 0 & 0 \\ 0 & 0.02 & 0 \\ 0 & 0 & 0.03 \end{pmatrix} \begin{pmatrix} 1 & 0.8 & 0.5 \\ 0.8 & 1 & 0.3 \\ 0.5 & 0.3 & 1 \end{pmatrix} \\ \times \begin{pmatrix} 0.04 & 0 & 0 \\ 0 & 0.02 & 0 \\ 0 & 0 & 0.03 \end{pmatrix} = \begin{pmatrix} 0.16 & 0.064 & 0.06 \\ 0.064 & 0.04 & 0.018 \\ 0.06 & 0.018 & 0.09 \end{pmatrix} \\ \times 10^{-2}$$

Note that \mathbf{V} is positive semidefinite if and only if \mathbf{C} is positive semidefinite. \mathbf{D} is always positive definite. Hence, the positive semidefiniteness of \mathbf{V} only depends on the way we construct the correlation matrix. It is quite a challenge to generate meaningful, positive semidefinite correlation matrices that are large enough for managers to be able to net the risks across all positions in a firm. Simplifying assumptions are necessary. For example, *RiskMetrics* (1996) uses a very simple methodology based on moving

averages in order to estimate extremely large positive definite matrices covering hundreds of risk factors for global financial markets. (This is discussed further below.)

EQUALLY WEIGHTED AVERAGES

This section describes how volatility and correlation are estimated and forecast by applying equal weights to certain historical time series data. We outline a number of pitfalls and limitations of this approach and as a result recommend that these models be used as an indication of the possible range for long-term volatility and correlation. As we shall see, these models are of dubious validity for short-term volatility and correlation forecasting.

In the following, for simplicity, we assume that the mean return is zero and that returns are measured at the daily frequency, unless specifically stated otherwise. A zero mean return is a standard assumption for risk assessments based on time series of daily data, but if returns are measured over longer intervals it may not be very realistic. Then the equally weighted estimate of the variance of returns is the average of the squared returns and the corresponding volatility estimate is the square root of this expressed as an annual percentage. The equally weighted estimate of the covariance of two returns is the average of the cross products of returns and the equally weighted estimate of their correlation is the ratio of the covariance to the square root of the product of the two variances.

Equal weighting of historical data was the first widely accepted statistical method for forecasting volatility and correlation of financial asset returns. For many years, it was the market standard to forecast average volatility over the next h days by taking an equally weighted average of squared returns over the previous h days. This method was called the historical volatility forecast. Nowadays, many different statistical forecasting techniques can be applied to

historical time series data so it is confusing to call this equally weighted method the historical method. However, this rather confusing terminology remains standard.

Perceived changes in volatility and correlation have important consequences for all types of risk management decisions, whether to do with capitalization, resource allocation, or hedging strategies. Indeed it is these parameters of the returns distributions that are the fundamental building blocks of market risk assessment models. It is therefore essential to understand what type of variability in returns the model has measured. The model assumes that an independently and identically distributed process generates returns. That is, both volatility and correlation are constant and the “square root of time rule” applies. This assumption has important ramifications and we shall take care to explain these very carefully.

Statistical Methodology

The methodology for constructing a covariance matrix based on equally weighted averages can be described in very simple terms. Consider a set of time series $\{r_{i,t}\}$ $i = 1, \dots, m$; $t = 1, \dots, T$. Here the subscript i denotes the asset or risk factor, and t denotes the time at which each return is measured. We shall assume that each return has a zero mean. Then an unbiased estimate of the unconditional variance of the i th returns variable at time t , based on the T most recent daily returns as:

$$\hat{\sigma}_{i,t}^2 = \frac{\sum_{l=1}^T r_{i,t-l}^2}{T} \quad (3)$$

The term “unbiased estimator” means the expected value of the estimator is equal to the true value.

Note that (3) gives an unbiased estimate of the variance but this is not the same as the square of an unbiased estimate of the standard deviation. That is, $\sqrt{E(\hat{\sigma}^2)} = \sigma$ but $E(\hat{\sigma}) \neq \sigma$. So really the hat ‘^’ should be written over the whole of σ^2 .

But it is generally understood that the notation $\hat{\sigma}^2$ is used to denote the estimate or forecast of a variance, and not the square of an estimate of the standard deviation. So, in the case that the mean return is zero, we have

$$E(\hat{\sigma}^2) = \sigma^2$$

If the mean return is not assumed to be zero we need to estimate this from the sample, and this places a (linear) constraint on the variance estimated from sample data. In that case, to obtain an unbiased estimate we should use

$$s_{i,t}^2 = \frac{\sum_{l=1}^T (r_{i,t-l} - \bar{r}_i)^2}{T-1} \quad (4)$$

where \bar{r}_i is the average return on the i th series, taken over the whole sample of T data points. The mean-deviation form above may be useful for estimating variance using monthly or even weekly data over a period for which average returns are significantly different from zero. However with daily data the average return is usually very small and since, as we shall see below, the errors induced by other assumptions are huge relative to the error induced by assuming the mean is zero, we normally use the form (3).

Similarly, an unbiased estimate of the unconditional covariance of two zero mean returns at time t , based on the T most recent daily returns is:

$$\hat{\sigma}_{i,j,t} = \frac{\sum_{l=1}^n r_{i,t-l} r_{j,t-l}}{T} \quad (5)$$

As mentioned above, we would normally ignore the mean deviation adjustment with daily data.

The equally weighted unconditional covariance matrix estimate at time t for a set of k returns is thus $\hat{\mathbf{V}}_t = (\hat{\sigma}_{i,j,t})$ for $i, j = 1, \dots, k$. Loosely speaking, the term “unconditional” refers to the fact that it is the overall or long-run or average variance that we are estimating, as opposed to a conditional variance that can

change from day to day and is sensitive to recent events.

As mentioned in the introduction, we use the term “volatility” to refer to the annualized standard deviation. The equally weighted estimates of volatility and correlation are obtained in two stages. First, one obtains an unbiased estimate of the unconditional covariance matrix using equally weighted averages of squared returns and cross products of returns and the same number n of data points each time. Then these are converted into volatility and correlation estimates by applying the usual formulas. For instance, if the returns are measured at the daily frequency and there are 250 trading days per year:

$$\text{Equally weighted volatility} = \hat{\sigma}_t \sqrt{250} \quad (6)$$

$$\text{Equally weighted correlation} = \hat{\varrho}_{ij,t} = \frac{\hat{\sigma}_{ij,t}}{\hat{\sigma}_{i,t} \hat{\sigma}_{j,t}}$$

In the equally weighted methodology the forecasted covariance matrix is simply taken to be the current estimate, there being nothing else in the model to distinguish an estimate from a forecast. The original risk horizon for the covariance matrix is given by the frequency of the data—daily returns will give the 1-day covariance matrix forecast, weekly returns will give the 1-week covariance matrix forecast, and so forth. Then, since the model assumes that returns are independently and identically distributed we can use the square root of time rule to convert a 1-day forecast into an h -day covariance matrix forecast, simply by multiplying each element of the 1-day matrix by h . Similarly, a monthly forecast can be obtained for the weekly forecast by multiplying each element by 4, and so forth.

Having obtained a forecast of variance, volatility, covariance, and correlation we should ask: How accurate is this forecast? For this we could provide either a confidence interval, that is, a range within which we are fairly certain that the true parameter will lie, or a standard error for our parameter estimate. *The stan-*

dard error gives a measure of precision of the estimate and can be used to test whether the true parameter can take a certain value, or lie in a given range. The next few sections show how such confidence intervals and standard errors can be constructed.

Confidence Intervals for Variance and Volatility

A confidence interval for the true variance σ^2 when it is estimated by an equally weighted average can be derived using a straightforward application of sampling theory. Assuming the variance estimate is based on n normally distributed returns with an assumed mean of zero, then $T\hat{\sigma}^2/\sigma^2$ will have a chi-squared distribution with T degrees of freedom (see Freund, 1998). A $100(1 - \alpha)\%$ two-sided confidence interval for $T\hat{\sigma}^2/\sigma^2$ would therefore take the form $(\chi_{1-\alpha/2,T}^2, \chi_{\alpha/2,T}^2)$ and a straightforward calculation gives the associated confidence interval for the variance σ^2 as:

$$\left(\frac{T\hat{\sigma}^2}{\chi_{\alpha/2,T}^2}, \frac{T\hat{\sigma}^2}{\chi_{1-\alpha/2,T}^2} \right) \quad (7)$$

For example, a 95% confidence interval for an equally weighted variance forecast based on 30 observations is obtained using the upper and lower chi-squared critical values:

$$\chi_{0.975,30}^2 = 16.791 \quad \text{and} \quad \chi_{0.025,30}^2 = 46.979$$

So the confidence interval is $(0.6386\hat{\sigma}^2, 1.7867\hat{\sigma}^2)$ and exact values are obtained by substituting in the value of the variance estimate.

Figure 1 illustrates the upper and lower bounds for a confidence interval for a variance forecast when the equally weighted variance estimate is one. We see that as the sample size T increases, the width of the confidence interval decreases, markedly so as T increases from low values.

We can turn now to the confidence intervals that would apply to an estimate of volatility. Recall that volatility, being the square root of the variance, is simply a monotonic decreasing

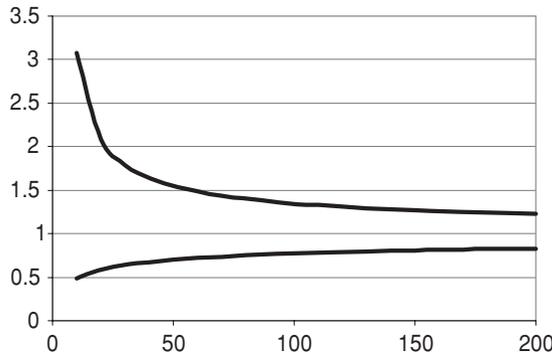


Figure 1 Confidence Interval for Variance Forecasts

transformation of the variance. Percentiles are invariant under any strictly monotonic increasing transformation. That is, if f is any monotonic increasing function of a random variable X , then:

$$P(c_l < X < c_u) = P(f(c_l) < f(X) < f(c_u)) \quad (8)$$

Property (8) provides a confidence interval for a historical volatility based on the confidence interval (7). Since \sqrt{x} is a monotonic increasing function of x , one simply takes the square root of the lower and upper bounds for the equally weighted variance. For instance if a 95% confidence interval for the variance is [16%, 64%] then a 95% for the associated volatility is [4%, 8%]. And, since x^2 is also monotonic increasing for $x > 0$, the converse also applies. Thus if a 95% confidence interval for the volatility is [4%, 8%] then a 95% for the associated variance is [16%, 64%].

Standard Errors for Equally Weighted Average Estimators

An estimator of any parameter has a distribution and a point estimate of volatility is just the expectation of the distribution of the volatility estimator. The distribution function of the equally weighted average volatility estimator is not just square root of the distribution function of the corresponding variance estimate. In-

stead, it may be derived from the distribution of the variance estimator via a simple transformation. Since volatility is the square root of the variance, the density function of the volatility estimator is

$$g(\hat{\sigma}) = 2\hat{\sigma}h(\hat{\sigma}^2) \quad \text{for } \hat{\sigma} > 0 \quad (9)$$

where $h(\hat{\sigma}^2)$ is the density function of the variance estimator. This follows from the fact that if y is a monotonic and differentiable function of x , then their probability densities $g(\cdot)$ and $h(\cdot)$ are related as $g(y) = |dx/dy|h(x)$ (see Freund, 1998). Note that when $y = \sqrt{x}$, $|dx/dy| = 2y$ and so $g(y) = 2yh(x)$.

In addition to the point estimate or expectation, one might also estimate the standard deviation of the distribution of the estimator. This is called the "standard error" of the estimate. The standard error determines the width of a confidence interval for a forecast and it indicates how reliable a forecast is considered to be. The wider the confidence interval, the more uncertainty there is in the forecast.

Standard errors for equally weighted average variance estimates are based on a normality assumption for the returns. Moving average models assume that returns are independent and identically distributed. Now assuming normality also, so that the returns are normally and independently distributed, denoted by $NID(0, \sigma^2)$, we apply the variance operator to (3). Note that if X_i are independent random variables ($i = 1, \dots, T$), then $f(X_i)$ are also independent for any monotonic differentiable function f . Hence, the squared returns are independent, and we have:

$$V(\hat{\sigma}_t^2) = \sum_{i=1}^T V(r_{t-i}^2)/T^2 \quad (10)$$

Since $V(X) = E(X^2) - E(X)^2$ for any random variable X , $V(r_t^2) = E(r_t^4) - E(r_t^2)^2$. By the zero mean assumption $E(r_t^2) = \sigma^2$ and assuming normality, $E(r_t^4) = 3\sigma^4$. Hence for every t :

$$V(r_t^2) = 3\sigma^4 - \sigma^4 = 2\sigma^4$$

and substituting this into (10) gives

$$V(\hat{\sigma}_t^2) = \frac{2\sigma^4}{T} \quad (11)$$

Hence, the standard error of an equally weighted average variance estimate based on T zero mean squared returns is $\sigma^2\sqrt{\frac{2}{T}}$ or simply $\sqrt{\frac{2}{T}}$, when expressed as a percentage of the variance. For instance, the standard error of the variance estimate is 20% when 50 observations are used in the estimate, and 10% when 200 observations are used in the estimate.

What about the standard error of the volatility estimator? To derive this, we first prove that for any continuously differentiable function f and random variable X :

$$V(f(X)) \approx f'(E(X))^2 V(X) \quad (12)$$

To show this, we take a second order Taylor expansion of f about the mean of X and then take expectations. See Alexander (2008a), Chapter 1. This gives:

$$E(f(X)) \approx f(E(X)) + \frac{1}{2}f''(E(X))V(X) \quad (13)$$

Similarly,

$$E(f(X)^2) \approx f(E(X))^2 + (f'(E(X)))^2 + f(E(X))f''(E(X))V(X) \quad (14)$$

again ignoring higher-order terms. The result (12) follows on noting that:

$$V(f(X)) = E(f(X)^2) - E(f(X))^2$$

We can now use (11) and (12) to derive the standard error of a historical volatility estimate. From (12) we have $V(\hat{\sigma}^2) \approx (2\hat{\sigma})^2 V(\hat{\sigma})$ and so:

$$V(\hat{\sigma}) \approx \frac{V(\hat{\sigma}^2)}{(2\hat{\sigma})^2} \quad (15)$$

Now using (11) in (15) we obtain the variance of the volatility estimator as:

$$V(\hat{\sigma}) = \left(\frac{1}{2\sigma^2}\right) \left(\frac{2\sigma^4}{T}\right) = \frac{\sigma^2}{2T} \quad (16)$$

so the standard error of the volatility estimator as a percentage of volatility is $(2T)^{-1/2}$. This result tells us that the standard error of the volatility estimator (as a percentage of volatility) is approximately one-half the size of the standard error of the variance (as a percentage of the variance).

Thus, as a percentage of the volatility, the standard error of the historical volatility estimator is approximately 10% when 50 observations are used in the estimate, and 5% when 200 observations are used in the estimate. The standard errors on *equally weighted moving average* volatility estimates become very large when only a few observations are used. This is one reason why it is advisable to use a long averaging period in historical volatility estimates.

It is harder to derive the standard error of an equally weighted average correlation estimate. However, it can be shown that

$$V(\hat{\rho}_{ij}) = \frac{1 - \rho^2}{T - 2} \quad (17)$$

and so we have the following t -distribution for the correlation estimate divided by its standard error:

$$\frac{\hat{\rho}_{ij}\sqrt{T-2}}{\sqrt{1-\hat{\rho}_{ij}^2}} \sim t_{T-2} \quad (18)$$

In particular, the significance of a correlation estimate depends on the number of observations that are used in the sample.

To illustrate testing for the significance of historical correlation, suppose that a historical correlation estimate of 0.2 is obtained using 38 observations. Is this significantly greater than zero? The null hypothesis is $H_0 : \rho = 0$, the alternative hypothesis is $H_1 : \rho > 0$, and the test statistic is (18). Computing the value of this statistic given our data gives

$$t = \frac{0.2 \times 6}{\sqrt{1 - 0.04}} = \frac{12}{\sqrt{96}} = \frac{3}{\sqrt{6}} = \sqrt{1.5} = 1.225$$

Even the 10% upper critical value of the t -distribution with 36 degrees of freedom is greater than this value (it is in fact 1.3). Hence

we cannot reject the null hypothesis: 0.2 is not significantly greater than zero when estimated from 38 observations. However, if the same value of 0.2 had been obtained from a sample with, say, 100 observations our t -value would have been 2.02, which is significantly positive at the 2.5% level because the upper 2.5% critical value of the t -distribution with 98 degrees of freedom is 1.98.

Equally Weighted Moving Average Covariance Matrices

An equally weighted “moving” average is calculated on a fixed size data “window” that is rolled through time, each day adding the new return and taking off the oldest return. The length of this window of data, also called the “look-back” period or averaging period, is the time interval over which we compute the average of the squared returns (for variance) or the average cross products of returns (for covariance). In the past, several large financial institutions have lost a lot of money because they used the equally weighted moving average model inappropriately. I would not be surprised if much more money was lost because of the inexperienced use of this model in the future. The problem is not the model itself—after all, it is a perfectly respectable statistical formula for an unbiased estimator—the problems arise from its inappropriate application within a time series context.

A (fallacious) argument goes as follows: Long-term predictions should be unaffected by short-term phenomena such as “volatility clustering” so it will be appropriate to take the average over a very long historic period. But short-term predictions should reflect current market conditions, which means that only the immediate past returns should be used. Some people use an historical averaging period of T days in order to forecast forward T days; others use slightly longer historical periods than the forecast period. For example, for a 10-day fore-

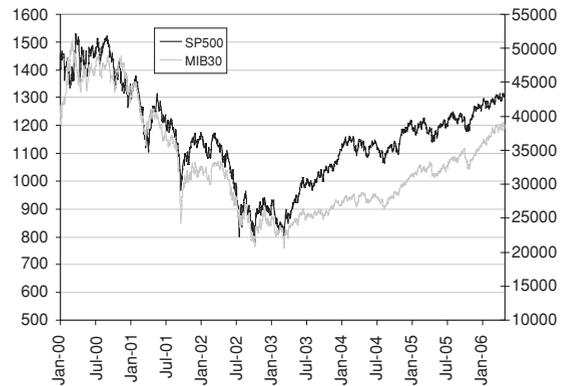


Figure 2 MIB 30 and S&P 100 Daily Close

cast, some practitioners might look back 30 days or more. But this apparently sensible approach actually induces a major problem. If one or more extreme returns is included in the averaging period, the volatility (or correlation) forecast can suddenly jump downward to a completely different level on a day when absolutely nothing happened in the markets. And prior to mysteriously jumping down, a historical forecast will be much larger than it should be.

Figure 2 illustrates the daily closing prices of the Italian MIB 30 stock index between the beginning of January 2000 and the end of April 2006 and compares these with the S&P 100 index prices over the same period. The prices were downloaded from Yahoo! Finance. We will show how to calculate the 30-day, 60-day, and 90-day historical volatilities of these two stock indexes and compare them graphically.

We construct three different equally weighted moving average volatility estimates for the MIB 30 index, with $T = 30$ days, 60 days and 90 days, respectively. The result is shown in Figure 3. Let us first focus on the early part of the data period and on the period after the September 11, 2001 (9/11), terrorist attack in particular. The Italian index reacted to the news far more than most other indexes. The volatility estimate based on 30 days of data jumped from 15% to nearly 50% in one day, and then continued to rise further, up to 55%. Then, suddenly, exactly 30 days after the event, 30-day volatility jumped down again

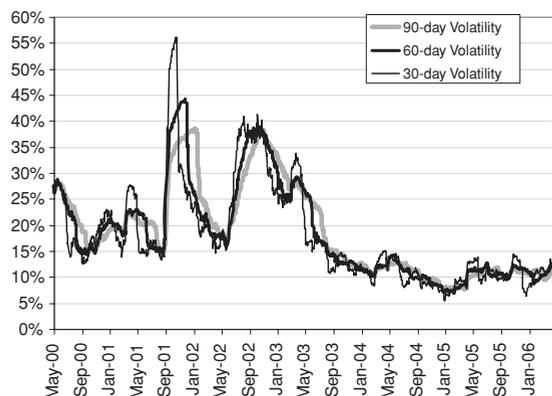


Figure 3 Equally Weighted Moving Average Volatility Estimates of the MIB 30 Index

to 30%. But nothing particular happened in the Italian markets on that day. The drastic fall in volatility was just a “ghost” of the 9/11 terrorist attack: It was no reflection at all of the real market conditions at that time.

Similar features are apparent in the 60-day and 90-day volatility series. Each series jumps up immediately after the 9/11 event, and then, either 60 or 90 days afterward, jumps down again. On November 9, 2001, the three different look-back periods gave volatility estimates of 30%, 43%, and 36%, but they are all based on the same underlying data and the same independent and identically distributed assumption for the returns! Other such ghost features are evident later in the period, for instance, in March 2001 and March 2003. Later on in the period, the choice of look-back period does not make so much difference: The three volatility estimates are all around the 10% level.

Case Study: Measuring the Volatility and Correlation of U.S. Treasuries

The interest rate covariance matrix is an important determinant of the value at risk (VaR) of a cash flow. In this section, we show how to estimate the volatilities and correlations of different maturity U.S. zero-coupon interest rates using the equal weighted moving average

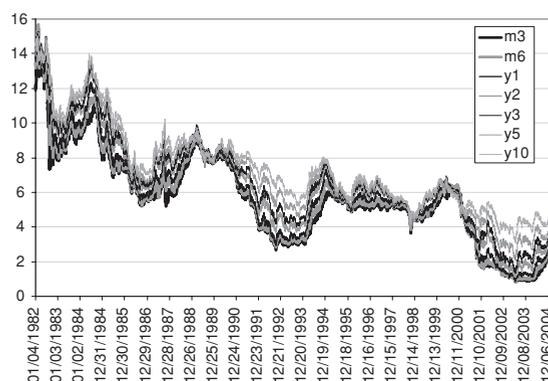


Figure 4 U.S. Treasury Rates

Source: <http://www.federalreserve.gov/releases/h15/data.htm>.

method. Consider daily data on constant maturity U.S. Treasury rates between January 4, 1982 and March 11, 2005. The rates are graphed in Figure 4.

It is evident that rates followed marked trends over the period. From a high of about 15% in 1982, by the end of the same period the short-term rates were below 3%. Also, periods where the term structure of interest rates is relatively flat are interspersed with periods when the term structure is upward sloping, sometimes with the long-term rates being several percent higher than the short-term rates. During the upward sloping yield curve regimes, especially the latter one from 2000 to 2005, the medium- to long-term interest rates are more volatile than the short-term rates, in absolute terms. However, it is not clear which rates are the most volatile in relative terms, as the short rates are much lower than the medium to long-term rates. There are three decisions that must be made:

Decision 1: How long an historical data period should be used?

Decision 2: Which frequency of observations should be used?

Decision 3: Should the volatilities and correlations be measured directly on absolute changes in interest rates, or should they be measured on relative changes and then the result converted into absolute terms?

Decision 1: How Long a Historical Data Period Should Be Used?

The equally weighted historical method gives an average volatility, or correlation, over the sample period chosen. The longer the data period, the less relevant that average may be today (that is, at the end of the sample). Looking at Figure 4, it may be thought that data from 2000 onward, and possibly also data during the first half of the 1990s, are relevant today. However, we may not wish to include data from the latter half of the 1990s, when the yield curve was flat.

Decision 2: Which Frequency of Observations Should Be Used?

This is an important decision, which depends on the end use of the covariance matrix. We can always use the square root of time rule to convert the holding period of a covariance matrix. For instance, a 10-day covariance matrix can be converted into a 1-day matrix by dividing each element by 10; and it can be converted into an annual covariance matrix by multiplying each element by 25. However, this conversion is based on the assumption that variations in interest rates are independent and identically distributed. Moreover, the data become more noisy when we use high-frequency data. For instance, daily variations may not be relevant if we only ever want to measure covariances over a 10-day period. The extra variation in the daily data is not useful, and the crudeness of the square root of time rule will introduce an error. To avoid the use of crude assumptions it is best to use a data frequency that corresponds to the holding period of the covariance matrix.

However, the two decisions above are linked. For instance, if data are quarterly, we need a data period of five or more years; otherwise, the standard error of the estimates will be very large. But then our quarterly covariance matrix represents an average over many years that may not be thought of as relevant today. If data are daily, then just one year of data provides plenty of observations to measure the historical model

volatilities and correlations accurately. Also, a history of one year is a better representation of today's markets than a history of five or more years. However, if it is a quarterly covariance matrix that we seek, we have to apply the square root of time rule to the daily matrix. Moreover, the daily variations that are captured by the matrix may not be relevant information at the quarterly frequency.

In summary, there may be a trade-off between using data at the relevant frequency and using data that are relevant today. It should be noted that such a trade-off between Decisions 1 and 2 above applies to the measurement of risk in all asset classes and not only to interest rates.

In interest rates, there is another decision to make before we can measure risk. Since the price value of a basis point (PV01) sensitivity vector is usually measured in basis points, an interest rate covariance matrix is also usually expressed in basis points. Hence, we have Decision 3.

Decision 3: Absolute versus Relative Measures

Should the volatilities and correlations be measured directly on absolute changes in interest rates, or should they be measured on relative changes and then the result converted into absolute terms?

If rates have been trending over the data period the two approaches are likely to give very different results. One has to make a decision about whether relative changes or absolute changes are the more stable. In these data, for example, an absolute change of 50 basis points in 1982 was relatively small, but in 2005 it would have represented a very large change. Hence, to estimate an average daily covariance matrix over the entire data sample, it may be more reasonable to suppose that the volatilities and correlations should be measured on relative changes and then converted to absolute terms.

Note, however, that a daily matrix based on the entire sample would capture a very long-term average of volatilities and correlations between daily U.S. Treasury rates, indeed it

is a 22-year average that includes several periods of different regimes in interest rates. Such a long-term average, which is useful for long-term forecasts, may be better based on lower frequency data (e.g., monthly). For a 1-day forecast horizon, we shall use only the data since January 1, 2000.

To make the choice for Decision 3, we take both the relative daily changes (the difference in the log rates) and the absolute daily changes (the differences in the rates, in basis-point terms). Then we obtain the standard deviation, correlation, and covariance in each case, and in the case of relative changes we translate the results into absolute terms. We now compare results based on relative changes with results based on absolute changes. The correlation matrix estimates based on the period January 1, 2000, to March 11, 2005, are shown in Table 1.

The matrices are similar. Both matrices display the usual characteristics of an interest rate term structure: Correlations are higher at the long end than the short end, and they decrease as the difference between the two maturities increases.

Table 1 Correlation of U.S. Treasuries

(a) Based on Relative Changes							
	m3	m6	y1	y2	y3	y5	y10
m3	1.00						
m6	0.77	1.00					
y1	0.53	0.84	1.00				
y2	0.44	0.69	0.88	1.00			
y3	0.42	0.66	0.84	0.97	1.00		
y5	0.39	0.62	0.79	0.91	0.96	1.00	
y10	0.32	0.54	0.71	0.82	0.88	0.95	1.00
(b) Based on Absolute Changes							
	m3	m6	y1	y2	y3	y5	y10
m3	1.00						
m6	0.79	1.00					
y1	0.54	0.81	1.00				
y2	0.40	0.67	0.87	1.00			
y3	0.37	0.62	0.83	0.97	1.00		
y5	0.33	0.57	0.77	0.92	0.95	1.00	
y10	0.26	0.48	0.69	0.84	0.88	0.95	1.00

Table 2 compares the volatilities of the interest rates obtained using the two methods. The figures in the last row of each table represent an average absolute volatility for each rate over the period January 1, 2000 to March 11, 2005. Basing this first on relative changes in interest rates, Table 2(a) gives the standard deviation of relative returns volatility in the first row. The long-term rates have the lowest standard deviations, and the medium-term rates have the highest standard deviations. These standard deviations are then annualized (by multiplying by $\sqrt{250}$, assuming each rate is independent and identically distributed) and multiplied by the level of the interest rate on March 11, 2005. There was a very marked upward sloping yield curve on March 11, 2005. Hence the long-term rates are more volatile than the short-term rates: For instance, the 3-month rate has an absolute volatility of about 76 basis points, but the absolute volatility of the 10-year rates is about 98 basis points.

Table 2(b) measures the standard deviation of absolute changes in interest rates over the period January 1, 2000 to March 11, 2005, and then converts this into volatility by multiplying by $\sqrt{250}$. We again find that the long-term rates are more volatile than the short-term rates; for instance, the six-month rate has an absolute volatility of about 62 basis points, but the absolute volatility of the five-year rates is about 106 bps. (It should be noted that it is quite unusual for long-term rates to be more volatile than short-term rates. But from 2000 to 2004 the U.S. Fed was exerting a lot of control on short-term rates, to bring down the general level of interest rates. However, the market expected interest rates to rise, because the yield curve was upward sloping during most of the period.) We find that correlations were similar, whether based on relative or absolute changes. But Table 2 shows there is a substantial difference between the volatilities obtained using the two methods. When volatilities are based directly on the absolute changes, they are slightly lower at the short end and substantially lower for the medium-term rates.

Table 2 Volatility of U.S. Treasuries

(a) Based on Relative Changes							
	m3	m6	y1	y2	y3	y5	y10
Standard deviation	0.0174	0.0172	0.0224	0.0267	0.0239	0.0187	0.0136
Yield curve on March 11, 2005	2.76	3.06	3.28	3.73	3.94	4.22	4.56
Absolute volatility (in basis points)	75.89	83.08	116.23	157.61	148.71	124.88	98.21
(b) Based on Absolute Changes							
	m3	m6	y1	y2	y3	y5	y10
Standard deviation	4.4735	3.9459	4.7796	6.4626	6.7964	6.7615	6.1738
Absolute volatility (in basis points)	70.73	62.39	75.57	102.18	107.46	106.91	97.62

Finally, we obtain the annual covariance matrix of absolute changes (in basis point terms) by multiplying the correlation matrix by the appropriate absolute volatilities and to obtain the one-day covariance matrix we divide by 250. The results are shown in Table 3. Depending on whether we base estimates of volatility and correlation on relative or absolute changes in interest rates, the covariance matrix can be very different. In this case, it is short-term and medium-term volatility estimates that are the most affected by the choice. Given that we have used the equally weighted average method-

Table 3 One-Day Covariance Matrix of U.S. Treasuries, in Basis Points

(a) Based on Relative Changes							
	m3	m6	y1	y2	y3	y5	y10
m3	23.04						
m6	19.46	27.61					
y1	18.85	32.26	54.04				
y2	20.87	36.29	64.50	99.36			
y3	18.98	32.86	58.28	91.14	88.46		
y5	14.75	25.84	45.95	71.94	71.01	62.38	
y10	9.67	17.70	32.45	51.07	51.29	46.47	38.58
(b) Based on Absolute Changes							
	m3	m6	y1	y2	y3	y5	y10
m3	20.01						
m6	13.96	15.57					
y1	11.65	15.30	22.84				
y2	11.69	17.01	26.86	41.77			
y3	11.17	16.76	26.96	42.73	46.19		
y5	9.89	15.21	25.03	40.09	43.81	45.72	
y10	7.17	11.71	20.25	33.34	36.92	39.55	38.12

ology to construct the covariance matrix, the underlying assumption is that volatilities and correlations are constant. Hence, the choice between relative or absolute changes depends on which are the more stable. In countries with very high interest rates, or when interest rates have been trending during the sample period, relative changes tend to be more stable than absolute changes.

In summary, there are four crucial decisions to be made when estimating a covariance matrix for interest rates:

1. Which statistical model should we employ?
2. Which historical data period should be used?
3. Should the data frequency be daily, weekly, monthly, or quarterly?
4. Should we base the matrix on relative or absolute changes in interest rates?

The first three decisions must also be made when estimating covariance matrices in other asset classes such as equities, commodities, and foreign exchange rates. There is a huge amount of model risk involved with the construction of covariance matrices; very different results may be obtained depending on the choice made.

Pitfalls of the Equally Weighted Moving Average Method

The problems encountered when applying this model stem not from the small jumps that are often encountered in financial asset prices, but

from the large jumps that are only rarely encountered. When a long averaging period is used, the importance of a single extreme event is averaged out within a large sample of returns. Hence, a moving average volatility estimate may not respond enough to a short, sharp shock in the market. This effect is clearly visible in 2002, where only the 30-day volatility rose significantly over a matter of a few weeks. The longer-term volatilities did rise, but it took several months for them to respond to the market falls in the MIB during mid-2002. At this point in time there was actually a cluster of volatility, which often happens in financial markets. The effect of the cluster was to make the longer-term volatilities rise, eventually, but then they took too long to return to normal levels. It was not until markets returned to normal in late 2003 that the three volatility series in Figure 2 are in line with each other.

When there is an extreme event in the market, even just one very large return will influence the T -day moving average estimate for exactly T days until that very large squared return falls out of the data window. Hence volatility will jump up, for exactly T days, and then fall dramatically on day $T + 1$, even though nothing happened in the market on that day. This type of ghost feature is simply an artifact of the use of equal weighting. The problem is that extreme events are just as important to current estimates, whether they occurred yesterday or a very long time ago. A single large, squared return remains just as important $T - 1$ days ago as it was yesterday. It will affect the T -day volatility or correlation estimate for exactly T days after that return was experienced, and to exactly the same extent. However, with other models we would find that volatility or correlation had long ago returned to normal levels. Exactly $T + 1$ days after the extreme event, the equally weighted moving average volatility estimate mysteriously drops back down to about the correct level—that is, provided that we have not had another extreme return in the interim!

Note that the smaller is T , the number of data points used in the data window, the more variable the historical volatility series will be. When any estimates are based on a small sample size they will not be very precise. The larger the sample size the more accurate the estimate, because sampling errors are proportional to $1/\sqrt{T}$. For this reason alone a short moving average will be more variable than a long moving average. Hence, a 30-day historic volatility (or correlation) will always be more variable than a 60-day historic volatility (or correlation) that is based on the same daily return data. Of course, if one really believes in the assumption of constant volatility that underlies this method, one should always use as long a history as possible, so that sampling errors are reduced.

It is important to realize that whatever the length of the historical averaging period and whenever the estimate is made, the equally weighted method is always estimating the same parameter: the unconditional volatility (or correlation) of the returns. But this is a constant—it does not change over the process. Thus, the variation in T -day historic estimates can only be attributed to sampling error: There is nothing else in the model to explain this variation. It is not a time-varying volatility model, even though some users try to force it into that framework.

The problem with the equally weighted moving average model is that it tries to make an estimate of a constant volatility into a forecast of a time-varying volatility. Similarly, it tries to make an estimate of a constant correlation into a forecast of a time-varying correlation. No wonder financial firms have lost a lot of money with this model! It is really only suitable for long-term forecasts of average volatility, or correlation, for instance over a period of between six months to several years. In this case, the look-back period should be long enough to include a variety of price jumps, with a relative frequency that represents the modeler expectations of the probability of future price jumps of that magnitude during the forecast horizon.

Using Equally Weighted Moving Averages

To forecast a long-term average for volatility using the equally weighted model, it is standard to use a large sample size T in the variance estimate. The confidence intervals for historical volatility estimators given earlier in this entry provide a useful indication of the accuracy of these long-term volatility forecasts and the approximate standard errors that we have derived earlier in this entry give an indication of variability in long-term volatility. Here, we saw that the variability in estimates decreased as the sample size increased. Hence, long-term volatility that is forecast from this model may prove useful.

When pricing options, it is the long-term volatility that is most difficult to forecast. Options trading often focuses on short-maturity options and long-term options are much less liquid. Hence, it is not easy to forecast a long-term implied volatility. Long-term volatility holds the greatest uncertainty, yet it is the most important determinant of long-term option prices.

We conclude this section with an interesting conundrum, considering two hypothetical historical volatility modelers, whom we shall call Tom and Dick, both forecasting volatility over a 12-month risk horizon based on equally weighted average of squared returns over the past 12 months of daily data. Imagine that it is January 2006 and that on October 15, 2005, the market crashed, returning -50% in the space of a few days. So some very large jumps occurred during the current data window, albeit three months ago.

Tom includes these extremely large returns in his data window, so his ex-post average of squared returns, which is also his volatility forecast in this model, will be very high. Because of this, Tom has an implicit belief that another jump of equal magnitude will occur during the forecast horizon. This implicit belief will continue until one year after the crash, when those

large negative returns fall out of his moving data window. Consider Tom's position in October 2006. Up to the middle of October he includes the crash period in his forecast but after that the crash period drops out of the data window and his forecast of volatility in the future suddenly decreases—as if he suddenly decided that another crash was very unlikely. That is, he drastically changes his belief about the possibility of an extreme return. So, to be consistent with his previous beliefs, should Tom now “bootstrap” the extreme returns experienced during October 2005 back into his data set?

And what about Dick, who in January 2006 does not believe that another market crash could occur in his 12-month forecast horizon? So, in January 2006, he should somehow filter out those extreme returns from his data. Of course, it is dangerous to embrace the possibility of bootstrapping in and filtering out extreme returns in data in an ad hoc way, before it is used in the model. However, if one does not do this, the historical model can imply a very strange behavior of the beliefs of the modeler.

In the Bayesian framework of uncertain volatility the equally weighted model has an important role to play. Equally weighted moving averages can be used to set the bounds for long-term volatility; that is, we can use the model to find a range $[\sigma_{min}, \sigma_{max}]$ for the long-term average volatility forecast. The lower bound σ_{min} can be estimated using a long period of historical data with all the very extreme returns removed and the upper bound σ_{max} can be estimated using the historical data where the very extreme returns are retained—and even adding some!

A modeler's beliefs about long-term volatility can be formalized by a probability distribution over the range $[\sigma_{min}, \sigma_{max}]$. This distribution would then be carried through for the rest of the analysis. For instance, upper and lower price bounds might be obtained for long-term exposures with option-like structures, such as warrants on a firm's equity or convertibles bonds.

This type of Bayesian method, which provides a price distribution rather than a single price, will be increasingly used in market risk management in the future.

EXPONENTIALLY WEIGHTED MOVING AVERAGES

An *exponentially weighted moving average* (EWMA) avoids the pitfalls explained in the previous section because it puts more weight on the more recent observations. Thus as extreme returns move further into the past as the data window slides along, they become less important in the average.

Statistical Methodology

An exponentially weighted moving average can be defined on any time series of data. Say that on date t we have recorded data up to time $t - 1$, so we have observations (x_{t-1}, \dots, x_1) . The exponentially weighted average of these observations is defined as:

$$\begin{aligned} \text{EWMA}(x_{t-1}, \dots, x_1) \\ = \frac{x_{t-1} + \lambda x_{t-2} + \lambda^2 x_{t-3} + \dots + \lambda^{t-2} x_1}{1 + \lambda + \lambda^2 + \dots + \lambda^{t-2}} \end{aligned}$$

where λ is a constant, $0 < \lambda < 1$, called the smoothing or the decay constant. Since $\lambda^T \rightarrow 0$ as $T \rightarrow \infty$ the exponentially weighted average places negligible weight on observations far in the past. And since $1 + \lambda + \lambda^2 + \dots = (1 - \lambda)^{-1}$ we have, for large t ,

$$\begin{aligned} \text{EWMA}(x_{t-1}, \dots, x_1) &\approx \frac{x_{t-1} + \lambda x_{t-2} + \lambda^2 x_{t-3} + \dots}{1 + \lambda + \lambda^2 + \dots} \\ &= (1 - \lambda) \sum_{i=1}^{\infty} \lambda^{i-1} x_{t-i} \end{aligned}$$

This is the formula that is used to calculate exponentially weighted moving average (EWMA) estimates of variance (with x being the squared return) and covariance (with x being the cross product of the two returns). As with equally weighted moving averages, it is

standard to use squared daily returns and cross products of daily returns, not in mean deviation form. That is:

$$\hat{\sigma}_t^2 = (1 - \lambda) \sum_{i=1}^{\infty} \lambda^{i-1} r_{t-i}^2 \quad (19)$$

and

$$\hat{\sigma}_{12,t} = (1 - \lambda) \sum_{i=1}^{\infty} \lambda^{i-1} r_{1,t-i} r_{2,t-i} \quad (20)$$

The above formulas may be rewritten in the form of recursions, more easily used in calculations:

$$\hat{\sigma}_t^2 = (1 - \lambda) r_{t-1}^2 + \lambda \hat{\sigma}_{t-1}^2 \quad (21)$$

and

$$\hat{\sigma}_{12,t} = (1 - \lambda) r_{1,t-1} r_{2,t-1} + \lambda \hat{\sigma}_{12,t-1} \quad (22)$$

An alternative notation used for the above is $V_\lambda(r_t)$, for $\hat{\sigma}_t^2$ and $\text{COV}_\lambda(r_{1,t}, r_{2,t})$ for $\hat{\sigma}_{12,t}$ when we want to make explicit the dependence on the *smoothing constant*.

One converts the variance to volatility by taking the annualized square root, the annualizing constant being determined by the data frequency as usual. Note that for the EWMA correlation the covariance is divided by the square root of the product of the two EWMA variance estimates, all with the same value of λ . Similarly for the EWMA beta the covariance between the stock (or portfolio) returns and the market returns is divided by the EWMA estimate for the market variance, both with the same value of λ . That is:

$$\hat{\rho}_{t,\lambda} = \frac{\text{COV}_\lambda(r_{1,t}, r_{2,t})}{\sqrt{V_\lambda(r_{1,t}) V_\lambda(r_{2,t})}} \quad (23)$$

and

$$\hat{\beta}_{t,\lambda} = \frac{\text{COV}_\lambda(X_t, Y_t)}{V_\lambda(X_t)} \quad (24)$$

Interpretation of λ

There are two terms on the right-hand side of (21). The first term $(1 - \lambda) r_{t-1}^2$ determines the intensity of reaction of volatility to market

events: The smaller is λ the more the volatility reacts to the market information in yesterday's return. The second term $\lambda\hat{\sigma}_{t-1}^2$ determines the persistence in volatility: Irrespective of what happens in the market, if volatility was high yesterday it will be still be high today. The closer that λ is to 1, the more persistent is volatility following a market shock.

Thus, a high λ gives little reaction to actual market events but great persistence in volatility, and a low λ gives highly reactive volatilities that quickly die away. An unfortunate restriction of exponentially weighted moving average models is that the reaction and persistence parameters are not independent: The strength of reaction to market events is determined by $1 - \lambda$, while the persistence of shocks is determined by λ . But this assumption is not empirically justified except perhaps in a few markets (e.g., major U.S. dollar exchange rates).

The effect of using a different value of λ in EWMA volatility forecasts can be quite substantial. Figure 5 compares two EWMA volatility estimates/forecasts of the S&P 100 index, with $\lambda = 0.90$ and $\lambda = 0.975$. It is not unusual for these two EWMA estimates to differ by as much as 10%.

So which is the best value to use for the smoothing constant? How should we choose λ ? This is not an easy question. (By contrast, in generalized autoregressive conditional het-

eroskedasticity (GARCH) models there is no question of how we should estimate parameters, because maximum likelihood estimation is an optimal method that always gives consistent estimators.) Statistical methods may be considered: For example, λ could be chosen to minimize the root mean square error between the EWMA estimate of variance and the squared return. But, in practice, λ is often chosen subjectively because the same value of λ has to be used for all elements in an EWMA covariance matrix. As a rule of thumb, we might take values of λ between about 0.75 (volatility is highly reactive but has little persistence) and 0.98 (volatility is very persistent but not highly reactive).

Properties of the Estimates

An EWMA volatility estimate will react immediately following an unusually large return, then the effect of this return on the EWMA volatility estimate gradually diminishes over time. The reaction of EWMA volatility estimates to market events therefore persists over time, and with a strength that is determined by the smoothing constant λ . The larger the value of λ , the more weight is placed on observations in the past and so the smoother the series becomes.

Figure 6 compares the EWMA volatility of the MIB index with $\lambda = 0.95$ and the 60-day equally

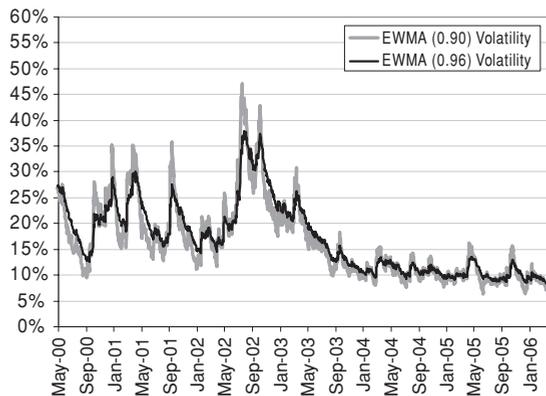


Figure 5 EWMA Volatility Estimates for SP100 with Different λ s

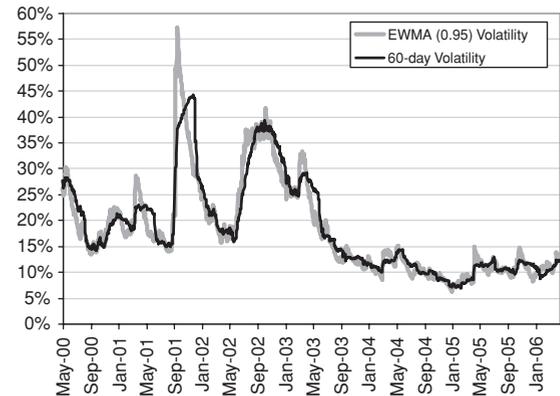


Figure 6 EWMA versus Equally Weighted Volatility

weighted volatility estimate. The difference between the two estimators is marked following an extreme market return. The EWMA estimate gives a higher volatility than the equally weighted estimate, but it returns to normal levels faster than the equally weighted estimate because it does not suffer from the ghost features discussed above.

One of the disadvantages of using EWMA to estimate and forecast covariance matrices is that the same value of λ is used for all the variances and covariances in the matrix. For instance, in a large matrix covering several asset classes, the same λ applies to all equity indexes, foreign exchange rates, interest rates, and/or commodities in the matrix. But why should all these risk factors have similar reaction and persistence to shocks? This constraint is commonly applied merely because it guarantees that the matrix will be positive semidefinite.

The EWMA Forecasting Model

The exponentially weighted average variance estimate (19), or in its equivalent form (21), is just a methodology for calculating $\hat{\sigma}_t^2$. That is, it gives a variance estimate at any point in time but there is no model as such that explains the behavior of the variance of returns, σ_t^2 at each time t . In this sense, we have to distinguish EWMA from a GARCH model, which starts with a proper specification of the dynamics of σ_t^2 and then proceeds to estimate the parameters of this model.

Without a proper model, it is not clear how we should turn our current estimate of variance into a forecast of variance over some future horizon. One possibility is to augment (21) by assuming it is the estimate associated with the model

$$\sigma_t^2 = (1 - \lambda) r_{t-1}^2 + \lambda \sigma_{t-1}^2 \quad r_t | I_{t-1} \sim N(0, \sigma_t^2) \quad (25)$$

An alternative is to assume a constant volatility, so the fact that our estimates are time vary-

ing is merely due to sampling error. In that case any EWMA variance forecast must be constant and equal to the current EWMA estimate. Similar remarks apply to the EWMA covariance, this time regarding EWMA as a simplistic version of bivariate normal GARCH. Similarly, the EWMA volatility (or correlation) forecast for all risk horizons is simply set at the current EWMA estimate of volatility (or correlation). The base horizon for the forecast is given by the frequency of the data—daily returns will give the one-day covariance matrix forecast, weekly returns will give the one-week covariance matrix forecast, and so forth. Then, since the returns are independent and identically distributed, the square root of time rule applies. So we can convert a one-day forecast into an h -day covariance matrix forecast by multiplying each element of the one-day EWMA covariance matrix by h .

Since the choice of λ itself is quite ad hoc, as discussed above, some users choose different values of λ for forecasting over different horizons. For instance, as discussed later in this entry, in the RiskMetrics™ methodology a relatively low value of λ is used for short-term forecasts and a higher value of λ is used for long-term forecasts. However, this is purely an ad hoc rule.

Standard Errors for EWMA Forecasts

In the previous section, we justified the assumption that the underlying returns are normally and independently distributed with mean zero and variance σ^2 . That is, for all t

$$E(r_t) = 0 \quad \text{and} \quad V(r_t) = E(r_t^2) = \sigma^2$$

In this section, we use this assumption to obtain standard errors for EWMA forecasts. From the above, and further from the normality assumption, we have:

$$V(r_t^2) = E(r_t^4) - E(r_t^2)^2 = 3\sigma^4 - \sigma^4 = 2\sigma^4$$

Now we can apply the variance operator to (21) and calculate the variance of the EWMA

variance estimator as:

$$V(\hat{\sigma}_t^2) = \frac{(1-\lambda)^2}{(1-\lambda^2)} V(r_t^2) = 2 \frac{1-\lambda}{1+\lambda} \sigma^4 \quad (26)$$

For instance, as a percentage of the variance, the standard error of the EWMA variance estimator is about 5% when $\lambda = 0.95$, 10.5% when $\lambda = 0.9$, and 16.2% when $\lambda = 0.85$.

A single point forecast of volatility can be very misleading. A forecast is always a distribution. It represents our uncertainty over the quantity that is being forecast. The standard error of a volatility forecast is useful because it can be translated into a standard error for a VaR estimate, for instance, or an option price. In any VaR model one should be aware of the uncertainty that is introduced by possible errors in the forecast of the covariance matrix. Similarly, in any mark-to-model value of an option, one should be aware of the uncertainty that is introduced by possible errors in the volatility forecast.

The RiskMetrics™ Methodology

Three very large covariance matrices, each based on a different moving average methodology, are available from www.riskmetrics.com. These matrices cover all types of assets including government bonds, money markets, swaps, foreign exchange, and equity indexes for 31 currencies and commodities. Subscribers have access to all of these matrices updated on a daily basis—and end-of-year matrices are also available to subscribers wishing to use them in scenario analysis. After a few days, the datasets are also made available free for educational use.

The RiskMetrics™ group is the market leader in market and credit risk data and modeling for banks, corporate asset managers, and financial intermediaries. It is highly recommended that readers visit the Web site (www.riskmetrics.com), where they will find a surprisingly large amount of information in the form of free publications and data. See the References at the end of this entry for details.

The three covariance matrices provided by the RiskMetrics group are each based on a history of daily returns in all the asset classes mentioned above. They are:

1. **Regulatory matrix:** This takes its name from the (unfortunate) requirement that banks must use at least 250 days of historical data for VaR estimation. Hence this metric is an equally weighted average matrix with $n = 250$. The volatilities and correlations constructed from this matrix represent forecasts of average volatility (or correlation) over the next 250 days.
2. **Daily matrix:** This is an EWMA covariance matrix with $\lambda = 0.94$ for all elements. It is not dissimilar to an equally weighted average with $n = 25$, except that it does not suffer from the ghost features caused by very extreme market events. The volatilities and correlations constructed from this matrix represent forecasts of average volatility (or correlation) over the next day.
3. **Monthly matrix:** This is an EWMA covariance matrix with $\lambda = 0.97$ for all elements and then multiplied by 25 (that is, using the square root of time rule and assuming 25 days per month). The volatilities and correlations constructed from this matrix represent forecasts of average volatility (or correlation) over the next 25 days.

The main difference between the three different methods is evidenced following major market movements: The regulatory forecast will produce a ghost effect of this event, and does not react as much as the daily or monthly forecasts. The most reactive is the daily forecast, but it also has less persistence than the monthly forecast.

Figure 7 compares the estimates for the FTSE 100 volatility based on each of the three RiskMetrics methodologies and using daily data from January 2, 1995, to June 23, 2006. As mentioned earlier in this entry, these estimates are assumed to be the forecasts over, respectively, one day, one month, and one year. In volatile

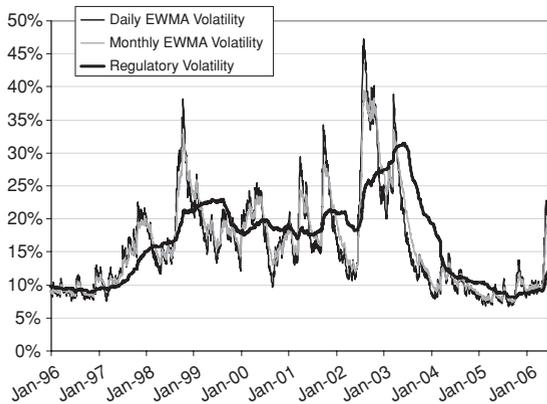


Figure 7 Comparison of the RiskMetrics “Forecasts” for FTSE100 Volatility

times, the daily and monthly estimates lie well above the regulatory forecast and the converse is true in more tranquil periods. For instance, during most of 2003, the regulatory estimate of average volatility over the next year was about 10% higher than both of the shorter-term estimates. However, it was falling dramatically during this period, and indeed the regulatory forecast of more than 20% volatility on average between June 2003 and June 2004 was entirely wrong. However, at the end of the period, in June 2006, the daily forecasts were above 20%, and the monthly forecasts were only just below this. However, the regulatory forecast over the next year was only slightly more than 10%.

During periods when the markets have been tranquil for some time, for instance during the whole of 2005, the three forecasts tend to agree more. But during and directly after a volatile period there are large differences between the regulatory forecasts and the two EWMA forecasts, and these differences are very difficult to justify. Neither the equally weighted average nor the EWMA methodology is based on a proper forecasting model. One simply assumes the current estimate is the volatility forecast. But the current estimate is a backward-looking measure based on recent historical data. So both of these moving average models make the assumption that the behavior of future volatility

is the same as its past behavior and this is a very simplistic view.

KEY POINTS

- The equally weighted moving average, or historical approach to estimating/forecasting volatilities and correlations, was the only statistical method used by practitioners until the mid-1990s.
- The historical method may provide a useful indication of the possible range for a long-term average, such as the average volatility or correlation over the next several years. However, its application to short-term forecasting suffers from at least four drawbacks: (1) The forecast of volatility/correlation over all future horizons is simply taken to be the current estimate of volatility, because the underlying assumption in the model is that returns are independent and identically distributed; (2) the only choice facing the user is on the data points to use in the data window; (3) following an extreme market move the forecasts of volatility and correlation will exhibit a so-called “ghost” feature of that extreme move, which will severely bias the volatility and correlation forecasts upward; and (4) the extent of this bias depends very much on the size of the data window.
- The bias issue associated with the historical approach was addressed by the RiskMetrics™ data and software suite. The choice of methodology helped to popularize the use of exponentially weighted moving averages (EWMA) by financial analysts.
- The EWMA approach provides useful forecasts for volatility and correlation over the very short term, such as over the new day or week. However, its use for longer-term forecasting is limited, and this methodology has two major problems: (1) The forecast of volatility/correlation over all future horizons is simply taken to be the current estimate of volatility, because the underlying assumption

in the model is that returns are independent and identically distributed, and (2) the only choice facing the user is about the value of the smoothing constant. With the EWMA approach, the forecasts produced depend crucially on this decision, yet there is no statistical procedure to choose for the value of the smoothing constant.

- Moving average models assume returns are independent and identically distributed, and the further assumption that they are normally distributed allows one to derive standard errors and confidence intervals for moving average forecasts. But empirical observations suggest that returns to financial assets are hardly ever independent and identically, let alone normally, distributed. For these reasons more and more practitioners are basing their forecasts on generalized autoregressive conditional heteroskedasticity (GARCH) models.
- There is no doubt that GARCH models produce superior volatility forecasts. It is only in GARCH models that the term structure volatility forecasts converge to the long-run average volatility—the other models produce constant volatility term structures. Moreover, the value of the EWMA smoothing constant is

chosen subjectively and the same smoothing constant must be used for all the returns, otherwise the covariance matrix need not be positive semidefinite. But GARCH parameters are estimated optimally and GARCH covariance matrices truly reflect the time-varying volatilities and correlations of the multivariate returns distributions.

REFERENCES

- Alexander, C. (2008a). *Market Risk Analysis*. Volume I: *Quantitative Methods in Finance*. Chichester, UK: John Wiley & Sons.
- Alexander, C. (2008b). *Market Risk Analysis*. Volume II: *Practical Financial Econometrics*. Chichester, UK: John Wiley & Sons.
- Freund, J. E. (1998). *Mathematical Statistics*. Englewood Cliffs: Pearson U.S. Imports & PHIPES.
- RiskMetrics (1996). *RiskMetrics Technical Document*, <http://www.riskmetrics.com/rmcovv.html>.
- RiskMetrics (1999). *Risk Management—A Practical Guide*, <http://www.riskmetrics.com/pracovv.html>.
- RiskMetrics (2001). *Return to RiskMetrics: The Evolution of a Standard*, <http://www.riskmetrics.com/r2rovv.html>.

Software for Financial Modeling

Introduction to Financial Model Building with MATLAB

DESSISLAVA A. PACHAMANOVA, PhD

Associate Professor of Operations Research, Babson College

Abstract: MATLAB is a modeling environment that allows for input and output processing, statistical analysis, simulation, and other types of model building for the purpose of analysis of a situation. MATLAB uses a number-array-oriented programming language; that is, a programming language in which vectors and matrices are the basic data structures. Reliable built-in functions, a wide range of specialized toolboxes, easy interface with widespread software like Microsoft Excel, and beautiful graphing capabilities for data visualization make implementation with MATLAB efficient and useful for the financial modeler.

MATLAB is an interactive computing environment for model development that also enables data visualization, data analysis, and numerical simulation. The core of the MATLAB environment was created as a number-array-oriented programming language; that is, as a programming language in which vectors and matrices are the basic data structures. (MATLAB stands for Matrix Laboratory.) Operations involving matrices and vectors can be performed efficiently within the core MATLAB *software* product. More specialized operations, such as statistical data analysis, optimization, and simulation, can be accessed by purchasing some of MATLAB's specialized toolboxes. Once a toolbox is installed, functions from the toolbox can be called in the same way as standard MATLAB functions, without any special additional syntax. MATLAB toolboxes that are useful for quantitative analysis in financial applications include:

- Statistics Toolbox
- Optimization Toolbox
- Global Optimization Toolbox
- Curve Fitting Toolbox
- Neural Network Toolbox
- Partial Differential Equation Toolbox

For example, the Statistics Toolbox contains data analysis tools (for multivariate analysis, statistical tests, statistical plots), random number generation tools, and quasi-random number generation tools, which are useful for implementing risk management and derivative pricing routines. The Optimization Toolbox contains solvers for linear, quadratic, nonlinear, and binary optimization, which can aid quantitative portfolio allocation schemes. The Global Optimization Toolbox contains randomized search optimization subroutines that can be used for solving complex (e.g., mixed-integer) optimization problems to near

optimality. It is useful, for example, for creating more complex portfolio allocation or trading routines. For more details and information about the other toolboxes, see the Mathworks website, <http://www.mathworks.com>.

MATLAB also has toolboxes that are specifically targeted at financial applications. These toolboxes include:

- Financial Toolbox
- Econometrics Toolbox
- Datafeed Toolbox
- Fixed-Income Toolbox
- Financial Derivatives Toolbox

For example, the Financial Toolbox contains specialized routines for computing frequently used financial quantities, such as present and future value, basic portfolio optimization, term structure of interest rates, bond prices, and derivative prices. It also contains functions that help with the manipulation of typical financial data sets, such as multivariate regression with missing data. Many of these routines can be implemented by using standard MATLAB functions, but the Financial Toolbox puts them together in a convenient package.

It is worth noting that most of the financial toolboxes require installation of one or more of the mathematics toolboxes listed earlier. For example, the Financial Toolbox requires the Statistics Toolbox and the Optimization Toolbox. The Financial Derivatives Toolbox requires the Statistics, Optimization, and Finance Toolboxes.

Another tool of interest to those who use Windows and Microsoft Excel extensively as the platform for their applications is Spreadsheet Link EX. Spreadsheet Link EX enables the manipulation of Microsoft Excel worksheets from within MATLAB and using MATLAB functions from within Excel. This is a useful toolbox that allows powerful MATLAB capabilities to be accessed through a familiar interface.

This entry provides brief pointers to important aspects of *modeling* in MATLAB. We discuss basic array construction and operations, func-

tions and scripts, as well as graphs. We also provide examples of MATLAB code for portfolio optimization schemes and for pricing a European call option by simulation.

When readers try to implement such routines themselves, they may find it useful to know that the MATLAB manual and online help contain abundant information and examples. Detailed documentation is also provided in MATLAB itself. For example, typing `help` at the prompt in MATLAB lists all major topics. Type `help name of function` at the prompt or in the box in the Help dialog box to access the documentation on that function in MATLAB. If unsure of which help topic is relevant, click on the button with question mark () in MATLAB's top menu. It provides richer search options.

THE MATLAB DESKTOP AND EDITOR

The standard MATLAB desktop window contains a **Workspace** window, a **Command History** window, and a **Command** window (see Figure 1). Depending on how you customize the MATLAB desktop window, however, you may see more or fewer windows. To check which windows are currently displayed and view other options, click on **Desktop** in the top MATLAB desktop window menu.

MATLAB commands are entered in the **Command** window. When a series of commands need to be given, it is more convenient to list them in an *M-file*, which is basically a file with instructions that MATLAB executes sequentially. Such files (*scripts*) are saved with the suffix “.m” and can be called from the prompt in the **Command** window typing their name (without the suffix “.m”). For example, if you create a file **OptimizePortfolio.m** with instructions on how to perform optimal portfolio allocation, you can call that file from the MATLAB command prompt by typing

```
>> OptimizePortfolio
```

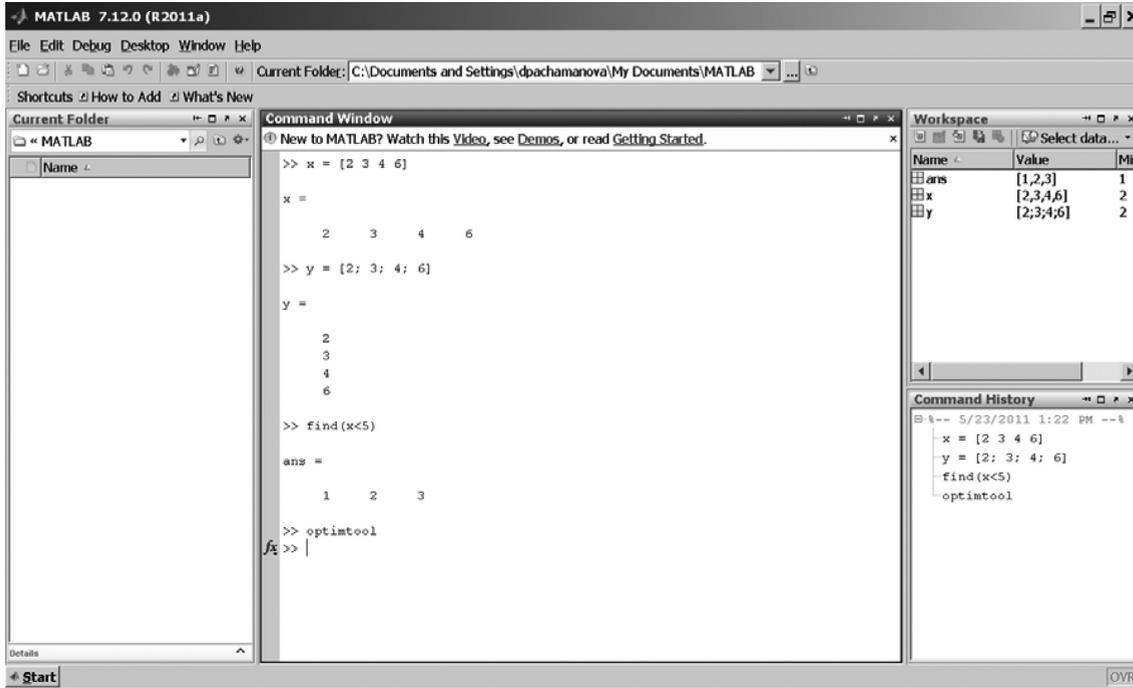


Figure 1 The Standard MATLAB Desktop

(If the file is saved in a directory other than the default MATLAB directory, you will need to make sure that MATLAB can find the file. Select **Desktop > Current Directory** from the top menu and navigate to the correct directory before typing the command at the prompt.)

To create an M-file, you can use any text editing program, such as WordPad, NotePad, and the open source editor Emacs. In general, it is convenient to use an editor that recognizes the MATLAB file type and provides helpful highlighting for parts of the code that have different characteristics. (For example, comments in the code appear in different colors than commands.) MATLAB's own editor can do that, and Emacs can be set up to recognize the MATLAB file format as well.

To call MATLAB's editor in order to create or edit M-files, select **Desktop > Editor** from the top menu. Alternatively, you can use the shortcut buttons at the top of the MATLAB desktop window: the button  to open the MATLAB

editor to write a new file, or the button  to open a file that has already been created.

BASIC OPERATIONS AND MATRIX ARRAY CONSTRUCTION

Basic Mathematical Operations

MATLAB can perform many kinds of different mathematical operations, such as addition (+), multiplication (* or .*), square root (sqrt or sqrtm), and power (^). These commands can be entered at the command prompt. For example, typing

```
>> 3*sqrt(4) + 15
```

and pressing Enter produces the output

```
ans =
    21
```

To suppress output, use the semicolon (;). For example, entering

```
>> 3*sqrt(4) + 15;
```

does not result in any visible output in the command window. However, MATLAB still performs the calculation. To see this, let us assign the value of the above expression to a variable, `ExpressionValue`:

```
>> ExpressionValue = 3*sqrt(4) + 15;
```

Then, typing `ExpressionValue` at the command prompt, you get

```
>> ExpressionValue
ExpressionValue =
    21
```

Constructing Vectors and Matrices

As mentioned earlier, MATLAB's core data structures are vectors and matrices. For example, the command

```
>> x = [2 3 4 6]
```

produces a horizontal vector array (one row) `x` that contains the numbers 2, 3, 4, and 6.

The semicolon (;) is used to create new rows. To create a vertical vector array `y` with the same entries, you can enter

```
>> y = [2; 3; 4; 6]
```

or press **Enter** after entering each number. (MATLAB treats semicolons and carriage returns in array declarations as new lines.) The different syntax is useful depending on the source for downloading the data that populate the *arrays*.

Matrices are declared similarly. For example, a 2-by-2 matrix `X` can be specified as

```
>> X = [1 2 3 4; 5 6 7 8]
X =
     1     2     3     4
     5     6     7     8
```

MATLAB is case-sensitive; that is, it will treat the matrix `X` and the vector `x` defined earlier as separate variables.

Special commands exist for declaring types of matrices that are used often. For example,

```
>> I = eye(3,3)
I =
     1     0     0
     0     1     0
     0     0     1
```

produces a 3×3 identity matrix.

Similarly, the commands `ones(n,m)` and `zeros(n,m)` can be used to declare matrices that contain only 0s or 1s of the desired dimension (`n × m`), and `diag(x)` can be used to create a matrix that has a vector `x` as its diagonal elements, and 0s everywhere else.

You can also “stack” matrices and vectors. For example,

```
>> Y = [x; X]
Y =
     2     3     4     6
     1     2     3     4
     5     6     7     8
```

Basic Array Operations

To transpose an array `A`, use the command `transpose(A)` or `A'`. This operation converts a horizontal vector into a vertical one and vice versa, and flips the elements of a matrix that contains real numbers in its entries around the diagonal, keeping the diagonal entries the same.

For example,

```
>> x'
ans =
     1     5
     2     6
     3     7
     4     8
```

To multiply two arrays, you can simply use the multiplication command `*`. Since the operation `*` performs a matrix multiplication, you

need to make sure that the matrix dimensions agree. For example, an error results in the case when the 1×4 array x is multiplied by the 2×4 array X :

```
>> x*X
???) Error using ==> mtimes
Inner matrix dimensions must agree.
```

To multiply x and X correctly, you can instead type

```
>> x*X'
ans =
    44    104
```

If you need to perform an element-by-element multiplication of two arrays (of equal sizes), use the `.*` operator. For example,

```
>> X.*X
ans =
     1     4     9    16
    25    36    49    64
```

Note that this is different from the matrix product. The matrix product would produce the following result:

```
>> X'*X
ans =
    26    32    38    44
    32    40    48    56
    38    48    58    68
    44    56    68    80
```

When a matrix array is multiplied by a number, all of the array's entries are multiplied by that number. Similarly, if a number is added to a matrix array, the number will be added to all of the elements of the matrix. For example,

```
>> 10+X
ans =
    11    12    13    14
    15    16    17    18
```

Extracting Information from Arrays

Suppose you have a matrix array `Data` with financial data on annual stock returns over

10 years for 1,000 companies traded on the New York Stock Exchange, and you would like to check the entry for the return on stock 253 in year 7. You are dealing with a 10×1000 matrix array in which each row is a time period and each column contains the returns on a particular stock. You are looking for the element in row 7, column 253 of this array. This can be requested with the command `Data(7,253)`.

Suppose now that you would like to extract information on all of stock 253's returns over the 10 years. This means that you are looking for the elements of column 253 of the matrix array. This can be requested with the command `Data(:,253)`. The colon operator replaces the row index to specify that elements with all indexes in the 253rd column should be produced. Similarly, if you would like to request all elements in the same row (e.g., the returns on all stocks in year 7), you can use the colon operator again: `Data(7,:)`.

To illustrate the output, let us use the matrix array X . To find out what the value of the element in row 1, column 3 is, enter

```
>> X(1,3)
ans =
     3
```

The third column of X is

```
>> X(:,3)
ans =
     3
     7
```

Similarly, the second row of X can be obtained as

```
>> X(2,:)
ans =
     5     6     7     8
```

IMPORTANT MATLAB FUNCTIONS

MATLAB supports a number of built-in functions. A function is written as a command and

takes arguments as inputs in parentheses. It processes the inputs by using operations hidden from the user and passes the final results back to the user. While we cannot cover many of the MATLAB functions in this brief introduction, we illustrate how functions work with an example of the function `find`, which can be useful in many situations.

`find` takes in an array and a condition as arguments and returns the indexes of elements within the array that satisfy the condition. In addition to traditional applications, `find` can be very helpful when dealing with missing data, which happens often with financial time series.

Suppose you want to find the indexes of the elements that are less than 5 of the 1×4 array `x` from the previous section. At the prompt, type

```
>> find(x<5)
```

The result is

```
ans =
     1     2     3
```

Now let us see how `find` works when the array is a matrix rather than a vector. Recall that `Y` was the matrix array obtained by stacking `x` and `X`. Suppose you want to find the indexes of the elements in the array that are less than 5. At the prompt, type

```
>> ind = find(Y<5)
```

MATLAB creates the following array:

```
ind =
     1
     2
     4
     5
     7
     8
    11
```

MATLAB treated the matrix array as a stacked-up collection of column vectors. The elements of the array `ind` correspond to the indexes of the elements in that long column vector. Obtaining the actual elements of `Y` that cor-

respond to these indexes can be accomplished by typing

```
>> Y(ind)
```

This produces the answer

```
ans =
     2
     1
     3
     2
     4
     3
     4
```

The indexing of an array as a sequence of stacked columns works well if the array is a vector, but it can get confusing if the array is a matrix. In the latter case, it is more intuitive to obtain the indexes as a row and column index. For example,

```
>> [indRow,indCol] = find(Y<5)
```

produces

```
indRow =
     1
     2
     1
     2
     1
     2
     2
indCol =
     1
     1
     2
     2
     3
     3
     4
```

This means that the following elements of `Y` have values less than 5: (row 1, column 1), (row 2, column 1), (row 1, column 2), and so on. Unfortunately, looking up the actual values of the elements of `Y` as `Y(indRow,indCol)` does not work.

CREATING USER-DEFINED FUNCTIONS

The compactness of the function syntax makes functions desirable when a user needs to call a certain sequence of commands often. For example, the Black-Scholes formula for pricing European options takes a number of steps to compute. It is convenient to have a function that returns one value—the option price—to the user after the user inputs values of factors that determine that price, such as the strike price, the time to maturity, the volatility, and so on.

Functions need to be written in M-files. Although general script M-files can contain any sequence of instructions that will be completed

when the name of the file is typed at the MATLAB prompt, function M-files need to start with a specific first line. That line contains the word “function” and a declaration of the function name, inputs, and outputs. The function name and the name of the M-file should be the same.

The Black-Scholes formula already exists in the Financial Toolbox, so it is convenient to see how the price is computed and discuss important aspects of writing user-defined functions. (We have skipped some lines in the code for the sake of brevity.) Users can view the source code for some of the advanced MATLAB functions in the toolboxes by entering *type function name* at the prompt.

```
>> type blsprice
function [call,put] = blsprice(S, X, r, T, sig, q)
% BLSPRICE Black-Scholes put and call option pricing.
% Compute European put and call option prices using a Black-Scholes model.
%
% [Call,Put] = blsprice(Price, Strike, Rate, Time, Volatility)
% [Call,Put] = blsprice(Price, Strike, Rate, Time, Volatility, % Yield)
%
% Optional Input: Yield
%
% Inputs:
% Price - Current price of the underlying asset.
% Strike - Strike (i.e., exercise) price of the option.
% Rate - Annualized continuously compounded risk-free rate of
% return over the life of the option, expressed as a positive decimal number.
% Time - Time to expiration of the option, expressed in years.
% Volatility - Annualized asset price volatility (i.e., annualized
% standard deviation of the continuously compounded asset return),
% expressed as a positive decimal number.
% Optional Input: Yield - Annualized continuously compounded yield of the
% underlying asset over the life of the option, expressed as a decimal
% number. If Yield is empty or missing. the default value is zero. For
% example, this could represent the dividend yield (annual dividend rate
% expressed as a percentage of the price of the security) or foreign
% risk-free interest rate for options written on stock indices and
% currencies, respectively.
%
```

```

% Outputs:
% Call      - Price (i.e., value) of a European call option.
% Put      - Price (i.e., value) of a European put option.
...
%
...

% Copyright 1995-2005 The MathWorks, Inc.
% $Revision: 1.8.2.5 $   $Date: 2005/09/18 16:19:06 $

%
% Input argument checking & default assignment.
%

if nargin < 5
    error('Finance:blsprice:InsufficientInputs', ...
        'Specify Price, Strike, Rate, Time, and Volatility.')
end

if (nargin < 6) || isempty(q)
    q = zeros(size(S));
end

message = blscheck('blsprice', S, X, r, T, sig, q);
error(message);
%
% Perform scalar expansion & guarantee conforming arrays.
%

try
    [S, X, r, T, sig, q] = finargsz('scalar', S, X, r, T, sig, q);
catch
    error('Finance:blsprice:InconsistentDimensions', ...
        'Inputs must be scalars or conforming matrices.')
end

%
% Record array dimensions for output argument formatting.
%

[nRows, nCols] = size(S);

call = nan(nRows * nCols, 1);
put  = nan(nRows * nCols, 1);

%

```

```

% Convert to column vectors for intermediate processing.
%
[S, X, r, T, sig, q] = deal(S(:), X(:), r(:), T(:), sig(:), q(:));

%
% Enforce some boundary conditions that produce warnings (e.g., logarithm
% of zero and divide by zero) and potential NaN's in the output option
% price arrays:
%
% (1) At expiration (i.e., T = 0), the price of all options is simply the
% greater of their intrinsic value and zero.
%
% (2) When the price of the underlying asset is zero (i.e., S = 0), the value
% of a call option is zero and the value of a put option is equal to its
% present value of the strike price (X). This boundary condition enforces
% the "absorbing barrier" property associated with the geometric Brownian
% motion diffusion process governing the price path of the underlying
% asset (S).
%
% (3) When the strike price is zero (i.e., X = 0), the value of a put option
% is zero and the value of a call option is equal to the price of the
% underlyer (S).
%

isTimeZero      = (T == 0);          % Expired options.
call(isTimeZero) = max(S(isTimeZero) - X(isTimeZero), 0);
put (isTimeZero) = max(X(isTimeZero) - S(isTimeZero), 0);

isStockZero     = (S == 0);
call(isStockZero) = 0;                % Worthless calls.
if any(isStockZero)
    put(isStockZero) = X(isStockZero) .* exp(-r(isStockZero).*T(isStockZero));
end
isStrikeZero    = (X == 0);
call(isStrikeZero) = S(isStrikeZero);
put (isStrikeZero) = 0;                % Worthless puts.

%
% Suppress a divide by zero warning ONLY for zero volatility conditions. Other
% warnings could be valuable.
%

state = warning; % Store the current state.

if any(sig == 0)
    warning('off', 'MATLAB:divideByZero')
end

```

```

%
% Now apply the general Black-Scholes European option pricing formulae,
% excluding the boundary cases handled above, and explicitly handling
% calculations that produce 0/0 = NaN's for the parameters of the
% cumulative normal distribution function (i.e., d1 & d2).
%
% NaN's occur when S = X, r = q, and Sigma = 0. This situation corresponds to
% at-the-money options written on riskless underlying assets. Such assets
% should earn the risk-free rate less the dividend yield. But when r = q, the
% net growth rate is also zero, resulting in 0/0 = NaN.
%

i = ~(isTimeZero | isStockZero | isStrikeZero);

d1 = log(S(i)./X(i)) + (r(i) - q(i) + sig(i).^2/2) .* T(i);
d1 = d1 ./ (sig(i).*sqrt(T(i)));
d2 = d1 - (sig(i).*sqrt(T(i)));

d1(isnan(d1)) = 0;
d2(isnan(d2)) = 0;

call(i) = S(i) .* exp(-q(i).*T(i)) .* normcdf( d1) - ...
    X(i) .* exp(-r(i).*T(i)) .* normcdf( d2);
put (i) = X(i) .* exp(-r(i).*T(i)) .* normcdf(-d2) - ...
    S(i) .* exp(-q(i).*T(i)) .* normcdf(-d1);

warning(state) % Restore the state.

%
% Reshape the outputs for the user.
%

call = reshape(call, nRows, nCols);
put = reshape(put , nRows, nCols);
% [EOF]

```

Some aspects of this function are very complicated for a beginner, but a review of the function syntax helps create a list of useful pointers to which you can refer when creating your own functions:

- The first line contains the word `function` followed by a specification of the outputs of the function `call` (in this case, `[call,put]`).

Note that a function can have more than one output. After calling the function, MATLAB computes the values for the outputs, and the variable `call` will contain the price of a European call option, while the variable `put` will contain the price of a European put option. Next, we have an equal sign followed by the name of the function (`blsprice`) and the arguments for the function (`S` for current stock

price, X for strike price, r for rate of return, T for time to maturity, sig for volatility, as well as the optional argument `yield` for continuous dividend yield).

- When the function is called with specific input values, you can assign the output to variables. For example,

```
>> [callOutput,putOutput]
= blsprice(110,100,0.10,2,0.40)
callOutput =
38.1757
putOutput =
10.0488
```

- The names of the input variables need to participate in calculations in the function. For example, S appears as the current stock price in the first line (`function [call,put] = blsprice(S, X, r, T, sig, q)`), and this is the same variable that is used to store the value of the stock price in the computations. Similarly, the names of the output variables (`call` and `put`) should appear somewhere in the text of the function and be assigned an expression, which can then be returned to the user.
- Note the abundance of the percentage sign (%) in the function code. This sign is used for writing comments that are ignored by MATLAB when executing the code. It is always a good idea to comment abundantly in order to be able to retrace your reasoning later. The first comment line is called “the H1 line,” and it is the line that is searched by the MATLAB built-in function `lookfor`. `lookfor` searches all MATLAB files containing a keyword in their first line. (This is useful if you are not sure which function to use for a specific purpose, and you would like to find the names of all functions that may be relevant.) Therefore, it is important to provide a meaningful description of your function in the first commented line. After the first line, you can continue with a more detailed description of the function and list references.

CONTROL FLOW STATEMENTS

M-files, whether of a generic or function kind, can contain more advanced operations than matrix manipulation. Next, we briefly review a couple of control flow statements that are often used in such files: the `for` loop and the `if` statement.

The general format of a `for` loop is

```
for n = array
commands
end
```

The commands inside the `for` loop are executed once for every value in the column in the array. (Typically, the array is a vector of numbers, so the loop is executed once for every number.) For example,

```
for n = 1:5
v(n) = sqrt(n);
end
```

results in

```
v =
1.0000 1.4142 1.7321 2.0000 2.2361
```

The array `1:5` is equivalent to `[1 2 3 4 5]`. MATLAB starts out with $n = 1$, computes its square root, and assigns it to `v(1)`. Then, it keeps repeating the process until it has computed `v(5)` for $n = 5$.

Loops in MATLAB are often necessary, but as a general rule MATLAB is more efficient in array operations than in loops. For example, the same effect (adding 10 to each element of the vector x) can be achieved in two ways:

```
for n = 1:4
x(n) = x(n) + 10;
end
```

and

```
>> x = x+10
```

Both of them result in

```
x =
12 13 14 16
```

The second command would typically be completed faster. Loops are not as inefficient as they used to be in older versions of MATLAB, however—the difference in speed between the two approaches has been greatly reduced in the latest versions of the software.

The `if` statement has the following general format:

```
if expression
commands
end
```

The commands are completed only if all elements in the expression are true. A somewhat more complex `if` statement is

```
if expression1
commands1
elseif expression2
commands2
else expression3
commands3
end
```

Commands1 are completed if `expression1` is true. If `expression1` is not true, MATLAB moves on and checks if `expression2` is true. If `expression2` is true, `commands2` are completed. If `expression2` is not true either, MATLAB moves to `expression3`. If `expression3` is true, `commands3` are completed; otherwise MATLAB exits. The `elseif` or `else` commands are optional in `if` statements.

There are several other useful control flow statements, such as the `while` loop, `switch-case` constructions, and `try-catch` blocks. See the MATLAB manual and help for more detail.

GRAPHS

MATLAB is well known for its beautiful graphing capabilities. The most common function for plotting two-dimensional (2-D) graphs is `plot`.

To illustrate how `plot` works, suppose we would like to plot the standard normal prob-

ability distribution. We will use the function `normpdf` (available from the Statistics Toolbox), which computes the probability density function (PDF) of a normal random variable.

The command

```
>> x = linspace(-6,6,100)
```

creates a vector `x` with 100 values, equally spaced between the minimum value -6 and the maximum value $+6$. (In reality, the normal distribution stretches from negative infinity to positive infinity, but it is highly unlikely that we will obtain realizations that are greater than 6 standard deviations away from the mean of 0, so we focus on plotting the center of the distribution.)

The command

```
>> y = normpdf(x)
```

computes the values of the normal probability distribution function for every value in the array for `x`.

To plot `x` versus `y`, use

```
>> plot(x,y)
```

The result is the graph in Figure 2.

You can play with the options for the graph. For example,

```
>> plot(x,y,'r:p'); title('Normal PDF');
xlabel('x'); ylabel('pdf')
```

plots the same graph as a red dotted line with a pentagram symbol, labels the horizontal (`x`) and the vertical (`y`) axes, and creates a title for the graph (see Figure 3).

To plot multiple graphs on the same picture, use the command `hold on` before you start and `hold off` when you are done with the instructions. For example, suppose we would like to plot the standard normal distribution and a standard `t`-distribution with 5 degrees of freedom on the same graph in order to compare them. The following sequence of commands accomplishes this.

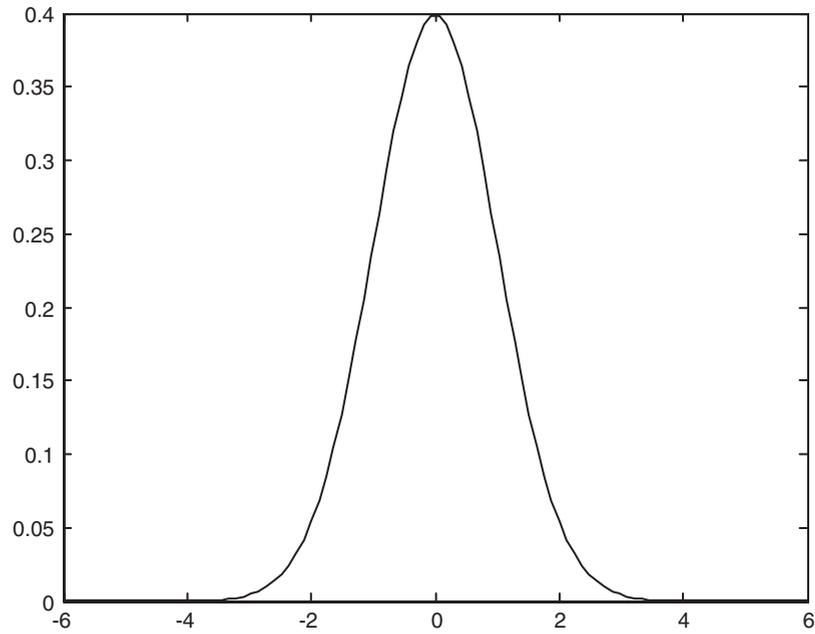


Figure 2 A Plot of the PDF of the Normal Distribution

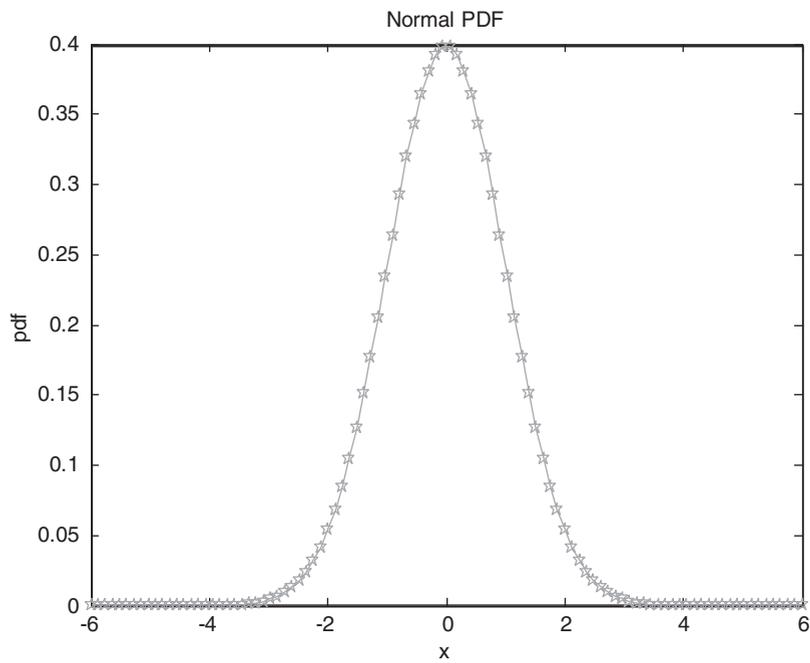


Figure 3 A Plot of the PDF of the Normal Distribution (with Modified Options)

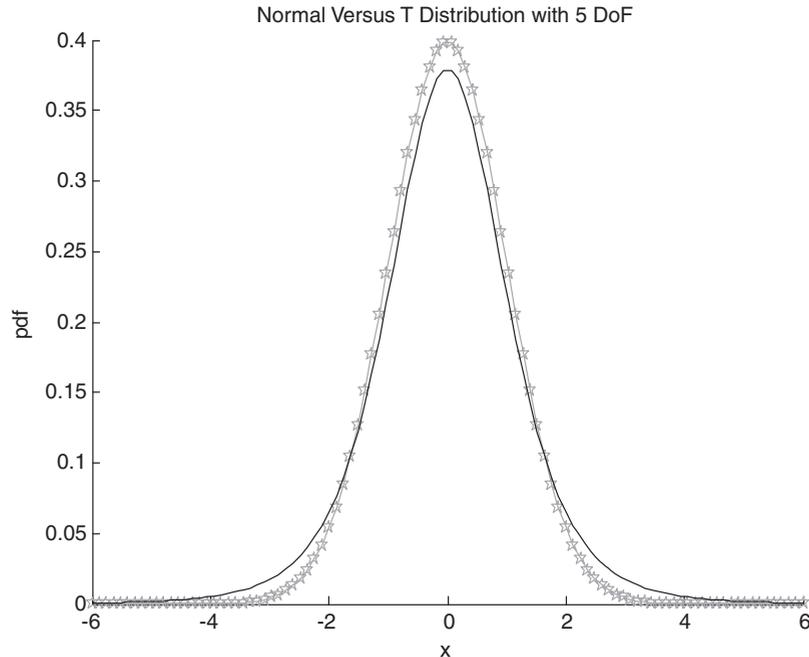


Figure 4 Illustration of hold on / hold off Effect

First, we declare a variable that follows a t-distribution with 5 degrees of freedom:

```
>> t=tpdf(x,5);
```

Then, we plot the graph:

```
>> hold on
>> plot(x,y,'r:p'); xlabel('x');
    ylabel('pdf')
>> plot(x,t);
>> title('Normal Versus T Distribution');
>> hold off
```

The results are displayed in Figure 4.

Alternatively, you can list several pairs of variables inside the plot function. For example,

```
>> plot(x,y,'r:p',x,t); xlabel('x');
    ylabel('pdf')
>> legend('Normal PDF','T PDF')
>> title('Normal Versus T Distribution
    with 5 DoF');
```

This script also creates a legend (Figure 5).

Legend, titles, and other graph attributes can be added and modified also after the basic plot command has been given and a graph window has popped up. To modify an existing graph's options, click on the corresponding items in the top menu of the graph window.

Suppose now that we would like to plot the two PDFs side by side in the same figure. To graph several separate graphs in the same figure, use the command subplot(number of rows, number of columns, index of graph within the graph array).

For example, the code

```
>> subplot(1,2,1), plot(x,y,'r:p');
    xlabel('x'); ylabel('pdf')
>> title('(a) Normal PDF')
>> subplot(1,2,2), plot(x,t);
    xlabel('x'); ylabel('pdf')
>> title('(b) T PDF')
```

produces the graph in Figure 6.

Finally, we briefly discuss three-dimensional (3-D) graphs. They can be created with

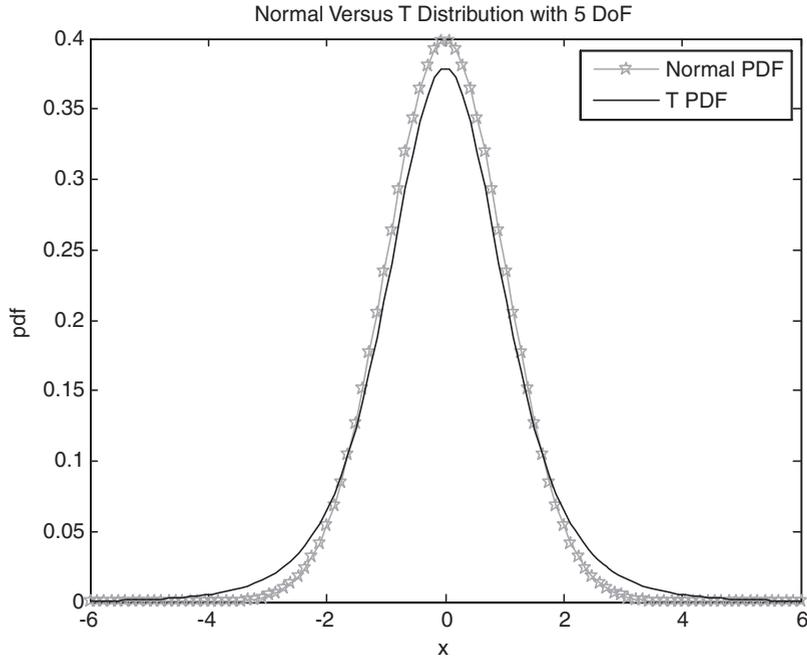


Figure 5 Changing Defaults and Plotting Multiple Graphs with the plot Function

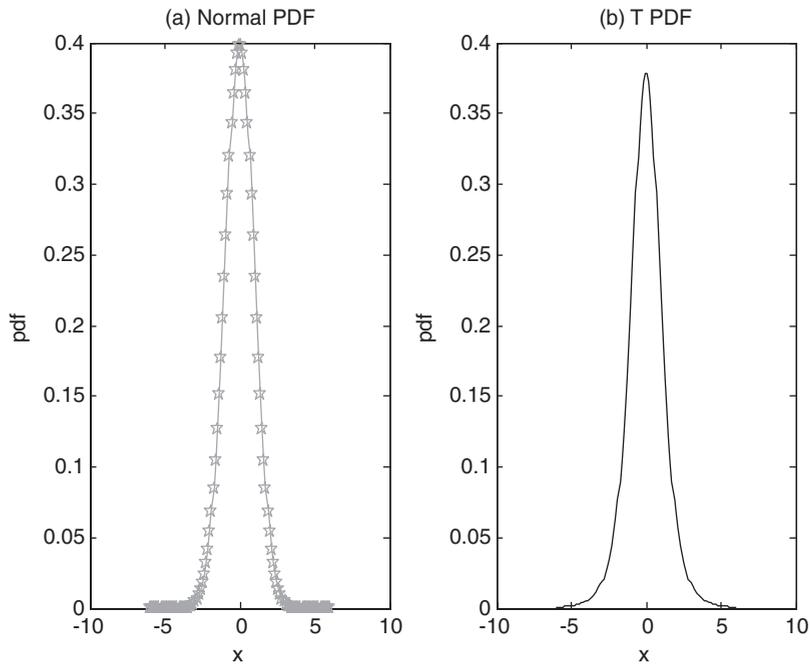


Figure 6 Multiple Plots within the Same Figure

commands like `plot3` and `surf`, and as a general matter are more complex.

The command `plot3`(first variable x , second variable y , third variable z) plots points in 3-D space whose three coordinates are given by the vectors or matrices (x, y, z) in the three arguments of the function. The arguments need to be arrays of equal sizes.

The command `surf`(x, y, z) plots a shaded surface using z as the height and (x, y) as the vectors or matrices that define the other two dimensions of the surface. When x and y are vector arrays, as is the case in most financial applications, the number of rows for z should be the length of the vector array y , and the number of columns for z should be the length of the vector array x .

For example, suppose we would like to plot a multivariate normal distribution function for two normal variables, x_1 and x_2 , that have means of 0 and are correlated with covariance matrix $[0.25 \ 0.3; 0.3 \ 1]$. (Note that this notation means that the variance of x_1 is 0.25 (the standard deviation of x_1 is 0.5), the variance of x_2 is 1 (the standard deviation of x_2 is 1), and the covariance of x_1 and x_2 is 0.3.

The multivariate normal distribution function can be computed with the MATLAB function `mvnpdf`(X, μ, Σ). The arguments μ and Σ are the vector array of average (expected) values for the normal random variables and their covariance matrix, respectively. In this case, we have two normal random variables, so $\mu = [0 \ 0]$ and $\Sigma = [0.25 \ 0.3; 0.3 \ 1]$. The first argument in the function (matrix X) provides the points at which the function should be evaluated. The function is evaluated for every row of X , taking the elements in that row as the coordinates of the point at which the function should be evaluated. Therefore, since in our example we are looking at two normal random variables, there should be two columns of the matrix X . We cannot simply provide two columns with, say, equally spaced values for x_1 and x_2 . If we do,

MATLAB would pair each entry of x_1 with the corresponding entry of x_2 , and will only use those combinations of coordinates, so the plot will look two-dimensional. The columns of X should provide a grid. In other words, we cannot simply provide possible coordinates along each axis and expect that MATLAB will know to take every combination of possible coordinates to obtain the points at which to plot the function. To create this grid of points, we need to go through a couple of steps.

First, we would use the function `[X1, X2] = meshgrid(x1, x2)`. It creates two matrices. The number of rows in the first matrix, X_1 , is the same as the number of elements in the vector y (i.e., the number of rows equals `length(y)`, another useful MATLAB command). Each row of the column X_1 contains identical entries: the entries of the vector x . The matrix X_2 contains the same number of columns as the number of elements in the vector x , and each column contains an identical copy of the vector y . While perhaps difficult to imagine at first, $X_1(i, j)$ and $X_2(i, j)$ cover all possible combinations of the elements of the original vectors, x and y .

The second step is to create the array `[X1(:), X2(:)]`. The colon operator `(:)` has multiple uses, but in the context of being used as an argument for a matrix, it takes all entries of a matrix, column by column, and lists them as a vector array. Therefore, the array `[X1(:), X2(:)]` would contain two columns with every possible combination of coordinates generated by the original list in the vector arrays x and y .

To summarize, here are the commands used to generate 30 points between -4 and 4 along each coordinate x_1 and x_2 , then to evaluate the multivariate normal PDF at each combination of coordinates:

```
>> x1 = linspace(-4,4,30);
    x2 = linspace(-4,4,30);
>> Sigma = [0.25 0.3; 0.3 1]; mu = [0 0];
>> [X1,X2] = meshgrid(x1,x2);
>> z = mvnpdf([X1(:),X2(:)],mu,Sigma);
```

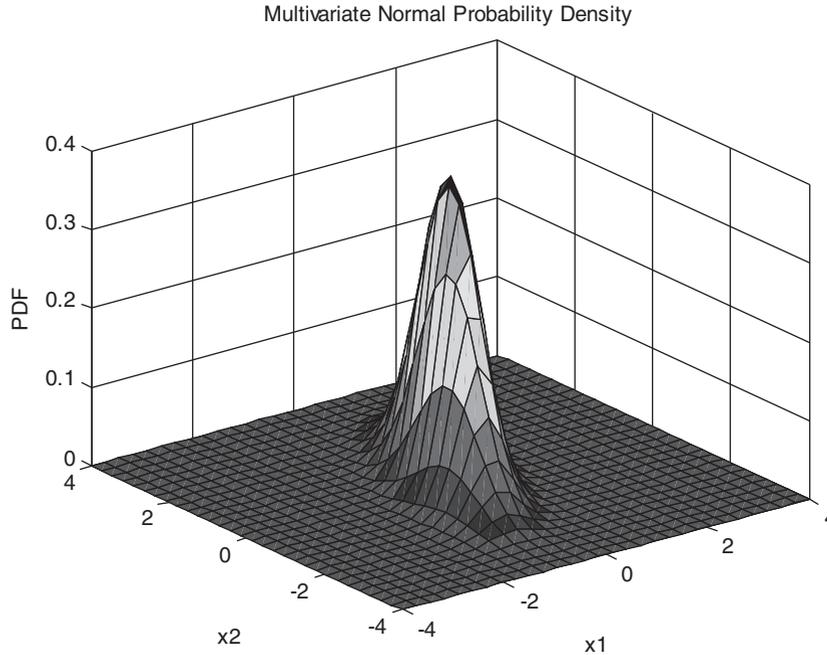


Figure 7 Three-Dimensional Plot of a Multivariate Normal Distribution

The output of this sequence of commands is a vertical array of values that represent the multivariate normal PDF evaluated at each combination of coordinates. (If you skip the semicolon at the end of the last row with the function `mvnpdf`, you can see what the output looks like. You can also use the command `size(z)` to check the dimensions of `z`.) Now we would like to plot these values. We will use the `surf` function.

The `surf` function's third argument, `z`, needs to be a matrix whose entries represent the values of the function to be plotted at each combination of coordinates. However, we obtained a vector of values for the PDF. We need to "reshape" that vector back into a matrix. This can be done with the command `Z = reshape(z, m, n)`. The function `reshape` takes the array `z` and goes through the elements of `z` columnwise. The first `m` elements of `z` become the first column of the new matrix `Z`, the next `m` elements of `z` become the second column of the matrix `Z`, and so forth until `n` columns for `Z` are created. In this example, we would like to create

`length(x1)` columns and `length(x2)` rows. (This may be a bit confusing, but, as we mentioned earlier, the function `surf` expects the third argument to be a matrix with the number of columns equal to the size of the first argument, and the number of rows equal to the size of the second argument.)

```
>> Z = reshape(z, length(x2), length(x1));
>> surf(x1, x2, Z);
>> title('Multivariate Normal
    Probability Density');
>> axis([-4 4 -4 4 0 0.4]);
>> xlabel('x1'); ylabel('x2');
    zlabel('PDF');
```

The resulting graph is in Figure 7.

IMPORTING DATA AND INTERACTING WITH SPREADSHEETS

MATLAB recognizes files with the extension `.dat` as data files. Such files should contain text structured in rows and columns. For example,

suppose that the file `returns.dat` contains daily annual returns on the stocks traded in the NYSE for 10 years. The command

```
>> load returns.dat
```

imports the data in the file into a data structure—a matrix array with rows and columns that can then be referenced using some of the commands we described earlier.

Many financial companies build their infrastructure around Microsoft Excel. The MATLAB core product contains some useful functions for importing Excel data and exporting MATLAB results to spreadsheets. The function

```
>> xlsread('fileName', 'sheetName',
           'range')
```

allows the user to read into MATLAB the data stored in file `fileName`, worksheet `sheetName`, cells in range `range`. Instead of a range in the spreadsheet, you can state an array name if you had named the array of cells in advance. Variations of this command exist; for instance

```
>> xlsread('fileName', -1)
```

allows the user to select the range in `fileName` directly, through interactive selection in Excel. Type `help xlsread` at the MATLAB command prompt for further information.

The function

```
>> xlswrite('fileName', output, 'sheetName',
           'cell')
```

allows the user to export MATLAB results (`output`) to a worksheet (`sheetName`) in an Excel file (`fileName`). MATLAB preserves the dimensions of the `output` and writes it to the spreadsheet starting at cell reference `cell`. For example, if `output` is a horizontal array of numbers, MATLAB will write the data in a row in the Excel file, starting at `cell`.

MATLAB operations work within the `xlswrite` command. For example, you can switch the array dimensions (transpose) the

output by using `output'` inside the parentheses of the `xlswrite` command if you desire different output formatting in the Excel spreadsheet.

More sophisticated capabilities exist through MATLAB's Excel Link. With Excel Link, you can call MATLAB's functions directly from within Excel, thus ensuring access to MATLAB's superior computational and graphical capabilities. Excel Link is purchased as a separate toolbox. It can then be made visible from within Excel by selecting it as one of Excel's Add-Ins. There are 11 commands (they all start with "ML") that allow for communicating data back and forth between Excel and MATLAB. For example, `=MLAppendMatrix()` creates or appends a matrix in MATLAB with data from an Excel spreadsheet.

A word of caution: Excel Link formulas are not case sensitive. For example, `MLAppendMatrix` and `mlappendmatrix` are the same. However, MATLAB functions and variables called through these links are case sensitive. For example, `x` and `X` would still be treated as two separate variables.

EXAMPLES

This section discusses several *scripts* and *functions* in MATLAB that can be used in financial applications. The goal is to illustrate the use of toolboxes in MATLAB and to provide concrete examples of some of the tools introduced earlier in the entry.

Optimization in MATLAB

Optimization is an area in applied mathematics that, most generally, deals with efficient algorithms for finding an optimal solution among a set of solutions that satisfy given constraints. The first application of optimization in finance was suggested by Harry Markowitz in 1952, in a seminal paper that outlined his mean-variance optimization framework for optimal asset allocation. Some other classical problems in finance

Table 1 MATLAB Optimization Toolbox Functions/Solvers Appropriate for Specific Types of Optimization Problems

		OBJECTIVE				
		Linear	Quadratic	Least squares	Smooth nonlinear	Nonsmooth
CONSTRAINTS	None	N/A	quadprog	\, isqcurvefit, isqnonlin	fminsearch, fminunc	fminsearch, *
	Bound	linprog	quadprog	isqcurvefit, isqlin, isqnonlin, isqnonne	fminbnd, fmincon, fseminf	*
	Linear	linprog	quadprog	isqlin	fmincon, fseminf	*
	Smooth nonlinear	fmincon	fmincon	fmincon	fmincon, fseminf	*
	Discrete	bintprog				

Note: Asterisk (*) is used to denote solvers that are available only through the Global Optimization Toolbox. Blank entries mean that there is currently no solver available. Technically, the Global Optimization Toolbox can be used for solving discrete problems as well; however, it requires additional programming.

that can be solved by optimization algorithms include:

1. Is there a possibility to make riskless profit given market prices of related securities?
2. How should trades be executed so as to reach a target allocation with minimum transaction costs?
3. Given a limited capital budget, which capital budgeting projects should be selected?
4. Given estimates for the costs and benefits of a multistage capital budgeting project, at what stage should the project be expanded/abandoned?

MATLAB’s Optimization Toolbox contains solvers for a range of optimization problems. MATLAB expects optimization formulations to be passed to its solvers in an array form and has functions that call specific solvers for specific types of optimization problems. (See Table 1 for a quick overview. See also MATLAB’s help for a complete listing.) If the Global Optimization Toolbox is available, the range of solvers is expanded to include randomized search algorithms.

The most often used solver in MATLAB is `fmincon`, which is the solver for general nonlinear optimization. However, if you know the type of problem you are trying to solve, you are always better off giving the optimization soft-

ware as much information as you can in order to make the optimization process more accurate and efficient. In financial applications, you are most likely to encounter situations in which you need `linprog` (a linear programming solver), `quadprog` (a quadratic programming solver), `bintprog` (a binary programming solver), and randomized search algorithms, such as `simulannealbnd` and `ga`.

We will use `linprog` and `quadprog` to solve two examples of portfolio allocation problems. Before we show the actual implementation, we need to explain how solvers are actually called in MATLAB. There are two ways to call the solvers: as functions directly from the command prompt (equivalently, from within M-files), or through the optimization tool.

The MATLAB optimization tool provides an interface between the solvers and the user. While using such an interface may not be optimal when solving sequences of optimization problems, as in the case of dynamic programming or stochastic programming, it is quite convenient when solving a single optimization problem, because it lists all available solvers, prompts the user for the different inputs that the optimization solvers expect, and allows for easy manipulation of the options. Options can be specified directly when a solver is called from the command prompt as well, but that is more difficult for MATLAB beginners.

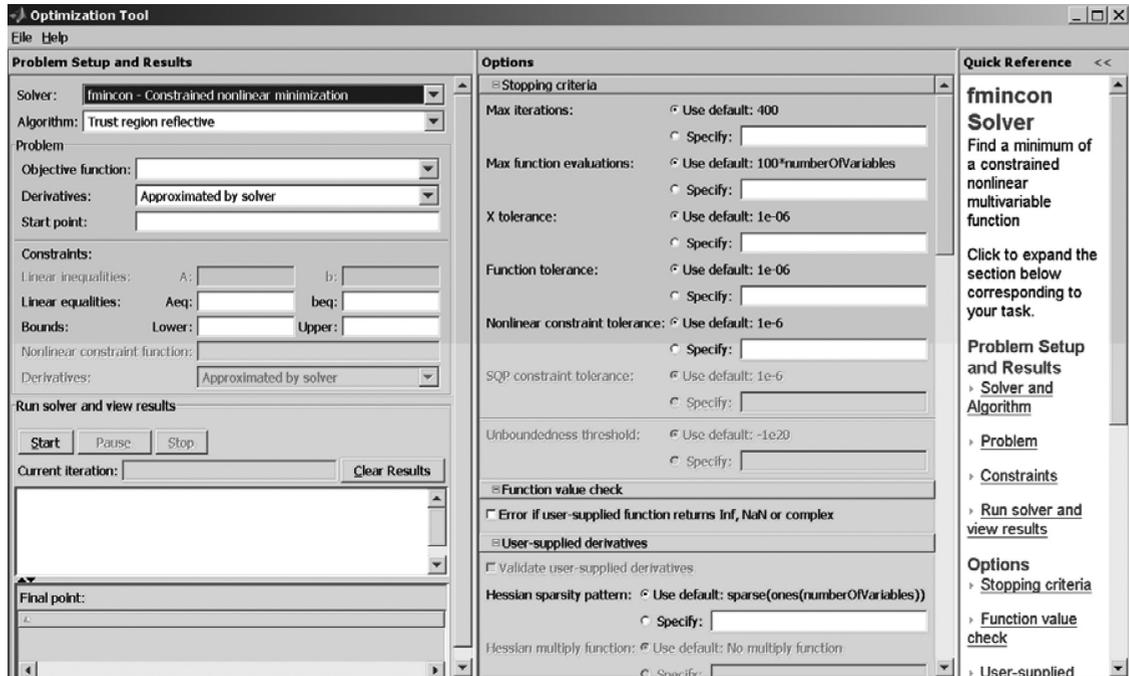


Figure 8 The Optimization Tool Interface in MATLAB

The optimization tool is called by typing `optimtool` at the MATLAB command prompt. The optimization tool dialog box is shown in Figure 8. The panel on the left-hand side is dedicated to the specification of the inputs: the type of solver that needs to be called, the arrays with the problem data, the starting point, and so on. The panel in the middle allows for changing the level of tolerance in the search for the optimal solution. For example, the Function tolerance is currently set at the default value of $1e-06$, which is 10^{-6} . This means that the selected algorithm will continue to iterate through solutions until the improvement in successive objective function values becomes smaller than 10^{-6} . Sometimes, such level of accuracy is unnecessary. For example, if our objective function is measured in dollars and cents (e.g., we are maximizing dollar return as in the simple portfolio allocation example we will discuss next), then technically we do not need precision beyond 2–3 digits after the decimal point. Therefore, we can speed up the algorithm by relaxing the re-

quirements on tolerance. Other useful options include level of display (whether to show iterations of the optimization algorithm or not) and function plots at intermediate stages.

Linear Optimization: Simple Portfolio Allocation

Let us consider a specific example to illustrate the use of the optimization function `linprog`. (For more details, see section 5.3.1 in Pachamanova and Fabozzi, 2010.)

The portfolio manager at a large university in the United States is tasked with investing a \$10 million donation to the university endowment. He has decided to invest these funds only in mutual funds and is considering the following four: an aggressive growth fund (Fund 1), an index fund (Fund 2), a corporate bond fund (Fund 3), and a money market fund (Fund 4), each with a different expected annual return and risk level. (The risk level measurement is deliberately simplified for the sake of this example.) The investment guidelines established

Table 2 Data for the Portfolio Manager’s Problem

Fund type	Growth	Index	Bond	Money market
Fund #	1	2	3	4
Expected return	20.69%	5.87%	10.52%	2.43%
Risk level	4	2	2	1
Max investment	40%	40%	40%	40%

by the Board of Trustees limit the percentage of the money that can be allocated to any single type of investment to 40% of the total amount. The data for the portfolio manager’s task are provided in Table 2. In addition, in order to contain the risk of the investment to an acceptable level, the amount of money allocated to the aggressive growth and the corporate bond funds cannot exceed 60% of the portfolio, and the aggregate average risk level of the portfolio cannot exceed 2. What is the optimal portfolio allocation for achieving the maximum expected return at the end of the year, if no short selling is allowed?

The vector of decision variables for this optimization problem can be defined as $x = (x_1, x_2, x_3, x_4)$: amounts (in \$) invested in Fund 1, 2, 3, and 4, respectively.

Let the vector of expected returns be $\mu = (20.69\%, 5.87\%, 10.52\%, 2.43\%)$. Then, the objective function can be written as

$$f(x) = \mu'x = (20.69\%) \cdot x_1 + (5.87\%) \cdot x_2 + (10.52\%) \cdot x_3 + (2.43\%) \cdot x_4$$

It represents the optimal expected dollar amount at the end of the year.

There are also several constraints.

1. The total amount invested should be \$10 million. This can be formulated as $x_1 + x_2 + x_3 + x_4 = 10,000,000$.
2. The total amount invested in Fund 1 and Fund 3 cannot be more than 60% of the total investment (\$6 million). This can be written as

$$x_1 + x_3 \leq 6,000,000$$

3. The average risk level of the portfolio cannot be more than 2. This constraint can be expressed as

$$4 \cdot (\text{proportion of investment with risk level 4}) + 2 \cdot (\text{proportion of investment with risk level 2}) + 1 \cdot (\text{proportion of investment with risk level 1}) \leq 2$$

or, mathematically,

$$\frac{4 \cdot x_1 + 2 \cdot x_2 + 2 \cdot x_3 + 1 \cdot x_4}{x_1 + x_2 + x_3 + x_4} \leq 2$$

In this particular example we know that the total amount $x_1 + x_2 + x_3 + x_4 = 10,000,000$, so the constraint can be formulated as

$$4 \cdot x_1 + 2 \cdot x_2 + 2 \cdot x_3 + 1 \cdot x_4 \leq 2 \cdot 10,000,000$$

1. The maximum investment in each fund cannot be more than 40% of the total amount (\$4,000,000). These constraints can be written as

$$x_1 \leq 4,000,000, x_2 \leq 4,000,000, x_3 \leq 4,000,000, x_4 \leq 4,000,000.$$

2. Given the no short selling requirement, the amounts invested in each fund cannot be negative. These are nonnegativity constraints: $x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0$.

The final optimization formulation can be written in matrix form. The objective function is

$$\max_{x_1, x_2, x_3, x_4} [0.2069 \quad 0.0587 \quad 0.1052 \quad 0.0243] \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

Let us organize the constraints into groups according to their signs. This will be useful when we input the constraints into MATLAB.

$$\text{Equality(=)} : [1 \quad 1 \quad 1 \quad 1] \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = 10,000,000$$

$$\text{Inequality(≤)} : \begin{bmatrix} 1 & 0 & 1 & 0 \\ 4 & 2 & 2 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \leq \begin{bmatrix} 6,000,000 \\ 20,000,000 \\ 4,000,000 \\ 4,000,000 \\ 4,000,000 \\ 4,000,000 \end{bmatrix}$$

$$\text{Nonnegativity}(\geq): \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

This is a linear optimization problem because all constraints and the objective function are linear. To solve linear optimization problems with MATLAB, use `linprog (f,A,b,Aeq,beq,lb,ub)`. The function arguments `f,A,b,Aeq,beq,lb,ub` correspond to the following

LP formulation:

$$\begin{array}{ll} \min_x & f'x \\ \text{s.t.} & Ax \leq b \\ & Aeq \cdot x = beq \\ & lb \leq x \leq ub \end{array}$$

Therefore, before calling `linprog`, you need to write the problem formulation in this particular form. We include the complete MATLAB script below.

```

1  numAssets = 4;
2  expReturnsVec = [0.2069 0.0587 0.1052 0.0243]';
3  %create placeholders for an array of decision variables
4  %(amounts to invest in
5  %each fund) and the optimal portfolio expected return (to be filled out
6  %after the optimization)
7
8  amountsVec = zeros(numAssets,1);
9  optReturn = [];
10
11 %vector of coefficients of objective function f since MATLAB expects
12 %minimization (and we are maximizing), take the negative of the function
13 %we are trying to maximize
14 f = -expReturnsVec;
15
16 %A, matrix of coefficients in constraints with inequalities so that
17 %Ax<=b
18 A = [1 0 1 0;
19      4 2 2 1;
20      1 0 0 0;
21      0 1 0 0;
22      0 0 1 0;
23      0 0 0 1];
24
25 %b
26 b = [6000000 20000000 4000000 4000000 4000000 4000000]';
27
28 %Aeq, matrix of coefficients in constraints with equalities so that
29 %Aeq*x=beq
30 Aeq = ones(1,numAssets);
31
32 %beq
33 beq = 10000000;
34
35 %lower bounds: nonnegativity requires that all decision variables are >= 0

```

```

36 lb = zeros(numAssets,1);
37
38 %upper bounds can be left infinite (although, technically, we cannot invest
39 %more than the $10m we have available)
40 ub = inf*ones(numAssets,1);
41
42 [amountsVec,optReturn] = linprog(f,A,b,Aeq,beq,lb,ub);
43
44 format('bank');
45
46 amountsVec
47 %revert to correct number for maximum return (reverse sign)
48 optReturn = -optReturn

```

The process for formulating the optimization problem is as follows. First, we ask ourselves what corresponds to the vector of decision variables x in the `linprog` formulation. In our example, x maps directly to the vector of amounts to invest in each asset. We then enter problem data, such as the expected returns vector `expReturnsVec`. We allocate empty arrays to store the values of the optimal solution `amountsVec` and the optimal value of the objective function `optReturn` after collecting the information from the solver.

Next, we create the input data for the `linprog` solver. The solver expects a vector of objective function coefficients f , which in our case is the vector of expected returns on the different assets. Note, however (line 14), that we specify f as `-expReturnsVec`. This is because MATLAB expects a minimization problem, and our objective function is to maximize expected revenue, so we need to convert our problem to the required form by minimizing the negative of the expression for the maximization objective. At the end (line 48), we take the negative of the optimal value for expected return found by the solver, so that we arrive at the actual optimal value for the maximization problem. The optimal values of the decision variables, which in this case are the amounts to invest, `amountsVec`, do not need to be modified af-

ter the optimization results are returned by the solver.

Lines 14–40 contain the specification of the other inputs in the problem. Note that we are in fact using the matrices of coefficients for the groups of constraints (inequality, equality, and nonnegativity) that we defined earlier. Namely, A (lines 18–23) is the matrix of left-hand-side inequality constraint coefficients; A_{eq} (line 30) is the matrix of left-hand-side equality constraint coefficients; b (line 26) is the vector of right-hand-side coefficients of the inequality constraints; and beq (line 33) is the vector of right-hand-side coefficients of the equality constraints (in our example, we have only one equality constraint). The lower bounds, lb (line 36), are the zeros from the right-hand-side of the nonnegativity constraints on the decision variables, so we create a vector array with size equal to the number of decision variables that contains only zeros. We have explicit upper bounds of \$4,000,000 on each decision variable since we cannot invest more than that amount in each individual fund, so we could have stated those bounds as the input vector ub . However, these bounds have already been included in the matrix A , so we do not need to state them again. Instead, we state the individual upper bounds as infinity, that is, as the product of the number `inf` (in MATLAB, that denotes

infinity) and a vector of ones. (See line 40 of the code.)

An equivalent formulation of the constraints from MATLAB's perspective would have been to specify the arrays *A*, *beq*, and *ub* as

```
A = [1 0 1 0;
     4 2 2 1]
b = [6000000 20000000]'
ub = 4000000*ones(numAssets,1)
```

with all other input arrays remaining the same.

After all inputs have been specified, the `linprog` solver is called (line 42). The syntax in line 42 outputs requests that the output from the optimization be stored in the arrays we specified at the beginning, `amountsVec` and `optReturn`. The results are then printed to the screen and are formatted according to `format('bank')` (line 44), which basically rounds numbers to two decimal places.

After running the M-file, we obtain the following output:

```
amountsVec =
    2000000.00
         0.00
    4000000.00
    4000000.00
optReturn =
    931800.00
```

If you prefer to solve the problem by using the optimization tool for solving this problem, you need to fill out the dialog box as shown in Figure 9. Select `linprog` as the solver from the drop-down menu at the top. Under **Algorithm**, you can either leave the default (**Large Scale**), or select **Medium scale – simplex**, which is appropriate because our problem is quite small. We entered the names of the arrays that correspond to the objective function coefficients and the constraint coefficients in the corresponding fields in the left panel of the dialog box.

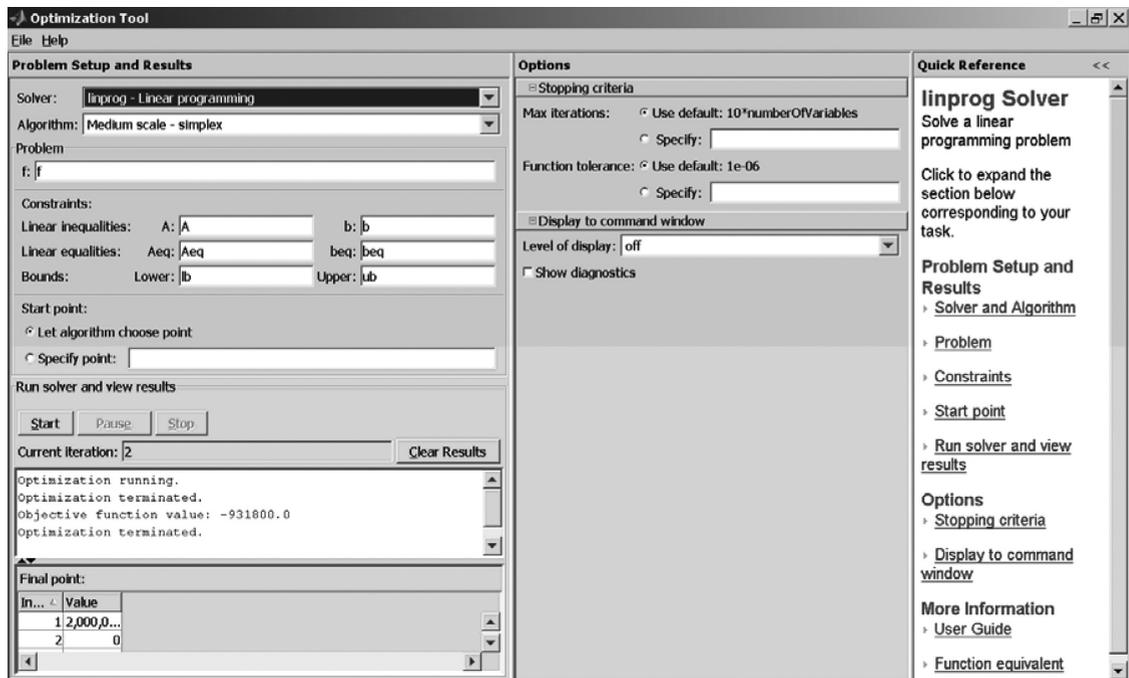


Figure 9 The Optimization Tool Dialog Box for the Portfolio Allocation Problem

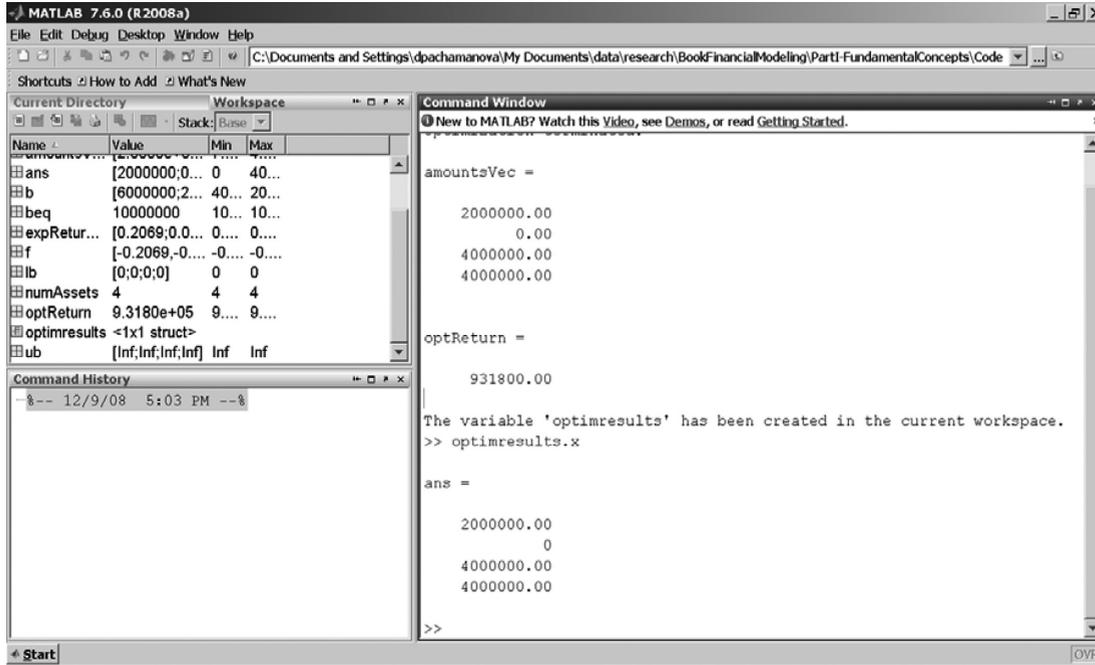


Figure 10 Handling the Structure of Optimization Results Exported from MATLAB's Optimization Tool

Note that these arrays must be prefilled; that is, they must be entered from the command prompt or read from a file before the problem is solved through the optimization tool; otherwise the solver will complain that these arrays are empty. You can make sure that the arrays f , A , b , A_{eq} , b_{eq} , lb , ub are filled in by checking first whether they are listed in the **Workspace** window at the upper left corner of the MATLAB desktop. Once all the input data are specified, click on the **Start** button in the left panel to solve the problem. The solution appears in the field below the **Start** button.

The optimization model can be saved as a script in an M-file by selecting **File | Generate M-file** from the main menu in the optimization tool. In addition, the optimization results can be exported to the workspace and further manipulated by selecting **File | Export to Workspace**. To export only the results, as opposed to the entire model, check **Export results to a MATLAB structure named: optimresults**. This creates a structure of results, `optimresults`, that

shows up in the **Workspace**. So, for example, the optimal solution (the portfolio allocation) can be read by typing `optimresults.x` at the command prompt. (See Figure 10.) Similarly, the optimal value of the objective function can be retrieved by typing `optimresults.fval` at the command prompt.

Quadratic Optimization: Mean-Variance Portfolio Allocation

The classical mean-variance portfolio optimization problem as introduced by Harry Markowitz (1952) is to minimize the variance of portfolio return subject to the constraint that the expected portfolio return is at a certain level. Let us consider a slight variation of the problem, in which we require that the expected return is at least at a certain level r_{target} . The mathematical formulation is

$$\begin{aligned}
 \min_w \quad & \mathbf{w}'\Sigma\mathbf{w} \\
 \text{s.t.} \quad & \mathbf{w}'\boldsymbol{\mu} \geq r_{\text{target}} \\
 & \mathbf{w}'\mathbf{1} = 1
 \end{aligned}$$

where \mathbf{w} is the vector of portfolio weights (to be determined), $\boldsymbol{\mu}$ is the vector of expected returns, $\boldsymbol{\Sigma}$ is the covariance matrix of returns, and $\mathbf{1}$ is a vector of ones of appropriate dimension.

The minimum variance portfolio allocation problem is a quadratic optimization problem with linear constraints. The `quadprog` function in MATLAB solves exactly problems of this kind:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \frac{1}{2} \mathbf{x}' \mathbf{H} \mathbf{x} + \mathbf{f}' \mathbf{x} \\ \text{s.t.} \quad & \mathbf{A} \mathbf{x} \leq \mathbf{b} \\ & \mathbf{A} \mathbf{e} \mathbf{q} \cdot \mathbf{x} = \mathbf{b} \mathbf{e} \mathbf{q} \\ & \mathbf{l} \mathbf{b} \leq \mathbf{x} \leq \mathbf{u} \mathbf{b} \end{aligned}$$

and is called with the command `quadprog(H, f, A, b, Aeq, beq, lb, ub)`.

It is easy to see how to match the two formulations:

- $\mathbf{x} = \mathbf{w}$
- $\mathbf{f} = \mathbf{0}$
- $\mathbf{H} = 2 \boldsymbol{\Sigma}$
- $\mathbf{A} = -\boldsymbol{\mu}'$
- $\mathbf{b} = -r_{\text{target}}$
- $\mathbf{A} \mathbf{e} \mathbf{q} = \mathbf{1}'$
- $\mathbf{b} \mathbf{e} \mathbf{q} = 1$
- $\mathbf{l} \mathbf{b} = -\text{infinity}$
- $\mathbf{u} \mathbf{b} = \text{infinity}$

```
numAssets = 2;
muVec = [9.1; 12.1];
SigmaMx = [272.25, -57.35;
           -57.35, 249.64];
```

```
targetReturn = 11;
```

```
%SINGLE OPTIMIZATION
```

```
%create the matrix X
H = 2*SigmaMx;
```

```
%create a vector of length numAssets with zeros
f = zeros(numAssets,1);
```

```
%create right hand and left hand side of inequality constraints
A = -transpose(muVec);
b = -targetReturn;
```

For example, the inequality constraint

$$\mathbf{w}' \boldsymbol{\mu} \geq r_{\text{target}}$$

in the mean-variance formulation is mapped to the inequality constraint assumed by the `quadprog` function

$$\mathbf{A} \mathbf{x} \leq \mathbf{b}$$

by rewriting the mean-variance constraint as

$$-\mathbf{w}' \boldsymbol{\mu} \leq -r_{\text{target}}$$

and setting $\mathbf{A} = -\boldsymbol{\mu}'$ and $\mathbf{b} = -r_{\text{target}}$.

Suppose we are given a portfolio with a number of stocks equal to `numAssets`, expected returns for the stocks stored in a vertical vector `muVec`, covariance matrix `SigmaMx`, and required expected return of `targetReturn`. Consider a simple portfolio of two stocks with expected returns of 9.1% and 12.1%, standard deviations of returns of 16.5% and 15.8%, and a correlation of -0.22 (covariance of -57.35). A MATLAB script that uses input data for the two stocks, calls the optimization solver for several instances of the problem with different values of `targetReturn`, and plots the efficient frontier looks as follows:

```

%create lower bounds array for asset weights (negative infinity)
lb = -inf*ones(numAssets,1);

%create upper bounds array for asset weights (infinity)
ub = inf*ones(numAssets,1);

%create right hand and left hand side of equality constraints
beq = [1];
Aeq = transpose(ones(numAssets,1));
[weights,variance] = quadprog(H,f,A,b,Aeq,beq,lb,ub);

%print results to screen
stdDev = sqrt(variance)
weights

%EFFICIENT FRONTIER
%loop through different values of the target portfolio returns, compute the
%optimal portfolio standard deviation, and plot the efficient frontier

iCounter = 1;

for iTRet = 9.5:0.5:12
    b = -iTRet;
    [weights,variance] = quadprog(H,f,A,b,Aeq,beq,lb,ub);
    y(iCounter) = iTRet;
    x(iCounter) = sqrt(variance);
    iCounter = iCounter + 1;
end

%plot efficient frontier
plot(x,y);
xlabel('Portfolio standard deviation');
ylabel('Portfolio expected return');
title('Efficient Frontier');

```

The command

```
[weights,variance] = quadprog(H,f,A,b,
Aeq,beq,lb,ub);
```

ensures that the optimal solution to the optimization problem is stored in a vector called `weights`, and the optimal objective function value (the minimum portfolio variance) is stored in the scalar `variance`. This is an example of using a MATLAB built-in function.

The portfolio standard deviation is computed as the square root of variance.

The MATLAB output from running the code above is as follows:

```
stdDev =
    10.4928
weights =
    0.3667
    0.6333
```

The script also contains an example of a `for` loop that runs the optimization problem for values of the target return between 9.5 and 12, increasing the target return by 0.5 at each iteration. The expected portfolio return and the optimal standard deviation obtained from the optimization output are stored in vectors `x` and `y`. The last few lines in the code plot the efficient frontier using the values stored in `x` and `y`, and label the graph.

Pricing a European Call Option by Simulation

Simulation is a technique for replicating uncertain processes and evaluating decisions under uncertain conditions. In the financial context, it typically involves generation of random numbers from particular probability distributions, using those for approximating the behavior of exogenous variables such as stock returns, and assessing outcomes of interest, such as the performance of a portfolio or the price of a financial instrument.

Through the Statistics Toolbox, MATLAB provides commands for generating the most commonly used random numbers directly. For example, a normal random variable can be simulated with

```
>> normrnd(mean, stdev, numRows,
           numColumns)
```

In the expression above, `mean` and `stdev` are the mean and the standard deviation of the normal random variable. `numRows` and `numColumns` specify the dimension of the array of random numbers we would like to generate.

We show how to use MATLAB's Statistics Toolbox to compute the price of a European call option with simulation under the assumptions that there are no transaction costs or market frictions, and the price of the underlying follows geometric Brownian motion. (The closed-form formula for pricing the option under these assumptions is the Black-Scholes for-

mula.) Option pricing by simulation was first suggested by Boyle (1977). For further details on the implementation and more examples, see Pachamano and Fabozzi (2010).

The evolution of the asset price at time t , S_t , can be described by the equation

$$dS_t = \mu S_t dt + \sigma S_t dW_t$$

where W_t is standard Brownian motion and μ and σ are the drift and the volatility of the process, respectively. For technical reasons (absence of arbitrage), when pricing an option, the drift μ is replaced by the risk-free rate r .

Under the assumption for the random process followed by the asset price, the value of the asset price S_T at time T given the asset price S_t at time t can be computed as

$$S_T = S_t e^{(r - \frac{1}{2}\sigma^2) \cdot (T-t) + \sigma \cdot \sqrt{(T-t)} \cdot \tilde{\varepsilon}}$$

where $\tilde{\varepsilon}$ is a standard normal random variable. (If the stock pays a continuously compounded dividend yield of q , then we use $(r - q - 0.5 \cdot \sigma^2)$ instead of $(r - 0.5 \cdot \sigma^2)$ as the drift term.)

The price of the option can be approximated by creating scenarios for the stock price S_T at time T , computing the discounted payoffs of the option, and finding the expected payoff of the option. Suppose we generate N scenarios for $\tilde{\varepsilon}$: $\varepsilon^{(1)}, \dots, \varepsilon^{(N)}$. Then, the price of a European call option with strike price K will be

$$C_t = e^{-r(T-t)} \cdot \sum_{n=1}^N \frac{1}{N} \cdot \max \left\{ S_t e^{(r - \frac{1}{2}\sigma^2) \cdot (T-t) + \sigma \cdot \sqrt{(T-t)} \cdot \varepsilon^{(n)}} - K, 0 \right\}$$

The expression above is the expected value of the option payoffs; that is, the weighted average of the option payoffs. The "weight," or the probability of each scenario, is $1/N$.

In MATLAB, we create a function `EuropeanCall` (stored in a file `EuropeanCall.m`), which follows.

```

function CEPrice = EuropeanCall(initPrice,K,r,T,sigma,q,numPaths)
%function for evaluating the price of a European call option using crude
%Monte Carlo
%initPrice is the initial price, K is the strike price, r is the annual interest
%rate, T is the time to maturity, sigma is the annual volatility, q is the
%continuous dividend yield, numPaths is the number of scenarios to generate for
%the evaluation

CEpayoffs = zeros(numPaths,1);

%compute a vector array of asset prices, one for each scenario
assetPrices = initPrice*exp((r-q-0.5*sigma^2)*T+sigma*sqrt(T)*
normrnd(zeros(1,numPaths),ones(1,numPaths))));

CEpayoffs = exp(-r*T)*max(assetPrices - K,0);

CEPrice = mean(CEpayoffs);

```

In the function, we generate the (random) end points of `numPaths` paths for the underlying stock price under the assumption that the price follows geometric Brownian motion. We use the Statistics Toolbox function `normrnd(mu,sigma)`, which in this case returns a vector array with the realizations of normal random variables. The array has the dimension of the `mu` and `sigma` vectors, which are vectors of zeros and ones, respectively, with length `numPaths`. Then, we generate a vector array of asset prices by calculating the asset price in each scenario. We use a nice feature in MATLAB, which is that we can pass an array (namely, `normrnd(zeros(1,numPaths),ones(1,numPaths))`) into a formula (namely, `initPrice*exp((r-q-0.5*sigma^2)*T + sigma*sqrt(T)*normrnd(zeros(1,numPaths),ones(1,numPaths)))`), and MATLAB automatically creates an array with results (`assetPrices`). In other programming languages, we would need to implement this by creating a `for` loop.

Finally, we calculate the option price `CEPrice` as the average of the payoffs of the option in each scenario by using the function `mean`.

Pricing a European Call Option Using a Sobol Sequence

In the function `EuropeanCall`, we used the MATLAB built-in function `normrnd` from the Statistics Toolbox with arguments that were arrays of zeros and ones to generate a set of realizations drawn from a standard normal probability distribution and compute a set of paths for the price of the underlying. Alternatively, we could have generated a set of quasirandom numbers that sometimes lead to a faster and more accurate estimation for the option price. (See the discussion in Chapter 14 of Pachamanova and Fabozzi, 2010; Chapter 6 in McLeish, 2005; or section 5.2.3 of Chapter 5 in Glasserman, 2004.) MATLAB's Statistics Toolbox contains built-in syntax for computing the elements of some low-discrepancy sequences, such as the Sobol sequence (Sobol, 1967). Namely, the function `sobolset(d)` computes a Sobol sequence of dimension `d`, and the sequence can then be retrieved with the command `net`. For example,

```
seq = sobolset(3); net(seq,5)
```

returns the first five elements of a Sobol sequence of dimension 3.

The calculation of the European call option price using the Sobol sequence is shown in the function `EuropeanCallSobol` below.

```
function SCEPrice = EuropeanCallSobol(initPrice,K,r,T,sigma,q,numPaths)
%function for evaluating the price of a European option using
%a Sobol sequence
%initPrice is the initial price, K is the strike price
%r is the annual interest rate, T is the time to maturity, sigma is the
%annual volatility
%q is the continuous dividend yield
%numPaths is the maximum number of scenarios to generate for the evaluation

SCEpayoffs = zeros(numPaths,1);
%use the sobolset function in the Statistics Toolbox to generate the
%sequence
seq = sobolset(1);
SobolPoints = net(seq,numPaths+1);
%drop the first element, which is 0
SobolPoints = SobolPoints(2:numPaths+1);

%compute a vector array of asset prices, one for each Sobol point
assetPrices = initPrice*exp((r-q-0.5*sigma^2)*T+sigma*sqrt(T)*
norminv(SobolPoints));

%compute a vector array of discounted payoffs, one for each scenario
%generated from a Sobol point

SCEpayoffs = exp(-r*T)*max(assetPrices - K,0);

%compute price of option
SCEPrice = mean(SCEpayoffs);
```

Again, in this function, we passed an array (`SobolPoints`) into a formula (`initPrice * exp((r-q-0.5*sigma^2)*T + sigma*sqrt(T)*norminv(SobolPoints))`), and MATLAB automatically created an array with results (`assetPrices`).

The Sobol sequence generated in the function is of dimension 1 and length `numPaths+1`. We created it with the commands

```
seq = sobolset(1);
SobolPoints = net(seq,numPaths+1);
```

and remove the first element, which is 0, with the command

```
SobolPoints = SobolPoints(2:numPaths+1);
```

(As explained in Chapter 14 of Pachamanova and Fabozzi [2010], it is common to drop some number of elements of low-discrepancy sequences. It takes a certain “warming up” for the low-discrepancy sequence to begin producing stable and accurate estimates.)

Computing the Black-Scholes Price of a European Option Using the Financial Toolbox

The price for the European option obtained in the ways described in the previous two sections is, of course, an approximation. It will vary slightly depending on the specific set of scenarios simulated with the `normrnd` function, or on the number of points generated with the Sobol sequence. The true option price under the stated assumptions is given by the Black-Scholes formula. (See Black and Scholes, 1973; Hull, 2008; or Pachamanova and Fabozzi, 2010.) As we mentioned earlier in this entry, the function `blsprice` in MATLAB's Financial Toolbox can compute this price. For example, for an initial price of 100, a strike price of 110, an interest rate of 6%, time to maturity of 1 year, and volatility 40%, the Black-Scholes price for the European call option will be computed by typing

```
>> blsprice(100, 110, 0.06, 1, 0.40)
```

at the MATLAB prompt. MATLAB returns

```
ans =
    14.4018
```

You should get a similar price by typing the names of the user-defined functions we wrote previously,

```
>> EuropeanCall(100,110,0.06,1,0.40,
    0,20,1000)
```

to compute it with simulation, or

```
>> EuropeanCallSobol(100,110,0.06,1,
    0.40,0,1000)
```

to compute it by using a Sobol sequence. Here we are requesting that the price be evaluated with 1,000 paths for the price of the underlying. The greater the number of paths, the closer

the estimates will be to the Black-Scholes price. For this example, we obtained 14.3772 for the option price by crude Monte Carlo simulation, and 14.0882 by using the Sobol low-discrepancy sequence. The variability for the option price estimated using the crude Monte Carlo simulation approach is large, so readers can expect answers that differ quite a bit.

KEY POINTS

- MATLAB uses a number-array-oriented programming language; that is, a programming language in which vectors and matrices are the basic data structures.
- Array operations are very efficient in MATLAB.
- Specialized MATLAB toolboxes provide additional capabilities, save time, and simplify model building. Some toolboxes build on the capabilities of other toolboxes and need to be purchased in groups.
- An M-file is a file with instructions that MATLAB executes sequentially. Such files are saved with the suffix ".m" and can be called from the prompt in MATLAB's Command window by typing their name without the suffix ".m".
- M-files can be scripts, that is, a simple listing of instructions for MATLAB, or functions, which take in a certain number of arguments and return a certain number of outputs.
- While general script M-files can contain any sequence of instructions that will be completed when the name of the file is typed at the MATLAB prompt, function M-files need to start with a specific first line. That line contains the word "function" and a declaration of the function name, inputs, and outputs. The function name and the name of the M-file should be the same.
- Control flow statements in MATLAB include `for` loops, `if` statements, `while` loops, `switch-case` constructions, and `try-catch` blocks.

- MATLAB has beautiful 2-D and 3-D graphing capabilities. The most common function for plotting 2-D graphs is `plot`.
- MATLAB has the ability to interact efficiently with Microsoft Excel. The core product contains commands that allow importing data from and exporting data to Excel.
- Spreadsheet Link EX is a useful toolbox that allows a more complex interface between MATLAB and Excel. With Spreadsheet Link EX, one can call MATLAB's functions directly from within Excel, thus ensuring access to MATLAB's superior computational and graphical capabilities.
- Optimization in MATLAB can be performed through the Optimization and the Global Optimization Toolboxes. These capabilities are especially useful for quantitative portfolio management.
- MATLAB expects optimization formulations to be passed to its solvers in an array form and has functions that can call specific solvers for specific types of optimization problems.
- The MATLAB Statistics Toolbox contains functions for

random number generation and can be used when performing financial simulations.

REFERENCES

- Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 3: 637–654.
- Boyle, P. (1977). Options: A Monte Carlo approach. *Journal of Financial Economics* 4, 3: 323–338.
- Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. New York: Springer-Verlag.
- Hull, J. (2008). *Options, Futures and Other Derivatives*, 7th Edition. Upper Saddle River, NJ: Prentice Hall.
- Markowitz, H. (1952). Portfolio selection. *Journal of Finance* 7: 77–91.
- McLeish, D. (2005). *Monte Carlo Simulation and Finance*. Hoboken, NJ: John Wiley & Sons.
- Pachamanova, D. A., and Fabozzi, F. J. (2010). *Simulation and Optimization in Finance: Modeling with MATLAB, @RISK, and VBA*. Hoboken, NJ: John Wiley & Sons.
- Sobol, I. (1967). The distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics* 7, 4: 86–112.

Introduction to Visual Basic for Applications

DESSISLAVA A. PACHAMANOVA, PhD

Associate Professor of Operations Research, Babson College

Abstract: Visual Basic for Applications (VBA) is a programming language environment that allows Microsoft Excel users to automate tasks, create their own functions, perform complex calculations, and interact with spreadsheets. Despite some important limitations, VBA adds useful capabilities to spreadsheet modeling and is a good tool to know for finance professionals for whom Microsoft Excel is the platform of choice.

This entry is a brief introduction to Visual Basic for Applications (VBA), the *programming language* environment that allows *Microsoft Excel* users to *automate* tasks, create their own functions, perform complex calculations, and interact with *spreadsheets*. We focus on features of VBA useful for financial applications. For a comprehensive introduction to VBA, good references are Walkenbach (2004) and Roman (2002). The Excel VBA help is also useful as a quick reference. All Excel commands listed in this entry are based on Microsoft Office 2007.

A SIMPLE EXAMPLE OF A VBA PROGRAM

Before we review some important characteristics of the VBA language, let us create a simple example of a VBA program. Excel has a tool for recording tasks performed in a spreadsheet, which can then be replayed as a macro. Macros

in Excel record a sequence of commands, so that you do not have to repeat the same set of instructions if you need to perform the task several times. Macros are in effect computer programs whose commands are hidden from the user, but can be seen if you open the VBA editor (VBE). You can access the VBE by using a shortcut, **Alt-F11**, in all versions of Excel. In Excel 2007, VBE can be accessed from the **Developer** tab. If the **Developer** tab is not visible, do the following to set it up: Click on the main MS Excel button , then **Excel Options**. Under the **Popular Options** tab, check **Show Developer Tab in Ribbon**. Once the **Developer** tab is available in Excel's top menu, you can click on the Visual Basic button in the ribbon associated with it to open the editor. (See Figure 1.)

Use the **Macro Security** button to enable macros. (It is always a good idea to return to the default—disabled macros—after you are finished working with macros.)

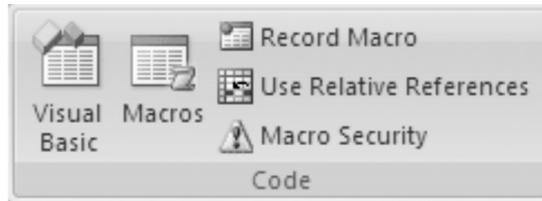


Figure 1 Visual Basic Button in the Developer Ribbon in Excel 2007

Open a new file and name it **ReturnCalc.xlsm**. (Excel 2007 will automatically make the file extension .xlsm if there are macros already in the file. Here, we do not have macros yet, so the default in Excel 2007 will be to save the file as .xlsx. To save the file with extension .xlsm, you need to select **Excel Macro-Enabled Workbook** from the drop-down menu next to **Save as Type** in the **Save** dialog box.)

We are trying to create the layout shown in Figure 2. First, enter the text in columns A and B; that is, enter stock prices for three points in time. Suppose we want to compute the realized cumulative return over the two time periods for any set of three stock prices in column B. We can do that by, for example, computing the realized returns over each of the two periods in column C, and then computing the cumulative return between times 1 and 3 in cell D5.

Let us record the entries and the calculations as a macro. To record a macro, click on **Record Macro** in the **Developer** tab. Delete the default name **Macro 1**, and replace it with something more meaningful, for example, **ReturnCalc**. Click **OK**. Once the macro recorder is on, do the following:

1. Enter $= (B3 - B2) / B2$ in cell C3 (this will compute the return for time period 1–2).
2. With the cursor in cell C3, enter **Ctrl-C** to copy the contents of cell C3, move the cursor to cell C4, and enter **Ctrl-V** to paste. This will fill cells C4 with the formula for computing the return between times 2 and 3.
3. Highlight cells C3–C4, right-click, select **Format Cells | Number | Percentage | Decimal Points 2** to format the returns as percentages.
4. Click on cell D3, enter $= (1 + C3)$. Then right-click, select **Format Cells | Number | Number | Decimal Points 2** to format the contents of the cell as a number.
5. Click on cell D4, enter $= D3 * (1 + C4)$.
6. Type **Total Return** in cell C5.
7. Enter $= D4 - 1$ in cell D5 to compute the total return over the five periods.
8. Highlight cells C5:D5. Right-click, then select **Format Cells | Border**. Select the double-line,

	A	B	C	D	E
1			Returns	Cumulative	
2	Stock price 1	100			
3	Stock price 2	110	10.00%	1.10	
4	Stock price 3	105	-4.55%	1.05	
5			Total return	5.00%	
6					

Figure 2 Macro Recording Example

then click the upper line of the cell in the **Border** window to make the double-line appear. Click **OK**.

9. Click on the stop button in the macro recorder to stop recording.

Now let us see what the macro does. You can use the file you created. Delete all contents from the array C3:D5. Press **Alt-F8** or, equivalently, click on the **Macro** button in the **Developer** tab. Select `ReturnCalc`, press **OK**. The spreadsheet

should fill up with the entries that you entered before. If you had changed the value of the stock price in any of the three cells in column B, the macro should calculate the correct corresponding value for total return in cell D5.

Behind the scenes, Excel recorded VBA code with instructions that tell Excel what functions to perform when you run the macro. You can see these instructions by opening the VBA editor and clicking on **Modules | Module 1**. The instructions look like this:

```

1 Sub ReturnCalc()
2 '
3 ' ReturnCalc Macro
4 ' Macro recorded month/day/year by you
5 '
6
7 '
8 Range("C3").Select
9 ActiveCell.FormulaR1C1 = "= (RC[-1]-R[-1]C[-1])/R[-1]C[-1]"
10 Range("C3").Select
11 Selection.Copy
12 Range("C4").Select
13 ActiveSheet.Paste
14 Range("C3:C4").Select
15 Selection.NumberFormat = "0.00%"
16 Range("D3").Select
17 ActiveCell.FormulaR1C1 = " = 1+RC[-1]"
18 Range("D3").Select
19 Selection.NumberFormat = "0.00"
20 Range("D4").Select
21 ActiveCell.FormulaR1C1 = " = R[-1]C*(1+RC[-1])"
22 Range("C5").Select
23 ActiveCell.FormulaR1C1 = "Total return"
24 Range("D5").Select
25 ActiveCell.FormulaR1C1 = " = R[-1]C-1"
26 Range("D5").Select
27 Selection.Style = "Percent"
28 Selection.NumberFormat = "0.00%"
29 Range("C5:D5").Select
30 Selection.Borders(xlDiagonalDown).LineStyle = xlNone
31 Selection.Borders(xlDiagonalUp).LineStyle = xlNone
32 Selection.Borders(xlEdgeLeft).LineStyle = xlNone
33 With Selection.Borders(xlEdgeTop)
34     .LineStyle = xlDouble

```

```

35     .Weight = xlThick
36     .ColorIndex = xlAutomatic
37 End With
38 Selection.Borders(xlEdgeBottom).LineStyle = xlNone
39 Selection.Borders(xlEdgeRight).LineStyle = xlNone
40 Selection.Borders(xlInsideVertical).LineStyle = xlNone
41 Range("D5").Select
42 End Sub

```

Knowing the actions we took to create the macro, it is relatively straightforward to trace what the program is doing at every step. To understand better how the macro works, however, and to know how to create such scripts without recording them in the spreadsheet, we need to understand some basic facts about VBA.

OBJECTS, PROPERTIES, AND METHODS

The most important fact about VBA is that it tries to act as an object-oriented language. (VBA does not quite qualify as an object-oriented language for technical reasons; however, for all practical purposes it is helpful to remember that VBA shares many of the same concepts as “real” object-oriented programming languages.) This means that it treats every component of Excel, such as a worksheet, a cell, a range of cells, and a chart, as an object. Objects are arranged in a hierarchy and have properties (attributes) that can be modified by entering the name of the object followed by dot and a specific command. In addition, objects are associated with actions (methods) that the objects can perform or have applied to them. You can view all objects by selecting **View | Object Browser** from the top menu in the VBE window. In Excel 2007, you can also view a detailed list of objects, their properties, and their methods by clicking on **Help** (pressing **F1**) and selecting **Excel Object Model Reference**.

The largest object, the object on top of the hierarchy, is Excel itself. It is the `Application` object. Worksheets, ranges, selections, charts, and

so on are all objects that are lower in the hierarchy. Objects in the same class are organized in collections. For instance, the `Workbooks` collection contains all workbooks (Excel files) that are currently open. Similarly, the `Worksheets` collection contains all Excel spreadsheets in the files that are currently open, the `Sheets` collection contains all Excel spreadsheets and charts in the files that are currently open, and so on. Thus, for example, to reference cell C3 in Worksheet **Return** in file (Workbook) **ReturnCalc.xlsm**, you would type

```

Application.Workbooks("ReturnCalc
    .xlsm").Sheets("Return").Range("C3")

```

This reference is rather long and, as we can see from the actual VBA code, it is not necessary, as long as the macro is saved within the active Excel workbook and the identification of the cell range that is referenced is unique. In our example, `Range("C3")` is sufficient to reference cell C3, because the objects higher in the hierarchy, such as the name of the worksheet and the name of the file, are implied in the reference.

An example of an action (method) that can be performed on an object is the command `Select`. The `Select` method applies to several objects, including `Worksheet`, `Chart`, and `Range`. Notice that it was used often in the macro we created, because clicking on a cell or highlighting on an array performed the action. For example, in line 14 we selected the range C3:C4. Similarly, in line 10 we selected the cell C3 with the command

```

Range("C3").Select

```

Then, the `Selection` property of an object in the background (the `Window` object) was used to return a `Range` object (representing the selected range on the spreadsheet) on which we can apply other methods, such as `Copy` (line 11 of the code):

```
Selection.Copy
```

VBA usually suggests actions and properties that can be used with an object, so you can select from a list.

Another example of modifying the properties of the object is in lines 14–15 of the VBA code. They request that the format of the cell range C3:C4 be changed to percentage with two digits after the decimal point. Namely, we selected the range C3:C4, and the `NumberFormat` property of the `Range` object that was returned by the `Selection` property was set to percentage with two digits after the decimal point.

While the code we created by recording a macro is helpful in understanding the basics of the VBA language, it can be confusing because it is unnecessarily verbose. For example, the same result as lines 14–15,

```
Range("C3:C4").Select
Selection.NumberFormat = "0.00%"
```

can be achieved with the command

```
Range("C3:C4").NumberFormat =
    "0.00%"
```

which modifies directly the property `NumberFormat` of the object `Range("C3:C4")`.

You can test that this is the case by deleting lines 14–15 in the VBA code in your file and replacing them with `Range("C3:C4").NumberFormat = "0.00%"`. Save the code by pressing **Ctrl-S** or selecting **Save** from the list under the main Excel button . Next, delete cells C3:D5 in the spreadsheet, and run the `ReturnCalc` macro again. The result and the formatting should be the same.

The effect of the `With/End` structure in lines 33–36 is another piece of code that can be repli-

cated easily through other commands; the advantage of the structure is that it allows you to reduce the number of listed objects in the code, and that it makes the code more readable. A `With/End` statement requires the specification of an object. Inside the `With/End` statement, one can omit mentioning the object with every modification of a property or application of a method to the object. In this particular example, lines 33–36 could be replaced with

```
Range("C5:D5").Borders(xlEdgeTop)
    .LineStyle = xlDouble
Range("C5:D5").Borders(xlEdgeTop)
    .Weight = xlThick
Range("C5:D5").Borders(xlEdgeTop)
    .ColorIndex = xlAutomatic
```

with the same effect as the `With/End` statement that references `Range("C5:D5").Borders(xlEdgeTop)`. However, the `With/End` statement is more concise.

In general, when writing VBA code you do not need to select cells explicitly in order to enter data into them. However, if you are new to VBA, it is helpful to record the macro first to see the code VBA suggests, and clean up afterward. In addition, it is a good idea to “comment out” the redundant statements at first, rather than deleting them. (Commenting out is done by entering an apostrophe (') at the front of the line of code that you wish VBA to ignore.) After commenting out overly verbose statements, save the macro by pressing **Ctrl-S**, make sure it still does what you would like it to do, and only then go back and delete the redundant statements.

A less verbose version of the VBA code is

```
Sub ReturnCalc()
    '
    ' ReturnCalc Macro
    ' Less verbose
    '
    Range("C3").Formula = "=" & (RC[-1]
        -R[-1]C[-1])/R[-1]C[-1] "
    Range("C3").Copy
    Range("C4").Select
```

```

ActiveSheet.Paste
Range("C3:C4").NumberFormat =
"0.00%"
Range("D3").Formula = "= 1+RC[-1]"
Range("D3").NumberFormat = "0.00"
Range("D4").Formula = "= R[-1]C*
(1+RC[-1])"
Range("C5").Formula = "Total
return"
Range("D5").FormulaR1C1 = "=
R[-1]C-1"
Range("D5").Style = "Percent"
Range("D5").NumberFormat = "0.00%"
With Range("C5:D5")
    .Borders(xlDiagonalDown)
    .LineStyle = xlNone
    .Borders(xlDiagonalUp)
    .LineStyle = xlNone
    .Borders(xlEdgeLeft).LineStyle
    = xlNone
    With .Borders(xlEdgeTop)
        .LineStyle = xlDouble
        .Weight = xlThick
        .ColorIndex = xlAutomatic
    End With
    .Borders(xlEdgeBottom).LineStyle
    = xlNone
    .Borders(xlEdgeRight).LineStyle
    = xlNone
    .Borders(xlInsideVertical)
    .LineStyle = xlNone
End With
Range("D5").Select
End Sub

```

Notice how the `With/End` structure was used to reduce the number of words we need to use, and how `With/End` structures can be nested inside one another. You can test that this code achieves the same effect by replacing the current code in the module in your file **ReturnCalc.xlsm**, saving the new code, and rerunning the macro `ReturnCalc`.

Before we end this section, we would like to mention a useful property of the `Range` object, `Offset(v, h)`. It points to a cell that is v cells

above or below (vertical direction) and h cells to the right or left (horizontal direction) from a specific cell. For example,

```

Range("C5").Select
ActiveCell.Offset(1,2) = 10

```

sets the value of the cell that is 1 cell down and 2 cells to the right from cell C5 (i.e., cell E6) to 10. Similarly,

```

Range("C5").Select
ActiveCell.Offset(-1,-2) = 20

```

sets the value of the cell that is 1 cell up and 2 cells to the left from cell C5 (i.e., cell A4) to 20.

We saw the idea of referencing cells above and below and to the left and right of the current cells in the example code at the beginning of this section. For example, the formula in line 9 of the original macro,

```

ActiveCell.FormulaR1C1 = "= (RC[-1]
-R[-1]C[-1])/R[-1]C[-1]"

```

uses the cell in the same row and one column to the left (`RC[-1]`) and the cell one row up and one column to the left (`R[-1]C[-1]`) to compute the value in the active cell. These kinds of commands help when one prefers to create relative references—in other words, to perform tasks relative to a prespecified location in the spreadsheet without changing the code when the starting location is changed.

The default in VBA is to record macros in absolute reference mode. To change the mode to relative references, make sure that the relative references button in the **Developer** tab ( Use Relative References) is “pressed in” before starting the macro recorder.

PROGRAMMING TIPS

While some desired formatting of an Excel spreadsheet can be recorded with the macro recorder, knowing basic programming in VBA opens up a whole lot of additional functionality. For example, suppose that you have a set of data on stock returns over several months and,

as often happens with real-world data, it is not recorded well—there are some duplicate rows. You could record a macro as you go through the spreadsheet and clean them by hand, but next time you have a set of data, duplicate entries will not be exactly in the same rows as the first set of data. How can you tell Excel to sort through the data and remove duplicate rows in *any* set of data? You need to construct a program from scratch and make the code general enough to enable the script to be transferable.

In the remainder of this section, we cover some basic VBA programming concepts. We discuss the difference between subroutines and user-defined functions, explain variable declaration in VBA, and introduce some important control flow statements. These concepts are not unique to VBA—versions of them exist in most programming languages.

Subroutines versus User-Defined Functions

Subroutines and user-defined *functions* in VBA are both blocks of code saved in modules. (If you do not see a module when you open VBE, select **Insert | Module** from the top menu in VBE to create one.) The difference is that subroutines are general scripts; that is, lists of instructions, whereas functions complete a list of instructions and return a value to the user. Subroutines have the general form

```
Sub ()
[commands]
End Sub
```

whereas functions have the form

```
Function FunctionName(list of inputs)
  As type [commands]
FunctionName = Return value
  'Computed from [commands]
End Function
```

The macro recorded at the beginning of this entry was an example of subroutine code. Next, we provide another small example in order to

illustrate the difference between a subroutine and a function. Do not worry about the details of the commands right now; we will explain each part of the code in subsequent sections.

Suppose we would like to calculate $n!$ (pronounced “ n factorial”), where n is an integer number the user provides as input. $n!$ is the product of all integer numbers less than or equal to n ; that is, $n! = 1 \cdot 2 \cdot \dots \cdot n$. Next, we provide several examples of subroutines and user-defined functions that accomplish this goal. The subroutine

```
Sub FactorialSub1()
'Compute factorial using control flow
statements

'Declare the variable that will
'store the value for factorial
  Dim Factorial As Integer
'Declare the variable that will
'store the number n
  Dim inNumber As Integer
'Declare the variable that will be
'used as counter in the loop
  Dim i As Integer

'Read in the number from cell B1,
'store it in inNumber
  inNumber = Range("B1").Value

'Calculate factorial
  Factorial = 1
  For i = 1 To inNumber
    Factorial = i * Factorial
  Next i

  Range("B2").Value = Factorial
End Sub
```

takes the number specified in cell B1, computes the factorial of that number, and sets the value of the cell B2 to the value of that factorial. To see how this subroutine works, copy the code in a new module in the VBE window of a new Excel file. Enter the number 5 in cell B1. Press **Alt-F8**, and select `FactorialSub1`. The

subroutine fills cell B2 with 120 ($5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$).

The function `FactorialFun1` whose code is provided next computes the same result, but works in a different way. It takes a number as an input (`inNumber`), and returns a number as an output (`FactorialFun1`). The output to be returned should have the same name as the function.

```
Function FactorialFun1(inNumber)
    As Integer
        Dim i As Integer
        'Calculate factorial
        FactorialFun1 = 1
        For i = 1 To inNumber
            FactorialFun1 = i * FactorialFun1
        Next i
End Function
```

Add this function to the module in the VBE in your file. To call this function, type in a cell in your spreadsheet (say, cell B3) = `FactorialFun1(B1)`. If the value in cell B1 was still 5 (the value you entered in the previous example), then the value of cell B3 will be 120. Notice that the syntax for calling your (user-defined) function is not different from the syntax for calling built-in Excel functions. In fact, Excel has a function for computing a factorial, = `Fact(number)`, and if you entered the expression = `Fact(B1)` in, say, cell B4 of your spreadsheet, you would get the same result (120).

Excel built-in functions can be used also inside VBA code with the prefix `Application`. It is worthwhile to note, though, that VBA itself has some built-in numeric functions. In particular, functions such as `Abs` (absolute value), `Exp` (exponential), `Int` (integer part), `Cos` (cosine), `Sin` (sine), `Log` (natural log), `Rnd` (random number generator), `Sign` (sign function), `Tan` (tangent), and `Sqr` (square root) can be used directly within VBA code without the prefix `Application`. Although it seems that this should make things easier, it may also be a source of confusion. Notice that Excel has equivalent nu-

merical functions for formulas that are entered into cells in spreadsheets, but the syntax for some of the functions is different. For example, the natural logarithm function in Excel is `Ln`, and the square root function is `Sqrt`. So, typing `Sqr` in your program in VBA is equivalent to typing `Application.Sqrt`. In practice, you would want to use the shorter syntax `Sqr`. It is important to be aware of inconsistencies between names of equivalent functions in Excel and VBA.

The subroutine `FactorialSub1()` and the function `FactorialFun2()` whose code is provided below illustrate how the factorial can be computed by calling the built-in Excel function `Fact`.

```
Sub FactorialSub2()
    'Compute factorial using Excel's FACT
    'function within a subroutine
    Range("B5") = Application.Fact_
        (Range("B1"))
End Sub
```

```
Function FactorialFun2(inNumber) As
Integer
    'Calculate factorial
    FactorialFun2 = Application_
        .Fact(inNumber)
End Function
```

Copy the code above in the module in your file. The subroutine `FactorialSub2()` uses the number entered in cell B1 in the spreadsheet as an input and calls the Excel function `Fact()` to compute the factorial of the value in cell B1. The function `FactorialFun2()` is called with an input argument that is a number and returns the factorial of that number. If you type = `FactorialFun2(B1)` in cell B6 and the value in cell B1 is still 5, you should obtain 120 in cell B6.

What is the advantage of using user-defined functions rather than subroutines? In some cases, you can only use one or the other. However, in cases in which both are possible, it may be preferable to structure the script as

a function as opposed to a subroutine. User-defined functions are more “transferable”—in other words, it is easier to use them in different places in the spreadsheet. There are some other conveniences—for example, check what happens when the number for n in cell B1 is changed from 5 to 6. Cell B3 (which contains the call to the function `FactorialFun1`) immediately updates to 720, which is the correct result. However, cells B2 and B5—those that are output ranges for the subroutines `FactorialSub1` and `FactorialSub2`—do not update until you rerun the macros associated with them.

Variable Declaration

Variables are a basic common concept in computer languages. They are used to store numerical and text data and handle intermediate output in subroutines and functions. For example, `inNumber` in the code for `FactorialSub1` was a variable that stored the value of n for which the factorial should be computed. There is no convention for naming variables, but a good practice is to give them meaningful names (rather than x , y , and z), so that your code is easier to follow. We prefer to start names of variables with small letters. If there is a second word in the name, that word starts with a capital letter. We also like to differentiate variables that store inputs (such as `inNumber`) and variables that record output (e.g., `outFactorialValue`).

Depending on their type, variables are handled differently and are allocated a different amount of memory. For example, we specified that `inNumber` should be an integer number by declaring it with the syntax `Dim variableName As variableType`:

```
Dim inNumber As Integer
```

Other types of variables include `String`, `Single`, `Double`, `Long`, `Boolean`, `Date`, `Object`, `Variant`, and so on. For example, when you need a variable that will hold a fractional (also called “floating point”)

value, then you should use the `Single` or `Double` data type. When you need a variable to store text data, use the `String` type. The `Variant` type can be used to replace any type; however, it also uses up the largest amount of space, so it is better to specify a particular type for a variable if you know it.

When specifying a variable type, make sure that you have enough space for the data you are planning to store in that variable. If the value gets too large for the variable type, your program may crash. For example, the `Integer` type can store values between $-32,768$ and $32,767$. If you need to store an integer number outside this range, use the `Long` variable type. Similarly, the `Single` (floating point) type can store numbers between $-3.402823E38$ and $-1.401298E-45$ for negative values, and numbers between $1.401298E-45$ and $3.402823E38$ for positive values.¹ If you need to work with fractional numbers outside this range, use the `Double` (floating point) variable type.

Variables can be grouped into arrays. For example,

```
Dim myArray(5) As Integer
```

declares an array of integers of size 6.

One of the most confusing things about VBA is the way it handles arrays. The default is to index the first element in arrays as 0, which is the convention in most programming languages, which is why the total number of elements in `myArray` is 6. However, in some special circumstances arrays are treated as starting with the index 1. To ensure consistency and minimize confusion, it is helpful to use the command

```
Option Base 1
```

at the beginning of the module, which makes sure that the indexing of arrays always starts at 1. If this option is stated, then declaring

```
Dim myArray(5) As Integer
```

will result in an array of 5 elements. Those elements can be referenced as `myArray(1)`, ..., `myArray(5)` later in the program.

You can specify arrays of multiple dimensions as well, for example,

```
Dim myMultiArray(5,2) As Integer
```

will result in an array of 5 rows and 2 columns.

You can also declare dynamic arrays, that is, arrays that do not have specific dimensions from the beginning. This may happen if, for example, you have a set of data and you need to read it in before you know how many elements it has. In that case, you would declare an array

```
Dim myDynamicArray() As Integer
```

which will be filled as necessary. Once the number of elements is counted, the array can be resized by using the command `ReDim`, for example,

```
ReDim myDynamicArray(10)
```

`ReDim` reinitializes (sets to empty) all values within an array. If you want to preserve the values that are already there, use `ReDim Preserve`, which preserves as many elements as can fit in the new array dimensions.

Working with arrays within VBA is cumbersome and prone to errors. Often, one needs to resort to loops (see the introduction to loops in the next section) to handle array operations. In many cases, it may be preferable to use built-in Excel array manipulation functions, such as `SUMPRODUCT`, which performs vector multiplication. As we mentioned earlier, such built-in Excel functions can be called with `Application.FunctionName`. For example, `Array3 = Application.SUMPRODUCT(Array1, Array2)` will fill a variable array `Array3` with the result of the elementwise multiplication and summation of the matrix arrays `Array1` and `Array2`.

VBA will assume that you are creating a new variable whenever you use an expression that is not one of the standard commands. Stating the type of variables you use in the program can save you a lot of headache. (Typically, variable declaration is done at the beginning of the program.)

We also strongly recommend that you write the statement `Option Explicit` in the first line of your modules. This statement makes sure that Excel will report an error if it encounters an undeclared variable in your code. (This also can be accomplished by checking **Require Variable Declaration** under **Tools | Options** in the top VBE menu.) While this may seem like an inconvenience, think about a situation in which you mistype the name of a variable somewhere in your program. If Excel is not in the `Option Explicit` mode, it will treat the mistyped name as a new variable, ignoring any value that your variable may have had at that point in the program, and you will get nonsensical output. If Excel reports an error instead, you will know to fix the typo.

Control Flow Statements: For and If

Control flow statements in VBA allow for building more sophisticated programs than simple input and output of data to Excel. We briefly review a couple of important control statements that are used in VBA code: an example of an iterative statement (the `For` loop) and an example of checking a condition (the `If` statement).

The general syntax of a `For` loop in VBA is as follows:

```
For i = 1 to n
  commands
Next i
```

The commands inside the `For` loop are executed once for every value of `n`. (One can also specify a step by writing `For i = 1 to n Step k`. For example, if `n = 10` and step `k = 2`, then the commands in the loop will be executed for `n = 1, 3, 5, 7, 9`.)

We saw an example of a `For` loop in the code for calculating the factorial of a number `n`. Let us walk through the `For` loop code inside `FactorialSub1`.

```
'Calculate factorial
Factorial = 1
```

```

For i = 1 To inNumber
    Factorial = i * Factorial
Next i

```

The initial value of `Factorial` is set to 1. Suppose the value for `inNumber` is 5. The loop starts at `i = 1`. During the first iteration, the new value of `Factorial` equals the current value of `i` (which is 1) times the current value of `Factorial` (which is 1 as well). At the end of the first run through the loop, the value of `Factorial` is 1. Next, the value of `i` is set to 2. The new value of `Factorial` equals the current value of `i` (which is 2) times the current value of `Factorial` (which is 1); that is, it equals 2. At the third iteration, the value of `i` is 3 and the current value of `Factorial` is 2; that is, the new value of `Factorial` is $3 \cdot 2 = 6$. And so on and so forth for the next values of `i`, which are 4 and 5. The value of `Factorial` keeps getting updated until it reaches 720 ($= 5!$) in the last iteration of the loop.

There are other commands that enable iterating through commands multiple times, such as the **Do While** and **Do Until**. See VBE's **Help** for description of the syntax and use of these alternatives.

The general form of the `If` statement is

```

If condition Then
    commands
End If

```

When the condition is true, the block of commands executes. More generally, you can use a statement of the kind

```

If condition1 Then
    commands1
ElseIf condition2 Then
    commands2
Else
    commands3
End If

```

`Commands1` will be executed if `condition1` is true. If `condition1` is not true, then (and only then) `condition2` will be checked. If

`condition2` is true, then `commands2` will be executed. If `condition2` is not true, then `commands2` will be executed.

When using `If` statements, one typically needs to compare values of variables and check whether conditions are true. Therefore, it is useful to know about the logical operators that allow for such comparisons and checks. The comparison operators are the following:

=	tests for equality
<>	tests for inequality
<	tests whether the variable to the left of it is less than the variable on the right
>	tests whether the variable to the right of it is less than the variable on the left
<= and >=	test for less than or equal to/ greater than or equal to

Additional useful operators are **AND**, **OR**, and **NOT**. **AND** allows checking whether more than one statement is true at the same time. **OR** returns a `True` result if at least one of the statements is true. **NOT** returns a `True` result if the statement is false.

To illustrate how we can use these operators, consider a couple of simple examples that involve three numerical variables, `var1`, `var2`, and `var3`. Let `var1 = 5`, `var2 = 10`.

The code

```

If (var1 <> var2) Then
    var3 = 100
Else
    var3 = -100
End If

```

checks whether the value for `var1` is different from the value of `var2`. If it is (i.e., the value of the logical statement (`var1 <> var2`) is `True`), then the value of `var3` is set to `-100`; otherwise the value of `var3` is set to `100`. In this example, the value of `var3` at the end of the loop is `100`, since the value for `var1` (5) is indeed different from the value of `var2` (10).

Consider also the example

```
If (var1 < 5) Or (var2 > = 7) Then
    var3 = 100
Else
    var3 = -100
End If
```

The code checks if at least one of the statements $(var1 < 5)$ and $(var2 \geq 7)$ is true. If at least one of them is true, then the value of `var3` is set to 100; otherwise the value of `var3` is set to -100. In our case, the first statement is false, because the value of `var1` is not less than 5 (it is equal to 5). However, the second statement is true: The value of `var2` (10) is indeed greater than or equal to 7. Therefore, the combined statement $(var1 < 5) \text{ Or } (var2 \geq 7)$ is true, and the value of `var3` will be set to 100.

User Interaction in VBA

While we covered the most fundamental concepts about the VBA language, it is fun to learn about some additional capabilities that enable your programs to interact better with

their users. For example, once you have created a macro, you can associate it with a button that the user can press every time he or she wants the macro to run. To do that, go to the **Developer** tab, select **Insert | Form Controls**, and click on the button. When Excel pops up in the Macro dialog box, click on the macro you would like to have associated with this button.

Sometimes, it is convenient to ask the user to input information through an *input dialog box*. This can be done with the command `InputBox("question for user", "title of the input box")`. For example,

```
inNumber = InputBox("Enter an
integer", "Factorial Calculation")
```

will prompt the user to enter an integer number and will save that number into the variable `inNumber`. The title of the input box will be `Factorial Calculation`.

Other useful user interaction tools include `Message Box (MsgBox)`, which allows you to report output not in a cell on the spreadsheet, but in a message box. To test how it works, let us go through the following modification of the factorial calculation program (save it in your file as subroutine `FactorialSubMsgBox()`):

```
1 Sub FactorialSubMsgBox()
2     Dim inNumber As Variant
3     Dim numberType As Boolean
4     Dim outFactorial As Integer
5
6     inNumber = InputBox("Enter an integer number", "Factorial Calculation")
7
8     numberType = IsNumeric(inNumber)
9
10    If numberType = True Then
11        outFactorial = Application.Fact(inNumber)
12        MsgBox ("The factorial of " & inNumber & " equals " & outFactorial)
13    ElseIf numberType = False Then
14        MsgBox ("Not a number. Please enter an integer number.")
15    End If
16 End Sub
```

On line 6, we ask the user to specify the number for which we want to compute the factorial. On line 8, we check whether this is indeed a number. Note that the variable `numberType` is specified as `Boolean`, which means that it can only take `True` or `False` values. If it is true, that is, if `inNumber` is indeed a number, then we call the Excel built-in function `Fact` to calculate the factorial of this number, and print the statement “**The factorial of the number the user entered is the result obtained**” in a message box on the screen. If it is not true, then we prompt the user to enter a number.

Note that in line 2, we specified the type of variable for `inNumber` as `Variant`, which allows it to be anything. If we had declared `inNumber` `As Integer` and had entered a letter instead of a number, Excel itself would have returned an error, complaining that there is a variable type mismatch between what was declared and what the actual value of the variable is. Thus, declaring the exact type of variable whenever we know the type is very important for minimizing errors in output.

DEBUGGING

VBA has useful *debugging* tools that allow you to look at the code in more detail if your programs do not work as expected. These tools can be accessed through commands under the **Debug** item in the top menu of the VBE.

The “Step Into” button  (shortcut F8) lets you execute your program step by step. When you are executing a program step-by-step, your program is in “break mode.” Every time you press F8, the “break” is moved to the next command. While the break is set on a particular command, placing the cursor over any variable above the break point will give you an updated stored value for that variable. This makes it easy to catch calculation errors and inconsistencies. You can “step over” (i.e., skip) some subroutines that you are not interested in double-checking (use the button  or the

shortcut **Shift-F8**) and “step out” of the break mode (use the button  or the shortcut **Ctrl-Shift-F8**). Equivalently, you can click on the **Reset** button in the top VBE menu () to get out of debug mode.

Rather than going through the program step-by-step, it is sometimes helpful in long programs to set breakpoints in advance, so that the program runs until it gets to a particular breakpoint. A breakpoint can be specified by placing the cursor at the place where it should be inserted, and clicking on the button  in the **Debug** menu (or using the shortcut F9). When the program gets to the breaking point, it automatically goes into break mode and allows you to follow the subsequent commands step-by-step and check the values of the variables at that point in the program. To remove a breakpoint, simply place the cursor in the corresponding line, and click on the breakpoint button again.

EXAMPLES

The best way to learn to program in VBA is to see and implement many examples. Let us discuss three examples of using VBA in financial applications. The first example is a function that computes the Black-Scholes price of a European call option. It shows how a function is created, how variables are declared, and how Excel functions are accessed from within VBA. The second example is a function that generates possible paths for an asset price assumed to follow geometric Brownian motion. It involves using the random number generator in VBA, manipulating arrays, and iterating with loops. The third example is a function that computes the price of a European call option by simulation. It illustrates how user-defined and Excel functions can be called from within VBA functions, and provides another example of array manipulation and loops in VBA. Further examples of VBA scripts for financial applications, such as calculating the price of an Asian option, or computing and graphing the mean-variance efficient

portfolio frontier (see Markowitz, 1952), can be found in Pachamanova and Fabozzi (2010). See also Jackson and Staunton (2001).

Pricing a European Call Option with the Black-Scholes formula

The Black-Scholes formula for a European call option (C) is as follows (Black and Scholes, 1973):

$$C = S_0 \cdot e^{-qT} \cdot \Phi(d_1) - K \cdot e^{-rT} \cdot \Phi(d_2)$$

where

$$d_1 = \frac{\ln(S_0/K) + (r - q + \sigma^2/2) \cdot T}{\sigma \cdot \sqrt{T}}$$

$$d_2 = d_1 - \sigma \cdot \sqrt{T}$$

K is the strike price

T is the time to maturity

q is the percentage of stock value paid annually in dividends

Φ denotes the cumulative probability density function for the normal distribution

The value for $\Phi(d)$ can be found in Excel by using the built-in formula =NORMDIST(d , 0, 1, 1) or, equivalently, the formula =NORMSDIST(d).

To illustrate the Black-Scholes option pricing formula, assume the following values:

Current stock price (S_0) = \$50

Strike price (K) = \$52

Time remaining to expiration (T) = 183 days = 0.5 years (183 days/365, rounded)

Stock return volatility (σ) = 0.25 (25%)

Short-term risk-free interest rate = 0.10 (10%)

Plugging into the formula, we obtain

$$d_1 = \frac{\ln(50/52) + (0.10 - 0 + 0.25^2/2) \cdot 0.5}{0.25 \cdot \sqrt{0.5}} = 0.1502$$

$$d_2 = 0.1502 - 0.25 \cdot \sqrt{0.5} = -0.0268$$

$$\Phi(0.1502) = 0.5597$$

$$\Phi(-0.0268) = 0.4893$$

$$C = 50 \cdot 1 \cdot 0.5597 - 52 \cdot e^{-0.10 \cdot 0.5} \cdot 0.4893 = \$3.79$$

Next, we provide the code of a VBA function that computes the price of a European call option with the Black-Scholes formula.

```
Function BSCallPrice(initPrice As _
    Double, _
    K As Double, _
    T As Double, _
    r As Double, _
    sigma As Double, _
    q As Double)
    'Computes the Black-Scholes price of a
    'European call option
    'initPrice is the initial price of the
    'stock
    'r is the interest rate
    'T is the time to maturity of the
    'option
    'sigma is the volatility of the stock
    'q is the continuous dividend yield

    Dim dOne As Double

    dOne = (Log(initPrice / K) + (r - q _
    + 0.5 * sigma ^ 2) * T) / _
    (sigma * Sqr(T))

    BSCallPrice = initPrice * Exp(-q * T) _
    * Application.NormSDist(dOne) - _
    K * Exp(-r * T) * Application. _
    NormSDist(dOne - sigma * Sqr(T))

End Function
```

In the code above, all input variables ($initPrice$, r , T , $sigma$ and q) are specified to be of type `Double`. A variable `dOne` is declared as type `Double` within the function. `dOne` stands for d_1 in the definition of the Black-Scholes formula above. It takes the value of the expression $(\text{Log}(\text{initPrice} / K) + (r - q + 0.5 * \text{sigma} ^ 2) * T) / (\text{sigma} * \text{Sqr}(T))$. (Note that this expression contained an underscore (“_”) in the code above. The underscore is used when transferring a part of an expression to a new line.) The price of the option is stored

	A	B	C	D	E
1	Black-Scholes option pricing formula (VBA)				
2					
3	Initial price	\$ 50.00			
4	Strike price	\$ 52.00			
5	Time to expiration	0.50			
6	Interest rate	10%			
7	Volatility	25%			
8	Dividend yield	0			
9					
10	Call price	\$ 3.79	=BSCallPrice(B3,B4,B5,B6,B7,B8)		

Figure 3 Example of Using the User-Defined Function BSCallPrice in a Spreadsheet

in BSCallPrice. The functions Log and Exp used in the calculation are VBA functions. We also call the Excel function NormSDist with the expression Application.NormSDist.

The function BSCallPrice can then be used in a spreadsheet. An example is given in Figure 3. The inputs are stored in cells B3:B8, and the function is called with arguments that are cell references to cells where the information is stored.

VBA is forgiving if you are sloppy in writing the function. For example, the code below (without any variable declarations) would have worked as well.

```
Function BSCallPrice(initPrice,K,T,r,
sigma,q)

dOne = (Log(initPrice / K) + (r - q
+ 0.5 * sigma ^ 2) * T) / _ (sigma
* Sqr(T))

BSCallPrice = initPrice * Exp(-q * T)
* Application.NormSDist(dOne) -
K * Exp(-r * T) * Application.
NormSDist(dOne - sigma * Sqr(T))

End Function
```

However, as we mentioned earlier, it is a good practice to keep your code well organized. It helps minimize errors and saves you time in the long run.

Generating Paths for the Price of an Asset That Follows Geometric Brownian Motion

In finance, the dynamics of asset price processes in discrete time increments are typically described by two kinds of models: trees (such as binomial trees) and random walks. When the time increment used to model the asset price dynamics becomes infinitely small, such processes are referred to as stochastic processes in continuous time. The ability to generate paths for asset prices following these processes is important for computing prices of securities that depend on the asset price under consideration, as well as for calculating various risk measures associated with holding the asset in a portfolio.

The most widely used stochastic process in finance is geometric Brownian motion. The evolution of the underlying asset price is described by the equation

$$dS_t = \mu S_t dt + \sigma S_t dW_t$$

where W_t is standard Brownian motion, and μ and σ are the drift and the volatility of the process, respectively. (See a more detailed introduction in Hull [2008] or Pachamanova and Fabozzi [2010].) It turns out that the value of the asset price S_T at time T given the asset price S_t at time t can be computed as

$$S_T = S_t e^{(\mu - \frac{1}{2}\sigma^2)(T-t) + \sigma \cdot \sqrt{(T-t)} \cdot \varepsilon}$$

where $\tilde{\varepsilon}$ is a standard normal random variable. If the stock pays a continuously compounded dividend yield of q , then we use $(\mu - q - 0.5\sigma^2)$ instead of $(\mu - 0.5\sigma^2)$ in the above formula.

Let us create a function, `GBMPaths`, that generates a prespecified number of paths (`numPaths`) for the asset. Each path consists

```
Function GBMPaths(initPrice As Double, _
    mu As Double, _
    sigma As Double, _
    T As Double, _
    q As Double, _
    numSteps As Integer, _
    numPaths As Integer)

    Randomize
    Dim iPath, iStep As Integer
    Dim paths() As Variant
    ReDim paths(1 To numSteps + 1, 1 To numPaths)

    For iPath = 1 To numPaths
        paths(1, iPath) = initPrice
        For iStep = 2 To numSteps + 1
            paths(iStep, iPath) = paths(iStep - 1, iPath) * _
                Exp((mu - q - 0.5 * sigma ^ 2) * (T / numSteps) + _
                    sigma * (T / numSteps) ^ (1 / 2) * _
                    (Application.NormSInv(Rnd)))
        Next
    Next
    GBMPaths = paths
End Function
```

Let us now see what this function does. First, we use the command `Randomize` to make sure that VBA creates a different sequence each time we generate normal random variables to compute the paths for the asset. (If you do not type `Randomize` before you use the VBA random generator function `Rnd`, `Rnd` will always return the same sequence of numbers.)

Next, we declare variables we will use in the function. The variables `iPath` and `iStep` will be counters for the number of paths and the number of steps we have generated so far. They are, of course, integers. The two-dimensional array `paths` will store the values of the asset

of a prespecified number of steps (`numSteps`). The value of the asset at each step is computed according to the formula for S_T above. In the formula, we replace time t with time 0 (i.e., the present), and time T with the time corresponding to the step.

The code of the function is

along each path and for each step. We use `ReDim` to specify the dimensions of the array.

We next use a `for` loop to populate the array `paths`. In fact, we have two nested `for` loops—one that iterates through the number for paths, and one that iterates through the points in each path. For each point `iStep` on each path `iPath`, we calculate the price of the asset and store it in `paths(iStep, iPath)`. The formula that computes the price of the asset contains the expression `Application.NormSInv(Rnd)`, which generates a value for the normal random variable $\tilde{\varepsilon}$ in the formula for S_T earlier in this section. `Rnd` is the VBA random

number generator—it returns a random number between 0 and 1. The reason we used the command `Randomize` at the beginning of the function is so that we could force `Rnd` to generate different sequences of random numbers every time you call the function `GBMPaths`. `NormSInv` is an Excel function that finds the number on the horizontal axis of the normal distribution that corresponds to the value for cumulative probability generated by `Rnd`. (See, for example, Chapter 4 in Pachamanova and Fabozzi [2010] for an explanation of how random numbers from different probability distributions are generated.) As in the previous example, in order to indicate to VBA that `NormSInv` is an Excel function, we use `Application`. in front of `NormSInv`.

The function returns a two-dimensional array, `GBMPaths` (which is equal to `paths`, as set in the second-to-last line of the function). Every column of the array contains a randomly generated path for the asset price; that is, it has `numSteps` values that represent the values of the asset price along that path.

Pricing a European Call Option by Simulation

Let us now use the function we created in the previous section to write a function that prices a European call option by simulation. While this is not the most efficient way to price a European call option by simulation, it will illustrate how user-defined functions are called within other functions, and how arrays are handled as outputs of a function.

As in the previous section, we will make the assumption that the asset price follows geometric Brownian motion, which means that the value of the asset price S_T at time T given the asset price S_t at time t can be computed as

$$S_T = S_t e^{(r - \frac{1}{2}\sigma^2) \cdot (T-t) + \sigma \cdot \sqrt{(T-t)} \cdot \tilde{\varepsilon}}$$

where $\tilde{\varepsilon}$ is a standard normal random variable. (When we generate asset price paths for the purpose of valuing an option, we use r (the risk-free rate) in place of the drift term μ . This is done for technical reasons (absence of arbitrage).) As in the previous section, if the stock pays a continuously compounded dividend yield of q , then we use $(r - q - 0.5 \cdot \sigma^2)$ instead of $(r - 0.5 \cdot \sigma^2)$ in the formula above.

The price of the option can be approximated by creating scenarios for the stock price S_T at time to maturity T , computing the discounted payoffs of the option, and finding the expected payoff of the option. (Option pricing by simulation was first suggested by Boyle, 1977. See also Boyle et al., 1997; Pachamanova and Fabozzi, 2010; Glasserman, 2004; or McLeish, 2005.)

Suppose we generate N scenarios for $\tilde{\varepsilon}$ at time T : $\varepsilon^{(1)}, \dots, \varepsilon^{(N)}$. Then, the price of a European call option with strike price K will be

$$C_t = e^{-r(T-t)} \cdot \sum_{n=1}^N \frac{1}{N} \times \max \left\{ S_t e^{(r - \frac{1}{2}\sigma^2) \cdot (T-t) + \sigma \cdot \sqrt{(T-t)} \cdot \varepsilon^{(n)}} - K, 0 \right\}$$

The expression above is the expected value of the option payoffs, that is, the weighted average of the option payoffs.

The VBA code of the function is given below.

```
Function EuropeanCall(initPrice As Double, _
    K As Double, _
    r As Double, _
    T As Double, _
    sigma As Double, _
    q As Double, _
    numSteps As Integer, _
    numPaths As Integer)
```

```

Dim iPath As Integer
Dim payoffs() As Variant
ReDim payoffs(1 To numPaths)
Dim paths() As Variant
ReDim paths(1 To numSteps + 1, 1 To numPaths)

paths = GBMPaths(initPrice, r - q, sigma, T, q, numSteps, numPaths)
For iPath = 1 To numPaths
    payoffs(iPath) = Exp(-r * T) * _
        Application.Max(paths(numSteps + 1, iPath) - K, 0)
Next

EuropeanCall = Application.Average(payoffs)
End Function

```

The variable declarations are similar to the declarations in the previous sections; however, now we have an additional array, `payoffs`, that will store the payoff of the option at the end of each generated path (that is, for each generated scenario). The dimension of the array is therefore $1 \times \text{numPaths}$.

After declaring the variables in the function, we call the function we created in the previous section, `GBMPaths`, and store the output in the array `paths`. This is achieved with the command

```
paths = GBMPaths(initPrice, r - q,
    sigma, T, q, numSteps, numPaths)
```

The arguments of the function `GBMPaths` were `initPrice`, `mu`, `sigma`, `T`, `q`, `numSteps` and `numPaths`. Note that when we call the function `GBMPaths` from within the function `EuropeanCall`, we input `r - q` in place of the argument `mu`.

After generating `numPaths` paths for the price of the underlying asset, we compute the payoffs of the option. We only need the payoffs at the time of maturity of the option, time T , so we only use `paths(numSteps + 1, iPath)` in the calculation.

The payoff along path `iPath` is calculated as the maximum of zero and the difference between the strike price K and the value of the underlying at the end of path `iPath` at time T .

We use the Excel function `Max` to compute the maximum and call it as `Application.Max`. Each payoff is discounted, and is added to the array `payoffs`.

After the array `payoffs` is filled, we compute the average of the payoffs to get the price of the option. We use the Excel function `Average`, which we call with the command `Application.Average`.

KEY POINTS

- Macros contain prerecorded tasks that can be performed in a spreadsheet. Macros are in effect computer programs whose commands are hidden from the user, but they can be seen if the VBA editor is open.
- The most important fact about VBA is that it tries to act as an object-oriented language. This means that it treats every component of Excel, such as a worksheet, a cell, a range of cells, and a chart, as an object.
- Objects are arranged in a hierarchy and have properties (attributes) that can be modified by entering the name of the object followed by a dot and a specific command. In addition, objects are associated with actions (methods) that the objects can perform or have applied to them.
- Subroutines and user-defined functions in VBA are both blocks of code saved in

modules. The difference is that subroutines are general scripts; that is, lists of instructions, whereas functions complete a list of instructions and return a value to the user.

- Variable types in VBA include Integer, String, Single, Double, Long, Boolean, Date, Object, and Variant. A different amount of memory is allocated to storing values of variables of different types.
- The default in VBA is to index the first element in arrays as 0, which is the convention in most programming languages. The command `Option Base 1` at the beginning of a module makes sure that the indexing of arrays starts at 1.
- Control flow statements such as `For` and `If` allow for building more sophisticated programs than simple input and output of data to Excel.
- Excel functions can be accessed from VBA by prefixing them with `Application`.
- VBA has some built-in numeric functions, but it is important to know that their syntax is not always the same as the syntax of the same function in Excel. For example, the function `Sqrt` (square root) in Excel is `Sqr` in VBA.
- Useful tools in Excel and VBA that allow for interaction with users include buttons, input dialog boxes, and message boxes.
- VBA has debugging tools that allow you to look at the code in more detail if your programs do not work as expected. These tools can be accessed through commands under the **Debug** item in the top menu of the VBE.

NOTE

1. The notation E in Excel denotes multiplication by 10 to a specific power. For example, 5E40 means $5 \cdot 10^{40}$, and 5E-45 means $5 \cdot 10^{-45}$.

REFERENCES

- Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 3: 637–654.
- Boyle, P. (1977). Options: A Monte Carlo approach. *Journal of Financial Economics* 4, 3: 323–338.
- Boyle, P., Broadie, M., and Glasserman, P. (1997). Monte Carlo methods for security pricing. *Journal of Economic Dynamics & Control* 21: 1267–1321.
- Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. New York: Springer-Verlag.
- Hull, J. (2008). *Options, Futures and Other Derivatives*, 7th ed. Upper Saddle River, NJ: Prentice Hall.
- Jackson, M., and Staunton, M. (2001). *Advanced Modelling in Finance Using Excel and VBA*. Hoboken, NJ: John Wiley & Sons.
- Markowitz, H. (1952). Portfolio selection. *Journal of Finance* 7: 77–91.
- McLeish, D. (2005). *Monte Carlo Simulation and Finance*. Hoboken, NJ: John Wiley & Sons.
- Pachamanova, D. A., and Fabozzi, F. J. (2010). *Simulation and Optimization in Finance: Modeling with MATLAB, @RISK, and VBA*. Hoboken, NJ: John Wiley & Sons.
- Roman, S. (2002). *Writing Excel Macros with VBA*, 2nd ed. Sebastopol, CA: O'Reilly Media.
- Walkenbach, J. (2004). *Excel 2003 Power Programming with VBA*. Hoboken, NJ: John Wiley & Sons.

Stochastic Processes and Tools

Stochastic Integrals

SERGIO M. FOCARDI, PhD

Partner, The Intertek Group

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Calculus is an important tool because it provides two key ideas for financial modeling: (1) the concept of instantaneous rate of change, and (2) a framework and rules for linking together quantities and their instantaneous rates of change. Calculus made the concept of infinitely small quantities precise with the notion of limit. If the rate of change can get arbitrarily close to a definite number by making the time interval sufficiently small, that number is the instantaneous rate of change. The instantaneous rate of change is the limit of the rate of change when the length of the interval gets infinitely small. This limit is referred to as the derivative of a function, or simply derivative. Starting from this definition and with the help of a number of rules for computing a derivative, it was shown that the instantaneous rate of change of a number of functions can be explicitly computed as a closed formula. The process of computing a derivative, referred to as differentiation, solves the problem of finding the steepness of the tangent to a curve; the process of integration solves the problem of finding the area below a given curve. A key result of calculus is the discovery that integration and derivation are inverse operations: Integrating the derivative of a function yields the function itself. Standard calculus deals with deterministic functions. As such, there are limits as to the application of integration of determinist functions to financial modeling such as pricing contingent claims. The major application of integration to financial modeling involves stochastic integrals. An understanding of stochastic integrals is needed to understand an important tool in contingent claims valuation: stochastic differential equations.

In elementary calculus, integration is an operation performed on single, deterministic functions; the end product is another single, deterministic function. Integration defines a process of *cumulation*: The integral of a function represents the area below the function. However, the usefulness of deterministic functions in financial modeling is limited. Given the amount of uncertainty, few laws in finan-

cial theory can be expressed through them. It is necessary to adopt an ensemble view, where the path of economic variables must be considered a realization of a stochastic process, not a deterministic path. We must therefore move from deterministic integration to *stochastic integration*. In doing so we have to define how to *cumulate random shocks in a continuous-time environment*. These concepts require rigorous

definition. In this entry, we define the concept and the properties of stochastic integration. Based on the concept of stochastic integration, an important tool used in financial modeling, stochastic differential equations can be understood.

Two observations are in order. First, although ordinary integrals and derivatives operate on functions and yield either individual numbers or other functions, stochastic integration operates on stochastic processes and yields either random variables or other stochastic processes. Therefore, while a definite integral is a number and an indefinite integral is a function, a stochastic integral is a random variable or a stochastic process. A differential equation—when equipped with suitable initial or boundary conditions—admits as a solution a single function while a stochastic differential equation admits as a solution a stochastic process.

Second, moving from a deterministic to a stochastic environment does not necessarily require leaving the realm of standard calculus. In fact, all the stochastic laws of financial theory could be expressed as laws that govern the distribution of transition probabilities. An example of this mathematical strategy is the application of the forward Komogorov differential equation or the Fokker-Planck differential equation to term structure modeling, which are deterministic partial differential equations that govern the probability distributions of prices. Nevertheless it is often convenient to represent uncertainty directly through stochastic integration and stochastic differential equations. This approach is not limited to financial theory: It is also used in the domain of the physical sciences. In financial theory, stochastic differential equations have the advantage of being intuitive: Thinking in terms of a deterministic path plus an uncertain term is easier than thinking in terms of abstract probability distributions. There are other reasons why stochastic calculus is the methodology of choice in economics and finance but easy intuition plays a key role.

For example, a risk-free bank account, which earns a deterministic instantaneous interest rate $f(t)$, evolves according to the deterministic law:

$$y = A \exp\left(\int f(t)dt\right)$$

which is the general solution of the differential equation:

$$\frac{dy}{y} = f(t)dt$$

The solution of this differential equation tells us how the bank account cumulates over time.

However, if the rate is not deterministic but is subject to volatility—that is, at any instant the rate is $f(t)$ plus a random disturbance—then the bank account evolves as a stochastic process. That is to say, the bank account might follow any of an infinite number of different paths: Each path cumulates the rate $f(t)$ plus the random disturbance. In a sense that will be made precise in this entry and with an understanding of stochastic differential equations, we must solve the following equation:

$$\frac{dy}{y} = f(t)dt \text{ plus random disturbance}$$

Here is where stochastic integration comes into play: It defines how the stochastic rate process is transformed into the stochastic account process. This is the direct stochastic integration approach.

It is possible to take a different approach. At any instant t , the instantaneous interest rate and the cumulated bank account have two probability distributions. We could use a partial differential equation to describe how the probability distribution of the cumulated bank account is linked to the interest rate probability distribution.

Similar reasoning applies to stock and derivative price processes. In continuous-time finance, these processes are defined as stochastic processes that are the solution of a stochastic differential equation. Hence, the importance of stochastic integrals in continuous-time finance theory should be clear.

Following some remarks on the informal intuition behind stochastic integrals, we proceed to define Brownian motions and outlines the formal mathematical process through which stochastic integrals are defined. A number of properties of stochastic integrals are then established. After introducing stochastic integrals informally, we go on to define more rigorously the mathematical process for defining stochastic integrals.

THE INTUITION BEHIND STOCHASTIC INTEGRALS

Let's first contrast ordinary integration with *stochastic integration*. A definite integral

$$A = \int_a^b f(x)dx$$

is a number A associated to each function $f(x)$ while an indefinite integral

$$y(x) = \int_a^x f(s)ds$$

is a function y associated to another function f . The integral represents the cumulation of the infinite terms $f(s)ds$ over the integration interval.

A *stochastic integral*, which we will denote by

$$W = \int_a^b X_t dB_t$$

or

$$W = \int_a^b X_t \circ dB_t$$

is a random variable W associated to a stochastic process if the time interval is fixed or, if the time interval is variable, is another stochastic process W_t . The stochastic integral represents the cumulation of the stochastic products $X_t dB_t$. The rationale for this approach is that we need

to represent how random shocks feed back into the evolution of a process. We can cumulate separately the deterministic increments and the random shocks only for linear processes. In nonlinear cases, as in the simple example of the bank account, random shocks feed back into the process. For this reason we define stochastic integrals as the cumulation of the product of a process X by the random increments of a *Brownian motion*.

Consider a stochastic process X_t over an interval $[S, T]$. Recall that a stochastic process is a real variable $X(\omega)_t$ that depends on both time and the state of the economy ω . For any given ω , $X(\cdot)_t$ is a path of the process from the origin S to time T . A stochastic process can be identified with the set of its paths equipped with an appropriate probability measure. A stochastic integral is an integral associated to each path; it is a random variable that associates a real number, obtained as a limit of a sum, to each path. If we fix the origin and let the interval vary, then the stochastic integral is another stochastic process.

It would seem reasonable, *prima facie*, to define the stochastic integral of a process $X(\omega)_t$ as the definite integral in the sense of Riemann-Stieltjes associated to each path $X(\cdot)_t$ of the process. If the process $X(\omega)_t$ has continuous paths $X(\cdot, \omega)$, the integrals

$$W(\omega) = \int_S^T X(s, \omega)ds$$

exist for each path. However, as discussed in the previous section, this is not the quantity we want to represent. In fact, we want to represent the cumulation of the stochastic products $X_t dB_t$. Defining the integral

$$W = \int_a^b X_t dB_t$$

pathwise in the sense of Riemann-Stieltjes would be meaningless because the paths of a Brownian motion are not of *finite variation*. If we define

stochastic integrals simply as the limit of $X_t dB_t$ sums, the stochastic integral would be infinite (and therefore useless) for most processes.

However, Brownian motions have bounded *quadratic variation*. Using this property, we can define stochastic integrals pathwise through an approximation procedure. The approximation procedure to arrive at such a definition is far more complicated than the definition of the Riemann-Stieltjes integrals. Two similar but not equivalent definitions of stochastic integral have been proposed, the first by the Japanese mathematician Kiyoshi Ito (1951), the second by the Russian physicist Ruslan Stratonovich in the 1960s.¹ The definition of stochastic integral in the sense of Ito integral or of Stratonovich stochastic replaces the increments Δx_i with the increments ΔB_i of a fundamental stochastic process called Brownian motion.² The increments ΔB_i represent the “noise” of the process.

The definition proceeds in the following three steps:

- *Step 1.* The first step consists in defining a fundamental stochastic process—the *Brownian motion*. In intuitive terms, a Brownian motion $B_t(\omega)$ is a continuous limit (in a sense that will be made precise in the following sections) of a simple *random walk*. A simple random walk is a discrete-time stochastic process defined as follows. A point can move one step to the right or to the left. Movement takes place only at discrete instants of time, say at time 1, 2, 3, . . . At each discrete instant, the point moves to the right or to the left with probability $\frac{1}{2}$.

The random walk represents the cumulation of completely uncertain random shocks. At each point in time, the movement of the point is completely independent from its past movements. Hence, the Brownian motion represents the cumulation of random shocks in the limit of continuous time and of continuous states. It can be demonstrated that a.s. each path of the Brownian motion is not of

bounded total variation but it has bounded quadratic variation.

Recall that the total variation of a function $f(x)$ is the limit of the sums

$$\sum |f(x_i) - f(x_{i-1})|$$

while the quadratic variation is defined as the limit of the sums

$$\sum |f(x_i) - f(x_{i-1})|^2$$

Quadratic variation can be interpreted as the *absolute volatility* of a process. Thanks to this property, the ΔB_i of the Brownian motion provides the basic increments of the stochastic integral, replacing the Δx_i of the Riemann-Stieltjes integral.

- *Step 2.* The second step consists in defining the stochastic integral for a class of simple functions called *elementary functions*. Consider the time interval $[S, T]$ and any partition of the interval $[S, T]$ in N subintervals: $S \equiv t_0 < t_1 < \dots < t_i < \dots < t_N \equiv T$. An elementary function ϕ is a function defined on the time t and the outcome ω such that it assumes a constant value on the i -th subinterval. Call $I[t_{i+1}, t_i]$ the indicator function of the interval $[t_{i+1}, t_i]$. The indicator function of a given set is a function that assumes value 1 on the points of the set and 0 elsewhere. We can then write an elementary function ϕ as follows:

$$\phi(t, \omega) = \sum_i \varepsilon_i(\omega) I[t_{i+1}, t_i]$$

In other words, the constants $\varepsilon_i(\omega)$ are random variables and the function $\phi(t, \omega)$ is a stochastic process made up of paths that are constant on each i -th interval.

We can now define the stochastic integral, in the sense of Ito, of elementary functions $\phi(t, \omega)$ as follows:

$$\begin{aligned} W &= \int_S^T \phi(t, \omega) dB_t(\omega) \\ &= \sum_i \varepsilon_i(\omega) [B_{i+1}(\omega) - B_i(\omega)] \end{aligned}$$

where B is a Brownian motion.

It is clear from this definition that W is a random variable $\omega \rightarrow W(\omega)$. Note that the *Ito integral* thus defined for elementary functions cumulates the products of the elementary functions $\phi(t, \omega)$ and of the increments of the Brownian motion $B_i(\omega)$.

It can be demonstrated that the following property, called *Ito isometry*, holds for Ito stochastic integrals defined for bounded elementary functions as above:

$$E \left[\left(\int_S^T \phi(t, \omega) dB_i(\omega) \right)^2 \right] = E \left[\int_S^T \phi(t, \omega)^2 dt \right]$$

The Ito isometry will play a fundamental role in Step 3.

- *Step 3.* The third step consists in using the Ito isometry to show that each function g which is square-integrable (plus other conditions that will be made precise in the next section) can be approximated by a sequence of elementary functions $\phi_n(t, \omega)$ in the sense that

$$E \left[\int_S^T [g - \phi_n(t, \omega)]^2 dt \right] \rightarrow 0$$

If g is bounded and has a continuous time-path, the functions $\phi_n(t, \omega)$ can be defined as follows:

$$\phi_n(t, \omega) = \sum_i g(t_i, \omega) I[t_{i+1}, t_i)$$

where I is the indicator function. We can now use the Ito isometry to define the stochastic integral of a generic function $f(t, \omega)$ as follows:

$$\int_S^T f(t, \omega) dB_i(\omega) = \lim_{n \rightarrow \infty} \int_S^T \phi_n(t, \omega) dB_i(\omega)$$

The Ito isometry ensures that the Cauchy condition is satisfied and that the above sequence thus converges.

In outlining the above definition, we omitted an important point that will be dealt with in the next section: The definition of the stochastic integral in the sense of Ito requires that the el-

ementary functions be without anticipation—that is, they depend only on the past history of the Brownian motion. In fact, in the case of continuous paths, we wrote the approximating functions as follows:

$$\phi_n(t, \omega) = \sum_i g(t_i, \omega) [B_{i+1}(\omega) - B_i(\omega)]$$

taking the function g in the left extreme of each subinterval.

However, the definition of stochastic integrals in the sense of Stratonovich *admits anticipation*. In fact, the stochastic integral in the sense of Stratonovich, written as follows

$$\int_S^T f(t, \omega) \circ dB_i(\omega)$$

uses the following approximation under the assumption of continuous paths:

$$\phi_n(t, \omega) = \sum_i g(t_i^*, \omega) [B_{i+1}(\omega) - B_i(\omega)]$$

where

$$t_i^* = \frac{t_{i+1} - t_i}{2}$$

is the midpoint of the i -th subinterval.

Whose definition—Ito's or Stratonovich's—is preferable? Note that neither can be said to be correct or incorrect. The choice of the one over the other is a question of which one best represents the phenomena under study. The lack of anticipation is one reason why the Ito integral is generally preferred in finance theory.

We have just outlined the definition of stochastic integrals leaving aside mathematical details and rigor. The following two sections will make the above process mathematically rigorous and will discuss the question of anticipation of information. While these sections are a bit technical and might be skipped by those not interested in the mathematical details of stochastic calculus, they explain a number of concepts that are key to the modern development of finance theory.

BROWNIAN MOTION DEFINED

The previous section introduced Brownian motion informally as the limit of a simple random walk when the step size goes to zero. This section defines Brownian motion formally. The term “Brownian motion” is due to the Scottish botanist Robert Brown who in 1828 observed that pollen grains suspended in a liquid move irregularly. This irregular motion was later explained by the random collision of the molecules of the liquid with the pollen grains. It is therefore natural to represent Brownian motion as a continuous-time stochastic process that is the limit of a discrete random walk.

Let’s now formally define Brownian motion and demonstrate its existence. Let’s first go back to the probabilistic representation of the economy. The economy can be represented as a probability space $(\Omega, \mathfrak{S}, P)$, where Ω is the set of all possible economic states, \mathfrak{S} is the event σ -algebra, and P is a probability measure. The economic states $\omega \in \Omega$ are not instantaneous states but represent full histories of the economy for the time horizon considered, which can be a finite or infinite interval of time. In other words, the economic states are the possible realization outcomes of the economy.

In this probabilistic representation of the economy, time-variable economic quantities—such as interest rates, security prices or cash flows as well as aggregate quantities such as economic output—are represented as stochastic processes $X_t(\omega)$. In particular, the price and dividend of each stock are represented as two stochastic processes $S_t(\omega)$ and $d_t(\omega)$.

Stochastic processes are time-dependent random variables defined over the set Ω . It is critical to define stochastic processes so that there is no anticipation of information, that is, at time t no process depends on variables that will be realized later. Anticipation of information is possible only within a deterministic framework. However the space Ω in itself does not contain any coherent specification of time. If we asso-

ciate random variables $X_t(\omega)$ to a time index without any additional restriction, we might incur the problem of anticipation of information. Consider, for instance, an arbitrary family of time-indexed random variables $X_t(\omega)$ and suppose that, for some instant t , the relationship $X_t(\omega) = X_{t+1}(\omega)$ holds. In this case there is clearly anticipation of information as the value of the variable $X_{t+1}(\omega)$ at time $t + 1$ is known at an earlier time t . All relationships that lead to anticipation of information must be treated as deterministic.

The formal way to specify in full generality the evolution of time and the propagation of information without anticipation is through the concept of *filtration*. The concept of filtration is based on identifying all events that are known at any given instant. It is the propagation of information assuming that it is possible to associate to each moment t a σ -algebra of events $\mathfrak{S}_t \subset \mathfrak{S}$ formed by all events that are known prior to or at time t . It is assumed that events are never “forgotten,” that is, that $\mathfrak{S}_t \subset \mathfrak{S}_s$, if $t < s$. An increasing sequence of σ -algebras, each associated to the time at which all its events are known, represents the propagation of information. This sequence (called a filtration) is typically indicated as \mathfrak{S}_t .

The economy is therefore represented as a probability space $(\Omega, \mathfrak{S}, P)$ equipped with a filtration $\{\mathfrak{S}_t\}$. The key point is that every process $X_t(\omega)$ that represents economic or financial quantities must be *adapted* to the filtration $\{\mathfrak{S}_t\}$, that is, the random variable $X_t(\omega)$ must be measurable with respect to the σ -algebras \mathfrak{S}_t . In simple terms, this means that each event of the type $X_t(\omega) \leq x$ belongs to \mathfrak{S}_t while each event of the type $X_s(\omega) \leq y$ for $t \leq s$ belongs to \mathfrak{S}_s . For instance, consider a process $P_t(\omega)$, which might represent the price of a stock. Any coherent representation of the economy must ensure that events such as $\{\omega: P_s(\omega) \leq c\}$ are not known at any time $t < s$. The filtration $\{\mathfrak{S}_t\}$ prescribes all events admissible at time t .

Why do we have to use the complex concept of filtration? Why can’t we simply identify

information at time t with the values of all the variables known at time t as opposed to identifying a set of events? The principal reason is that in a continuous-time continuous-state environment any individual value has probability zero; we cannot condition on single values as the standard definition of conditional probability would become meaningless. In fact, in the standard definition of conditional probability, the probability of the conditioning event appears in the denominator and cannot be zero.

It is possible, however, to reverse this reasoning and construct a filtration starting from a process. Suppose that a process $X_t(\omega)$ does not admit any anticipation of information, for instance because the $X_t(\omega)$ are all mutually independent. We can therefore construct a filtration \mathfrak{F}_t as the strictly increasing sequence of σ -algebras generated by the process $X_t(\omega)$. Any other process must be adapted to \mathfrak{F}_t .

Let's now go back to the definition of the Brownian motion. Suppose that a probability space $(\Omega, \mathfrak{F}, P)$ equipped with a filtration \mathfrak{F}_t is given. A *one-dimensional standard Brownian motion* is a stochastic process $B_t(\omega)$ with the following properties:

- $B_t(\omega)$ is defined over the probability space $(\Omega, \mathfrak{F}, P)$.
- $B_t(\omega)$ is continuous for $0 \leq t < \infty$.
- $B_0(\omega) = 0$.
- $B_t(\omega)$ is adapted to the filtration \mathfrak{F}_t .
- The increments $B_t(\omega) - B_s(\omega)$ are independent and normally distributed with variance $(t-s)$ and zero mean.

The above conditions³ state that the standard Brownian motion is a stochastic process that starts at zero, has continuous paths and normally distributed increments whose variance grows linearly with time. Note that in the last condition the increments are independent of the σ -algebra \mathfrak{F}_s and not of the previous values of the process. As noted above, this is because any single realization of the process has probability zero and it is therefore impossible to use the standard concept of conditional proba-

bility: Conditioning must be with respect to a σ -algebra \mathfrak{F}_s . Once this concept has been firmly established, one might speak loosely of independence of the present values of a process from its previous values. It should be clear, however, that what is meant is independence with respect to a σ -algebra \mathfrak{F}_s .

Note also that the filtration \mathfrak{F}_t is an integral part of the above definition of the Brownian motion. This does not mean that, given any probability space and any filtration, a standard Brownian motion with these characteristics exists. For instance, the filtration generated by a discrete-time continuous-state random walk is insufficient to support a Brownian motion. The definition states only that we call a one-dimensional standard Brownian motion a mathematical object (if it exists) made up of a probability space, a filtration, and a time dependent random variable with the properties specified in the definition.

However, it can be demonstrated that Brownian motions exist by constructing them. Several construction methodologies have been proposed, including methodologies based on the Kolmogorov extension theorem or on constructing the Brownian motion as the limit of a sequence of discrete random walks. To prove the existence of the standard Brownian motion, we will use the *Kolmogorov extension theorem*.

The Kolmogorov theorem can be summarized as follows. Consider the following family of probability measures

$$\begin{aligned} &\mu_{t_1, \dots, t_m}(H_1 \times \dots \times H_m) \\ &= P[(X_{t_1} \in H_1, \dots, X_{t_m} \in H_m), H_i \in \mathcal{B}^n] \end{aligned}$$

for all $t_1, \dots, t_k \in [0, \infty)$, $k \in N$ and where the H s are n -dimensional Borel sets. Suppose that the following two consistency conditions are satisfied

$$\begin{aligned} &\mu_{t_{\sigma(1)}, \dots, t_{\sigma(m)}}(H_1 \times \dots \times H_m) \\ &= \mu_{t_1, \dots, t_m}(H_{\sigma^{-1}(1)} \times \dots \times H_{\sigma^{-1}(m)}) \end{aligned}$$

for all permutations σ on $\{1, 2, \dots, k\}$, and

$$\begin{aligned} & \mu_{t_1, \dots, t_k}(H_1 \times \dots \times H_k) \\ &= \mu_{t_1, \dots, t_k, t_{k+1}, \dots, t_m}(H_1 \times \dots \times H_k \times R^n \times \dots \times R^n) \end{aligned}$$

for all m . The Kolmogorov extension theorem states that, if the above conditions are satisfied, then there is (1) a probability space $(\Omega, \mathfrak{F}, P)$ and (2) a stochastic process that admits the probability measures

$$\begin{aligned} & \mu_{t_1, \dots, t_m}(H_1 \times \dots \times H_m) \\ &= P[(X_{t_1} \in H_1, \dots, X_{t_m} \in H_m), H_i \in \mathcal{B}^n] \end{aligned}$$

as finite dimensional distributions.

The construction is lengthy and technical and we omit it here, but it should be clear how, with an appropriate selection of finite-dimensional distributions, the Kolmogorov extension theorem can be used to prove the existence of Brownian motions. The finite-dimensional distributions of a one-dimensional Brownian motion are distributions of the type

$$\begin{aligned} & \mu_{t_1, \dots, t_k}(H_1 \times \dots \times H_k) \\ &= \int p(t, x, x_1) p(t_2 - t_1, x_1, x_2) \dots \\ & \quad p(t_k - t_{k-1}, x_{k-1}, x_k) dx_1 \dots dx_k H_1 \times \dots \times H_k \end{aligned}$$

where

$$p(t, x, y) = (2\pi t)^{-\frac{1}{2}} \exp\left(-\frac{|x - y|^2}{2t}\right)$$

and with the convention that the integrals are taken with respect to the Lebesgue measure. The distribution $p(t, x, x_1)$ in the integral is the initial distribution. If the process starts at zero, $p(t, x, x_1)$ is a Dirac delta, that is, it is a distribution of mass 1 concentrated in one point.

It can be verified that these distributions satisfy the above consistency conditions; the Kolmogorov extension theorem therefore ensures that a stochastic process with the above finite dimensional distributions exists. It can be demonstrated that this process has normally distributed independent increments with variance that grows linearly with time. It is

therefore a one-dimensional Brownian motion. These definitions can be easily extended to an n -dimensional Brownian motion.

In the initial definition of a Brownian motion, we assumed that a filtration \mathfrak{F}_t was given and that the Brownian motion was adapted to the filtration. In the present construction, however, we reverse this process. Given that the process we construct has normally distributed, stationary, independent increments, we can define the filtration \mathfrak{F}_t as the filtration \mathfrak{F}_t^B generated by $B_t(\omega)$. The independence of the increments of the Brownian motion guarantee the absence of anticipation of information. Note that if we were given a filtration \mathfrak{F}_t larger than the filtration \mathfrak{F}_t^B , $B_t(\omega)$ would still be a Brownian motion with respect to \mathfrak{F}_t .

In stochastic differential equations, there are two types of solutions of stochastic differential equations—strong and weak—depending on whether the filtration is given or generated by the Brownian motion. The implications of these differences for economics and finance will be discussed in the same section.

The above construction does not specify uniquely the Brownian motion. In fact, there are infinite stochastic processes that start from the same point and have the same finite dimensional distributions but have totally different paths. However, it can be demonstrated that only one Brownian motion has continuous paths a.s. (*a.s.* means almost surely; that is, for all paths except a set of measure zero). This process is called the *canonical Brownian motion*. Its paths can be identified with the space of continuous functions.

The Brownian motion can also be constructed as the continuous limit of a discrete random walk. Consider a simple random walk W_i where i are discrete time points. The random walk is the motion of a point that moves Δx to the right or to the left with equal probability $1/2$ at each time increment Δx . The total displacement X_i at time i is the sum of i independent increments each distributed as a Bernoulli variable. Therefore the random variable X has a binomial

distribution with mean zero and variance:

$$\frac{\Delta^2 x}{\Delta t}$$

Suppose that both the time increment and the space increment approach zero: $\Delta t \rightarrow 0$ and $\Delta x \rightarrow 0$. Note that this is a very informal statement. In fact what we mean is that we can construct a sequence of random walk processes W_i^n , each characterized by a time step and by a time displacement. It can be demonstrated that if

$$\frac{\Delta^2 x}{\Delta t} \rightarrow \sigma$$

(i.e., the square of the spaced interval and the time interval are of the same order) then the sequence of random walks approaches a Brownian motion. Though this is intuitive as the binomial distributions approach normal distributions, it should be clear that it is far from being mathematically obvious.

Figure 1 illustrates 100 realizations of a Brownian motion approximated as a random walk. The exhibit clearly illustrates that the standard

deviation grows with the square root of the time as the variance grows linearly with time. In fact, as illustrated, most paths remain confined within a parabolic region.

PROPERTIES OF BROWNIAN MOTION

The paths of a Brownian motion are rich structures with a number of surprising properties. It can be demonstrated that the paths of a canonical Brownian motion, though continuous, are nowhere differentiable. It can also be demonstrated that they are fractals of fractal dimension $3/2$. The fractal dimension is a concept that measures quantitatively how a geometric object occupies space. A straight line has fractal dimension one, a plane has fractal dimension two, and so on. Fractal objects might also have intermediate dimensions. This is the case, for example, of the path of a Brownian motion, which is so jagged that, in a sense, it occupies more space than a straight line.

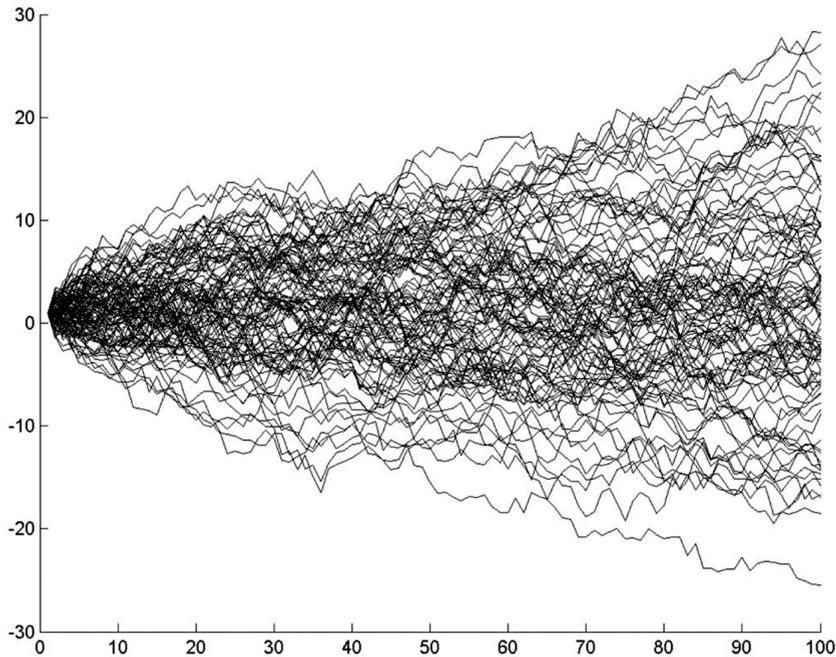


Figure 1 Illustration of 100 Paths of a Brownian Motion Generated as an Arithmetic Random Walk

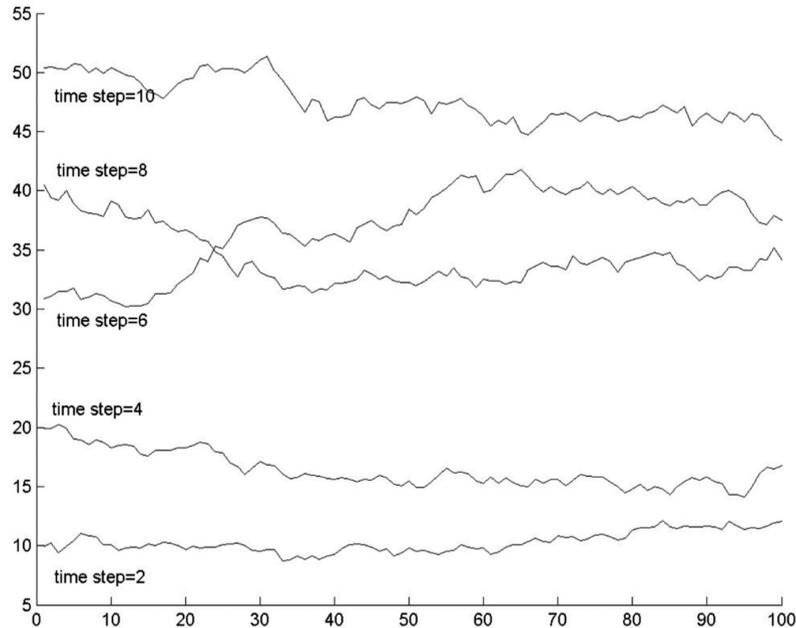


Figure 2 Illustration of the Fractal Properties of the Paths of a Brownian Motion

Note: Five paths of a Brownian motion are generated as random walks with different time steps and then magnified.

The fractal nature of Brownian motion paths implies that each path is a self-similar object. This property can be illustrated graphically. If we generate random walks with different time steps, we obtain jagged paths. If we allow paths to be graphically magnified, all paths look alike regardless of the time step with which they have been generated. In Figure 2, sample paths are generated with different time steps and then portions of the paths are magnified. Note that they all look perfectly similar.

This property was first observed by Mandelbrot (1963) in sequences of cotton prices in the 1960s. In general, if one looks at asset or commodity price time series, it is difficult to recognize their time scale. For instance, weekly or monthly time series look alike. (Recent empirical and theoretical research work has made this claim more precise.)

Let's consider a one-dimensional standard Brownian motion. If we wait a sufficiently long period of time, every path except a set of paths of measure zero will return to the ori-

gin. The path between two consecutive passages through zero is called an *excursion* of the Brownian motion. The distribution of the maximum height attained by an excursion and of the time between two passages through zero or through any level have interesting properties. The distribution of the time between two passages through zero has infinite mean. This is at the origin of the so-called St. Petersburg paradox described by the Swiss mathematician Bernoulli. The paradox consists of the following. Suppose a player bets increasing sums on a game that can be considered a realization of a random walk. As the return to zero of a random walk is a sure event, the player is certain to win—but while the probability of winning is one, the average time before winning is infinite. To stay the game, the capital required is also infinite. Difficult to imagine a banker ready to put up the money to back the player.

The distribution of the time to the first passage through zero of a Brownian motion is not Gaussian. In fact, the probability of a very long

waiting time before the first return to zero is much higher than in a normal distribution. It is a fat-tailed distribution in the sense that it has more weight in the tail regions than a normal distribution. The distribution of the time to the first passage through zero of a Brownian motion is an example of how fat-tailed distributions can be generated from Gaussian variables.

STOCHASTIC INTEGRALS DEFINED

Let's now go back to the definition of stochastic integrals, starting with one-dimensional stochastic integrals. Suppose that a probability space $(\Omega, \mathfrak{F}, P)$ equipped with a filtration \mathfrak{F}_t is given. Suppose also that a Brownian motion $B_t(\omega)$ adapted to the filtration \mathfrak{F}_t is given. We will define Ito integrals following the three-step procedure outlined earlier in this entry. We have just completed the first step defining Brownian motion. The second step consists in defining the Ito integral for elementary functions.

Let's first define the set $\Phi(S, T)$ of functions $\phi(S, T) \equiv \{f(t, \omega): [(0, \infty) \times \Omega \rightarrow R]\}$ with the following properties:

- Each f is jointly $\mathcal{B} \times \mathfrak{F}$ measurable.
- Each $f(t, \omega)$ is adapted to \mathfrak{F}_t .
- $E \left[\int_S^T f^2(t, \omega) dt \right] < \infty$ (this condition can be weakened).

This is the set of paths for which we define the Ito integral.

Consider the time interval $[S, T]$ and, for each integer n , partition the interval $[S, T]$ in subintervals: $t_0 < t_1 < \dots < t_n < \dots < t_N \equiv T$ in this way:

$$t_k = t_k^n = \begin{cases} k2^{-n} & \text{if } S \leq k2^{-n} \leq T \\ S & \text{if } k2^{-n} < S \\ T & \text{if } k2^{-n} > T \end{cases}$$

This rule provides a family of partitions of the interval $[S, T]$ which can be arbitrarily refined.

Consider the elementary functions $\phi(t, \omega) \in \Phi$ which we write as

$$\phi(t, \omega) = \sum_i \varepsilon_i(\omega) I[t_{i+1} - t_i]$$

As $\phi(t, \omega) \in \Phi$, $\varepsilon_i(\omega)$ are \mathfrak{F}_{t_i} measurable random variables.

We can now define the stochastic integral, in the sense of Ito, of elementary functions $\phi(t, \omega)$ as

$$W = \int_S^T \phi(t, \omega) dB_t(\omega) = \sum_{i \geq 0} \varepsilon_i(\omega) [B_{t_{i+1}}(\omega) - B_{t_i}(\omega)]$$

where B is a Brownian motion. Note that the $\varepsilon_i(\omega)$ and the increments $B_j(\omega) - B_i(\omega)$ are independent for $j > i$. The key aspect of this definition that was not included in the informal outline is the condition that the $\varepsilon_i(\omega)$ are \mathfrak{F}_{t_i} measurable.

For bounded elementary functions $\phi(t, \omega) \in \Phi$ the Ito isometry holds

$$E \left[\left(\int_S^T \phi(t, \omega) dB_t(\omega) \right)^2 \right] = E \left[\int_S^T \phi(t, \omega)^2 dt \right]$$

The demonstration of the Ito isometry rests on the fact that

$$E[\varepsilon_i \varepsilon_j (B_{t_{i+1}} - B_{t_i})(B_{t_{j+1}} - B_{t_j})] = \begin{cases} 0 & \text{if } i \neq j \\ E(\varepsilon_i^2) & \text{if } i = j \end{cases}$$

This completes the definition of the stochastic integral for elementary functions.

We have now completed the introduction of Brownian motions and defined the Ito integral for elementary functions. Let's next introduce the approximation procedure that allows us to define the stochastic integral for any $\phi(t, \omega)$. We will develop the approximation procedure in the following three additional steps that we will state without demonstration:

- *Step 1.* Any function $g(t, \omega) \in \Phi$ that is bounded and such that all its time paths $\phi(\cdot, \omega)$ are

continuous functions of time can be approximated by

$$\phi_n(t, \omega) = \sum_i g(t_i, \omega) I[t_{i+1} - t_i]$$

in the sense that:

$$E \int_S^T [(g - \phi_n)^2 dt] \rightarrow 0, \quad n \rightarrow \infty, \forall \omega$$

where the intervals are those of the partition defined above. Note that $\phi_n(t, \omega) \in \Phi$ given that $g(t, \omega) \in \Phi$.

- *Step 2.* We release the condition of time-path continuity of the $\phi_n(t, \omega)$. It can be demonstrated that any function $h(t, \omega) \in \Phi$ which is bounded but not necessarily continuous can be approximated by functions $g_n(t, \omega) \in \Phi$, which are bounded and continuous in the sense that

$$E \left[\int_S^T (h - g_n)^2 dt \right] \rightarrow 0$$

- *Step 3.* It can be demonstrated that any function $f(t, \omega) \in \Phi$, not necessarily bounded or continuous, can be approximated by a sequence of bounded functions $h_n(t, \omega) \in \Phi$ in the sense that

$$E \left[\int_S^T (f - h_n)^2 dt \right] \rightarrow 0$$

We now have all the building blocks to complete the definition of Ito stochastic integrals. In fact, by virtue of the above three-step approximation procedure, given any function $f(t, \omega) \in \Phi$, we can choose a sequence of elementary functions $\phi_n(t, \omega) \in \Phi$ such that the following property holds:

$$E \left[\int_S^T (f - \phi_n)^2 dt \right] \rightarrow 0$$

Hence we can define the Ito stochastic integral as follows:

$$I[f](\omega) = \int_S^T f(t, \omega) dB_t(\omega) = \lim_{n \rightarrow \infty} \left[\int_S^T \phi_n(t, \omega) dt \right]$$

The limit exists as

$$\int_S^T \phi_n(t, \omega) dB_t(\omega)$$

forms a Cauchy sequence by the Ito isometry, which holds for every bounded elementary function.

Let's now summarize the definition of the Ito stochastic integral: Given any function $f(t, \omega) \in \Phi$, we define the Ito stochastic integral by

$$I[f](\omega) = \int_S^T f(t, \omega) dB_t(\omega) = \lim_{n \rightarrow \infty} \left[\int_S^T \phi_n(t, \omega) dt \right]$$

where the functions $\phi_n(t, \omega) \in \Phi$ are a sequence of elementary functions such that

$$E \left[\int_S^T (f - \phi_n)^2 dt \right] \rightarrow 0$$

The multistep procedure outlined above ensures that the sequence $\phi_n(t, \omega) \in \Phi$ exists. In addition, it can be demonstrated that the Ito isometry holds in general for every $f(t, \omega) \in \Phi$

$$E \left[\left(\int_S^T f(t, \omega) dB_t(\omega) \right)^2 \right] = E \left[\int_S^T f(t, \omega)^2 dt \right]$$

SOME PROPERTIES OF ITO STOCHASTIC INTEGRALS

Suppose that $f, g \in \Phi(S, T)$ and let $0 < S < U < T$. It can be demonstrated that the following

properties of Ito stochastic integrals hold:

$$\int_s^T f dB_t = \int_s^u f dB_t + \int_u^T f dB_t \text{ for a.a. } \omega$$

$$E \left[\int_s^T f dB_t \right] = 0$$

$$\int_s^T (cf + dg) dB_t = c \int_s^T f dB_t + d \int_s^T g dB_t,$$

for a.a. ω, c, d constants

If we let the time interval vary, say $(0, t)$, then the stochastic integral becomes a stochastic process:

$$I_t(\omega) = \int_0^t f dB_t$$

It can be demonstrated that a continuous version of this process exists. The following three properties can be demonstrated from the definition of integral:

$$\int_0^t dB_s = B_t$$

$$\int_0^t s dB_s = tB_t - \int_0^t B_s ds$$

$$\int_0^t B_s dB_s = \frac{1}{2} B_t^2 - \frac{1}{2} t$$

The last two properties show that, after performing stochastic integration, deterministic terms might appear.

KEY POINTS

- Stochastic integration provides a coherent way to represent that instantaneous uncer-

tainty (or volatility) cumulates over time. It is thus fundamental to the representation of financial processes such as interest rates, security prices, or cash flows as well as aggregate quantities such as economic output.

- Stochastic integration operates on stochastic processes and produces random variables or other stochastic processes.
- Stochastic integration is a process defined on each path as the limit of a sum. However, these sums are different from the sums of the Riemann-Lebesgue integrals because the paths of stochastic processes are generally not of bounded variation.
- Stochastic integrals in the sense of Ito are defined through a process of approximation.
- Step 1 consists in defining Brownian motion, which is the continuous limit of a random walk.
- Step 2 consists in defining stochastic integrals for elementary functions as the sums of the products of the elementary functions multiplied by the increments of the Brownian motion.
- Step 3 extends this definition to any function through approximating sequences.

NOTES

1. The publications of Stratonovich can be found in Romanovski (2007).
2. A history of stochastic integrations and financial mathematics is provided by Jarrow and Protter (2004). For a more detailed discussion of stochastic integration, see Protter (1990).
3. The set of conditions defining a Brownian motion can be more parsimonious. If a process has stationary, independent increments and continuous paths a.s. it must have normally distributed increments. A process with stationary independent increments and with paths that are continuous to the right and limited to the left (the *cadlag* functions) is called a Levy process.

REFERENCES

- Ito, K. (1951). On stochastic differential equations. *Memoirs, American Mathematical Society* 4: 1–51.
- Jarrow, R., and Protter, P. (2004). A short history of stochastic integration and mathematical finance: The early years, 1880–1970. *IMS Lecture Notes Monograph* 45: 1–17.
- Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business* 36: 394–419.
- Protter, P. (1990). *Stochastic Integration and Differential Equations*. New York: Springer.
- Romanovski, Y. M. (2007). *Professor R. L. Stratonovich: Reminiscences of Relatives, Colleagues and Friends*. Moscow-Izhevsk: Publishing House of Computer Research Institute.

Stochastic Differential Equations

SERGIO M. FOCARDI, PhD
Partner, The Intertek Group

FRANK J. FABOZZI, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

Abstract: In nontechnical terms, differential equations are equations that express a relationship between a function and one or more derivatives (or differentials) of that function. It would be difficult to overemphasize the importance of differential equations in financial modeling where they are used to express laws that govern the evolution of price probability distributions, the solution of economic variational problems (such as intertemporal optimization), and conditions for continuous hedging (such as in the Black-Scholes equation). The two broad types of differential equations are ordinary differential equations and partial differential equations. The former are equations or systems of equations involving only one independent variable; the latter are differential equations or systems of equations involving partial derivatives. When one or more of the variables is a stochastic process, we have the case of stochastic differential equations and the solution is also a stochastic process. An assumption must be made about driving noise in a stochastic differential equation. In most applications, it is assumed that the noise term follows a Gaussian random variable, although types of random variables can be assumed.

Stochastic differential equations solve the problem of giving meaning to a differential equation where one or more of its terms are subject to random fluctuations. For instance, consider the following deterministic equation:

$$\frac{dy}{dt} = f(t)y$$

We know from differential equations that, by separating variables, the general solution of this equation can be written as follows:

$$y = A \exp \left[\int f(t) dt \right]$$

A stochastic version of this equation might be obtained, for instance, by perturbing the term f , thus resulting in the *stochastic differential equation*

$$\frac{dy}{y} = [f(t) + \varepsilon] dt$$

where ε is a random noise process.

As with stochastic integrals, in defining stochastic differential equations it is necessary to adopt an ensemble view: The solution of a stochastic differential equation is a stochastic process, not a single function. In this entry, we first provide the basic intuition behind

stochastic differential equations and then proceed to formally define the concept and the properties.

THE INTUITION BEHIND STOCHASTIC DIFFERENTIAL EQUATIONS

Let's go back to the equation

$$\frac{dy}{dt} = [f(t) + \varepsilon]y$$

where ε is a continuous-time noise process. It would seem reasonable to define a continuous-time noise process informally as the continuous-time limit of a zero-mean, IID sequence, that is, a sequence of independent and identically distributed variables with zero mean. In a discrete time setting, a zero-mean, IID sequence is called a *white noise*. We could envisage defining a continuous-time white noise as the continuous-time limit of a discrete-time white noise. Each path of ε is a function of time $\varepsilon(\cdot, \omega)$. It would therefore seem reasonable to define the solution of the equation pathwise, as the family of functions that are solutions of the equations

$$\frac{dy}{dt} = [f(t) + \varepsilon(t, \omega)]y$$

where each equation corresponds to a specific white noise path.

However, this definition would be meaningless in the domain of ordinary functions. In other words, it would generally not be possible to find a family of functions $y(\cdot, \omega)$ that satisfy the above equations for each white-noise path and that form a reasonable stochastic process.

The key problem is that it is not possible to define a white noise process as a zero-mean stationary stochastic process with independent increments and continuous paths. Such a process does not exist in the domain of ordinary functions.¹ In discrete time the white noise process is obtained as the first-difference process of a random walk. *Random walk* is an integrated

nonstationary process, while its first-difference process is a stationary IID sequence.

The continuous-time limit of the random walk is the *Brownian motion*. However, the paths of a Brownian motion are not differentiable. As a consequence, it is not possible to take the continuous-time limit of first differences and to define the white noise process as the derivative of a Brownian motion. In the domain of ordinary functions in continuous time, the white noise process can be defined only through its integral, which is the Brownian motion. The definition of stochastic differential equations must therefore be recast in integral form.

A sensible definition of a stochastic differential equation must respect a number of constraints. In particular, the solution of a stochastic differential equation should be a "perturbation" of the associated deterministic equation. In the above example, for instance, we want the solution of the stochastic equation

$$\frac{dy}{dy} = [f(t) + \varepsilon(t, \omega)]dt$$

to be a perturbation of the solution

$$y = A \exp\left(\int f(t)dt\right)$$

of the associated deterministic equation

$$\frac{dy}{y} = f(t)dt$$

In other words, the solution of a stochastic differential equation should tend to the solution of the associated deterministic equation in the limit of zero noise. In addition, the solutions of a stochastic differential equation should be the continuous-time limit of some discrete-time process obtained by discretization of the stochastic equation.

A formal solution of this problem was proposed by Kiyoshi Itô (1951) and, in a different setting, by Ruslan Stratonovich in the 1960s. Itô and Stratonovich proposed to give meaning to a stochastic differential equation through its integral equivalent. The Itô definition proceeds in two steps: In the first step, Itô processes are

defined; in the second step, stochastic differential equations are defined.

- *Step 1: Definition of Itô processes.* Given two functions $\varphi(t, \omega)$ and $\psi(t, \omega)$ that satisfy usual conditions to be defined later, an *Itô process*—also called a **stochastic integral**—is a stochastic process of the form:

$$Z(t, \omega) = \int_0^t \varphi(s, \omega) ds + \int_0^t \psi(s, \omega) dB_s(s, \omega)$$

An Itô process is a process that is the result of the sum of two summands: The first is an ordinary integral, the second an Itô integral. Itô processes are stable under smooth maps, that is, any smooth function of an Itô process is an Itô process that can be determined through the Itô formula (see Itô processes below).

- *Step 2: Definition of stochastic differential equations.* As we have seen, it is not possible to write a differential equation plus a white-noise term that admits solutions in the domain of ordinary functions. However, we can meaningfully write an integral stochastic equation of the form

$$X(t, \omega) = \int_0^t \varphi(s, X) ds + \int_0^t \psi(s, X) dB_s$$

It can be demonstrated that this equation admits solutions in the sense that, given two functions φ and ψ , there is a stochastic process X that satisfies the above equation. We stipulate that the above integral equation can be written in differential form as follows:

$$dX(t, \omega) = \varphi(t, X)dt + \psi(t, X)dB_t$$

Note that this is a definition; a stochastic differential equation acquires meaning only through its integral form. In particular, we *cannot* divide both terms by dt and rewrite the equation as follows:

$$\frac{dX(t, \omega)}{dt} = \varphi(t, X) + \psi(t, X) \frac{dB_t}{dt}$$

The above equation would be meaningless because the Brownian motion is not differentiable.

This is the difficulty that precludes writing stochastic differential equations adding white noise pathwise. The differential notation of a stochastic differential equation is just a shorthand for the integral notation.

However, we can consider a discrete approximation:

$$\Delta X(t, \omega) = \varphi^*(t, X)\Delta t + \psi^*(t, X)\Delta B_t$$

Note that in this approximation the functions $\varphi^*(t, X)$, $\psi^*(t, X)$ will not coincide with the functions $\varphi(t, X)$, $\psi(t, X)$. Using the latter would (in general) result in a poor approximation.

The following sections will define Itô processes and stochastic differential equations and study their properties.

ITÔ PROCESSES

Let's now formally define Itô processes and establish key properties, in particular the Itô formula. In the previous section we stated that an Itô process is a stochastic process of the form

$$Z(t, \omega) = \int_0^t a(s, \omega) ds + \int_0^t b(s, \omega) dB(s, \omega)$$

To make this definition rigorous, we have to state the conditions under which (1) the integrals exist, and (2) there is no anticipation of information. Note that the two functions a and b might represent two stochastic processes and that the Riemann-Stieltjes integral might not exist for the paths of a stochastic process. We have therefore to demonstrate that both the Itô integral and the ordinary integral exist. To this end, we define Itô processes as follows.

Suppose that a one-dimensional Brownian motion B_t is defined on a probability space $(\Omega, \mathfrak{F}, P)$ equipped with a filtration \mathfrak{F}_t . The filtration might be given or might be generated by the Brownian motion B_t . Suppose that both a and b are adapted to \mathfrak{F}_t and jointly measurable in $\mathfrak{F} \times \mathfrak{R}$. Suppose, in addition, that the

following two integrability conditions hold:

$$P \left[\int_0^t b^2(s, \omega) ds < \infty \text{ for all } t \geq 0 \right] = 1$$

and

$$P \left[\int_0^t |a(s, \omega)| ds < \infty \text{ for all } t \geq 0 \right] = 1$$

These conditions ensure that both integrals in the definition of Itô processes exist and that there is no anticipation of information. We can therefore define the Itô process as the following stochastic process:

$$Z(t, \omega) = \int_0^t a(s, \omega) ds + \int_0^t b(s, \omega) dB_s(s, \omega)$$

Itô processes can be written in the shorter differential form as

$$dZ_t = a dt + b dB_t$$

It should be clear that the latter formula is just a shorthand for the integral definition.

THE ONE-DIMENSIONAL ITÔ FORMULA

One of the most important results concerning Itô processes is a formula established by Itô that allows one to explicitly write down an Itô process that is a function of another Itô process. Itô's formula is the stochastic equivalent of the change-of-variables formula of ordinary integration. We will proceed in two steps. First we will introduce Itô's formula for functions of Brownian motion and then for functions of general Itô processes. Suppose that the function $g(t, x)$ is twice continuously differentiable in $[0, \infty) \times R$ and that B_t is a one-dimensional Brownian motion. The function $Y_t = g(t, B_t)$ is a stochastic process. It can be demonstrated that the process $Y_t = g(t, B_t)$ is an Itô process of the following form

$$dY_t = \left(\frac{\partial g}{\partial t}(t, B_t) + \frac{1}{2} \frac{\partial^2 g}{\partial x^2}(t, B_t) \right) dt + \frac{\partial g}{\partial x}(t, B_t) dB_t$$

The preceding is Itô's formula in the case the underlying process is a Brownian motion. For example, let's suppose that $g(t, x) = x^2$. In this case we can write

$$\frac{\partial g}{\partial t} = 0, \quad \frac{\partial g}{\partial x} = 2x, \quad \frac{\partial^2 g}{\partial x^2} = 2$$

Inserting the above in Itô's formula we see that the process B_t^2 can be represented as the following Itô process

$$dY_t = dt + 2B_t dB_t$$

or, explicitly in integral form

$$Y_t = t + 2 \int_0^t B_s dB_s$$

The nonlinear map $g(t, x) = x^2$ introduces a second term in dt .

Let's now generalize Itô's formula. Suppose that X_t is an Itô process given by $dX_t = a dt + b dB_t$. As X_t is a stochastic process, that is, a function $X(t, \omega)$ of both time and the state, it makes sense to consider another stochastic process Y_t , which is a function of the former, $Y_t = g(t, X_t)$. Suppose that g is twice continuously differentiable on $[0, \infty) \times R$.

It can then be demonstrated (we omit the detailed proof) that Y_t is another Itô process that admits the representation

$$dY_t = \frac{\partial g}{\partial t}(t, X_t) dt + \frac{\partial g}{\partial x}(t, X_t) dX_t + \frac{1}{2} \frac{\partial^2 g}{\partial x^2}(t, X_t) (dX_t)^2$$

where differentials are computed formally according to the rules known as Box algebra

$$dt \cdot dt = dt \cdot dB_t = dB_t \cdot dt = 0, \quad dB_t \cdot dB_t = dt$$

Itô's formula can be written (perhaps more) explicitly as

$$dY_t = \left(\frac{\partial g}{\partial t} + \frac{\partial g}{\partial x} a + \frac{1}{2} \frac{\partial^2 g}{\partial x^2} b^2 \right) dt + \frac{\partial g}{\partial x} b dB_t$$

This formula reduces to the ordinary formula for the differential of a compound function in the case where $b = 0$ (that is, when there is no noise).

As a second example of application of Itô's formula, consider the geometric Brownian motion:

$$dX_t = \mu X_t dt + \sigma X_t dB_t$$

where μ, σ are real constants, and consider the map $g(t, x) = \log x$. In this case, we can write

$$\frac{\partial g}{\partial t} = 0, \frac{\partial g}{\partial x} = \frac{1}{x}, \frac{\partial^2 g}{\partial x^2} = -\frac{1}{x^2}$$

and Itô's formula yields

$$dY_t = d \log X_t = \left(\mu - \frac{1}{2} \sigma^2 \right) dt + \sigma dB_t$$

STOCHASTIC DIFFERENTIAL EQUATIONS

An Itô process defines a process $Z(t, \omega)$ as the sum of the time integral of the process $a(t, \omega)$ plus the Itô integral of the process $b(t, \omega)$. Suppose that two functions $\varphi(t, x), \psi(t, x)$ that satisfy conditions established below are given. Given an Itô process $X(t, \omega)$, the two processes $\varphi(t, X), \psi(t, X)$ admit respectively a time integral and an Itô integral. It therefore makes sense to consider the following Itô process:

$$Z(t, \omega) = \int_0^t \varphi[s, X(s, \omega)] ds + \int_0^t \psi[s, X(s, \omega)] dB_s$$

The term on the right side transforms the process X into a new process Z . We can now ask if there are stochastic processes X that are mapped into themselves such that the following stochastic equation is satisfied:

$$X(t, \omega) = \int_0^t \varphi[s, X(s, \omega)] ds + \int_0^t \psi[s, X(s, \omega)] dB_s$$

The answer is positive under appropriate conditions. It is possible to prove the following theorem of existence and uniqueness. Suppose that a one-dimensional Brownian motion B_t is defined on a probability space $(\Omega, \mathfrak{F}, P)$ equipped with a filtration \mathfrak{F}_t and that B_t is adapted to the filtration \mathfrak{F}_t . Suppose also that the two measurable functions $\varphi(t, x), \psi(t, x)$ map $[0,$

$T] \times R \rightarrow R$ and that they satisfy the following conditions:

$$\begin{aligned} |\varphi(t, x)|^2 + |\psi(t, x)|^2 &\leq C(1 + |x|)^2, \\ t &\in [0, T], x \in R \end{aligned}$$

and

$$\begin{aligned} |\varphi(t, x) - \varphi(t, y) + |\psi(t, x) - \psi(t, y)| \\ \leq D|x - y|, t \in [0, T], x \in R \end{aligned}$$

for appropriate constants C, D . The first condition is known as the linear growth condition, the last condition is the Lipschitz condition. Suppose that Z is a random variable independent of the σ -algebra \mathfrak{F}_∞ generated by B_t for $t \geq 0$ such that $E(|Z|^2) < \infty$. Then there is a unique stochastic process X , defined for $0 \leq t \leq T$, with time-continuous paths such that $X_0 = Z$ and such that the following equation is satisfied:

$$\begin{aligned} X(t, \omega) = X_0 + \int_0^t \varphi[s, X(s, \omega)] ds \\ + \int_0^t \psi[s, X(s, \omega)] dB_s \end{aligned}$$

The process X is called a strong solution of the above equation.

The above equation can be written in differential form as follows:

$$dX(t, \omega) = \varphi[t, X(t, \omega)] dt + \psi[t, X(t, \omega)] dB_t$$

The differential form does not have an independent meaning; a differential stochastic equation is just a short albeit widely used way to write the integral equation.

The key requirement of a strong solution is that the filtration \mathfrak{F}_t is given and that the functions φ, ψ are adapted to the filtration \mathfrak{F}_t . From the economic (or physics) point of view, this requirement translates the notion of causality. In simple terms, a strong solution is a functional of the driving Brownian motion and of the "inputs" φ, ψ . A strong solution at time t is determined only by the "history" up to time t of the inputs and of the random shocks embodied in the Brownian motion.

These conditions can be weakened. Suppose that we are given only the two functions $\varphi(t, x)$, $\psi(t, x)$ and that we must construct a process X_t , a Brownian motion B_t , and the relative filtration so that the above equation is satisfied. The equation still admits a unique solution with respect to the filtration generated by the Brownian motion B . It is, however, only a weak solution in the sense that, though there is no anticipation of information, it is not a functional of a given Brownian motion. (See, for example, Karatzas and Shreve [1991].) Weak and strong solutions do not necessarily coincide. However, any strong solution is also a weak solution with respect to the same filtration.

Note that the solution of a differential equation is a stochastic process. Initial conditions must therefore be specified as a random variable and not as a single value as for ordinary differential equations. In other words, there is an initial value for each state. It is possible to specify a single initial value as the initial condition of a stochastic differential equation. In this case the initial condition is a random variable where the probability mass is concentrated in a single point.

We omit the detailed proof of the theorem of uniqueness and existence. Uniqueness is proved using the Itô isometry and the Lipschitz condition. One assumes that there are two different solutions and then demonstrates that their difference must vanish. The proof of existence of a solution is similar to the proof of existence of solutions in the domain of ordinary equations. The solution is constructed inductively by a recursive relationship of the type

$$X^{(k+1)}(t, \omega) = \int_0^t \varphi[s, X^k(s, \omega)] ds + \int_0^t \psi[s, X^k(s, \omega)] dB_s$$

It can be shown that this recursive relationship produces a sequence of processes that converge to the unique solution.

GENERALIZATION TO SEVERAL DIMENSIONS

The concepts and formulas established so far for Itô (and Stratonovich) integrals and processes can be extended in a straightforward but often cumbersome way to multiple variables. The first step is to define a d -dimensional Brownian motion.

Given a probability space $(\Omega, \mathfrak{F}, P)$ equipped with a filtration $\{\mathfrak{F}_t\}$, a d -dimensional standard Brownian motion $B_t(\omega)$, is a stochastic process with the following properties:

- $B_t(\omega)$ is a d -dimensional process defined over the probability space $(\Omega, \mathfrak{F}, P)$ that takes values in R^d .
- $B_t(\omega)$ has continuous paths for $0 \leq t \leq \infty$.
- $B_0(\omega) = 0$.
- $B_t(\omega)$ is adapted to the filtration \mathfrak{F}_t .
- The increments $B_t(\omega) - B_s(\omega)$ are independent of the σ -algebra \mathfrak{F}_s and have a normal distribution with mean zero and covariance matrix $(t - s)I_d$, where I_d is the identity matrix.

The above conditions state that the standard Brownian motion is a stochastic process that starts at zero, has continuous paths, and has normally distributed increments whose variances grow linearly with time.

The next step is to extend the definition of the Itô integral in a multidimensional environment. This is again a straightforward but cumbersome extension of the one-dimensional case. Suppose that the following $r \times d$ -dimensional matrix is given:

$$\mathbf{v} = \begin{bmatrix} v_{11} & \cdot & v_{1d} \\ \cdot & \cdot & \cdot \\ v_{r1} & \cdot & v_{rd} \end{bmatrix}$$

where each entry $v_{ij} = v_{ij}(t, \omega)$ satisfies the following conditions:

1. v_{ij} are $\mathfrak{B}^d \times \mathfrak{S}$ measurable.
2. v_{ij} are \mathfrak{F}_t -adapted.
3. $P[\int_0^t (v_{ij})^2 ds < \infty \text{ for all } t \geq 0] = 1$.

Then, we define the multidimensional Itô integral

$$\int_0^t \mathbf{v} dB = \int_0^t \begin{bmatrix} v_{11} & \cdot & v_{1d} \\ \cdot & \cdot & \cdot \\ v_{r1} & \cdot & v_{rd} \end{bmatrix} \begin{bmatrix} dB_1 \\ \cdot \\ dB_d \end{bmatrix}$$

as the r -dimensional column vector whose components are the following sums of one-dimensional Itô integrals:

$$\sum_{i=1}^d \int_0^t v_{ij}(s, \omega) dB_j(s, \omega)$$

Note that the entries of the matrix are functions of time and state: They form a vector of stochastic processes. Given the previous definition of Itô integrals, we can now extend the definition of Itô processes to the multidimensional case. Suppose that the functions u and v satisfy the conditions established for the one-dimensional case. We can then form a multidimensional Itô process as the following vector of Itô processes:

$$\begin{aligned} dX_1 &= u_1 dt + v_{11} dB_1 + \dots + v_{1d} dB_d \\ \dots & \\ dX_{1r} &= u_r dt + v_{r1} dB_1 + \dots + v_{rd} dB_d \end{aligned}$$

or, in matrix notation

$$d\mathbf{X} = \mathbf{u}dt + \mathbf{v}dB$$

After defining the multidimensional Itô process, multidimensional stochastic equations are defined in differential form in matrix notation as follows:

$$\begin{aligned} d\mathbf{X}(t, \omega) &= \mathbf{u}[t, X_1(t, \omega), \dots, X_d(t, \omega)]dt \\ &+ \mathbf{v}[t, X_1(t, \omega), \dots, X_d(t, \omega)]dB \end{aligned}$$

Consider now the multidimensional map: $g(t, x) \equiv [g_1(t, x), \dots, g_d(t, x)]$, which maps the process X into another process $Y = g(t, X)$. It can be demonstrated that Y is a multidimensional Itô process whose components are defined

according to the following rules:

$$\begin{aligned} dY_k &= \frac{\partial g_k(t, X)}{\partial t} dt + \sum_i \frac{\partial g_k(t, X)}{\partial X_i} dX_i \\ &+ \frac{1}{2} \sum_{i,j} \frac{\partial^2 g_k(t, X)}{\partial X_i \partial X_j} dX_i dX_j \end{aligned}$$

$$dB_i dB_j = 1 \text{ if } i = j, \text{ 0 if } i \neq j, dB_i dt = dt dB_i = 0$$

SOLUTION OF STOCHASTIC DIFFERENTIAL EQUATIONS

It is possible to determine an explicit solution of stochastic differential equations in the linear case and in a number of other cases that can be reduced to linear equations through functional transformations. Let's first consider linear stochastic equations of the form:

$$\begin{aligned} dX_t &= [A(t)X_t + a(t)]dt + \sigma(t)dB_t, \quad 0 \leq t < \infty \\ X_0 &= \xi \end{aligned}$$

where B is an r -dimensional Brownian motion independent of the d -dimensional initial random vector ξ and the $(d \times d)$, $(d \times d)$, $(d \times r)$ matrices $A(t)$, $a(t)$, $\sigma(t)$ are nonrandom and time dependent.

The simplest example of a linear stochastic equation is the equation of an arithmetic Brownian motion with drift, written as follows:

$$\begin{aligned} dX_t &= \mu dt + \sigma dB_t, \quad 0 \leq t < \infty \\ X_0 &= \xi, \mu, \sigma \text{ constants} \end{aligned}$$

In linear equations of this type, the stochastic part enters only in an additive way through the terms $\sigma_{ij}(t)dB_t$. The functions $\sigma(t)$ are sometimes called the instantaneous variances and covariances of the process. In the example of the arithmetic Brownian motion, μ is called the drift of the process and σ the volatility of the process.

It is intuitive that the solution of this equation is given by the solution of the associated deterministic equation, that is, the ordinary differential equation obtained by removing the stochastic part, plus the cumulated random

disturbances. Let's first consider the associated deterministic differential equation

$$\frac{dx}{dt} = A(t)x + a(t), 0 \leq t < \infty$$

where $x(t)$ is a d -dimensional vector with initial conditions $x(0) = \xi$.

It can be demonstrated that this equation has an absolutely continuous solution in the domain $0 \leq t < \infty$. To find its solution, let's first consider the matrix differential equation

$$\frac{d\Phi}{dt} = A(t)\Phi, 0 \leq t < \infty$$

This matrix differential equation has an absolutely continuous solution in the domain $0 \leq t < \infty$. The matrix $\Phi(t)$ that solves this equation is called the fundamental solution of the equation. It can be demonstrated that $\Phi(t)$ is a nonsingular matrix for each t . Lastly, it can be demonstrated that the solution of the equation:

$$\frac{dx}{dt} = A(t)x + a(t), 0 \leq t < \infty$$

with initial condition $x(0) = \xi$, can be written in terms of the fundamental solution as follows:

$$x(t) = \Phi(t) \left[x(0) + \int_0^t \Phi^{-1}(s)a(s)ds \right], 0 \leq t < \infty$$

Let's now go back to the stochastic equation

$$dX_t = [A(t)X_t + a(t)]dt + \sigma(t)dB_t, 0 \leq t < \infty$$

$$X_0 = \xi$$

Using Itô's formula, it can be demonstrated that the above linear stochastic equation admits the following unique solution:

$$X(t) = \Phi(t) \left[\xi + \int_0^t \Phi^{-1}(s)a(s)ds + \int_0^t \Phi^{-1}(s)\sigma(s)dB_s \right], 0 \leq t < \infty$$

This effectively demonstrates that the solution of the linear stochastic equation is the solution of the associated deterministic equation plus

the cumulated stochastic term

$$\int_0^t \Phi^{-1}(s)\sigma(s)dB_s$$

To illustrate this, below we now specialize the above solutions in the case of arithmetic Brownian motion, Ornstein-Uhlenbeck processes, and geometric Brownian motion.

The Arithmetic Brownian Motion

The *arithmetic Brownian motion* in one dimension is defined by the following equation:

$$dX_t = \mu dt + \sigma dB_t$$

In this case, $A(t) = 0$, $a(t) = \mu$, $\sigma(t) = \sigma$ and the solution becomes

$$X = \mu t + \sigma B$$

The Ornstein-Uhlenbeck Process

The *Ornstein-Uhlenbeck process* in one dimension is a mean-reverting process defined by the following equation:

$$dX_t = -\alpha X_t dt + \sigma dB_t$$

It is a mean-reverting process because the drift is pulled back to zero by a term proportional to the process itself. In this case, $A(t) = -\alpha$, $a(t) = 0$, $\sigma(t) = \sigma$ and the solution becomes

$$X_t = X_0 + e^{-\alpha t} + \sigma \int_0^t e^{-\alpha(t-s)} dB_s$$

The Geometric Brownian Motion

The *geometric Brownian motion* in one dimension is defined by the following equation:

$$dX = \mu X dt + \sigma X dB$$

This equation can be easily reduced to the previous linear case by the transformation:

$$Y = \log X$$

Let's apply Itô's formula

$$dY_t = \left(\frac{\partial g}{\partial t} + \frac{\partial g}{\partial x} a + \frac{1}{2} \frac{\partial^2 g}{\partial x^2} b^2 \right) dt + \frac{\partial g}{\partial x} b dB_t$$

where

$$g(t, x) = \log x, \quad \frac{\partial g}{\partial t} = 0, \quad \frac{\partial g}{\partial x} = \frac{1}{x}, \quad \frac{\partial^2 g}{\partial x^2} = -\frac{1}{x^2}$$

We can then verify that the logarithm of the geometric Brownian motion becomes an arithmetic Brownian motion with drift

$$\mu' = \mu - \frac{1}{2}\sigma^2$$

The geometric Brownian motion evolves as a lognormal process:

$$X_t = x_0 \exp \left\{ \left(\mu - \frac{1}{2}\sigma^2 \right) t + \sigma B_t \right\}$$

KEY POINTS

- Stochastic differential equations give meaning to ordinary differential equations where some terms are subject to random perturbation.
- Following Itô and Stratonovich, stochastic differential equations are defined through their integral equivalent: The differential notation is just a shorthand.
- Itô processes are the sum of a time integral plus an Itô integral.
- Itô processes are closed with respect to smooth maps: A smooth function of an Itô process is another Itô process defined through the Itô formula.
- Stochastic differential equations are equations established in terms of Itô processes.
- Linear equations can be solved explicitly as the sum of the solution of the associated deterministic equation plus a stochastic cumulative term.

NOTE

1. It is possible to define a "generalized white noise process" in the domain of "tempered distributions." See Oksendal (1992).

REFERENCES

- Itô, K. (1951). On stochastic differential equations. *Memoirs, American Mathematical Society* 4: 1–51.
- Karatzas, I., and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*. New York: Springer.
- Oksendal, B. (1992). *Stochastic Differential Equations*, 3rd Edition. Berlin: Springer.

Stochastic Processes in Continuous Time

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair-Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

YOUNG SHIN KIM, PhD

Research Assistant Professor, School of Economics and Business Engineering, University of Karlsruhe and KIT

MICHELE LEONARDO BIANCHI, PhD

Research Analyst, Specialized Intermediaries Supervision Department, Bank of Italy

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: The dynamic of a financial asset's returns and prices can be expressed using a deterministic process if there is no uncertainty about its future behavior, or with a stochastic process in the more likely case when the value is uncertain. Stochastic processes in continuous time are the most used tool to explain the dynamics of a financial asset's returns and prices. They are the building blocks with which to construct financial models for portfolio optimization, derivatives pricing, and risk management. Continuous-time processes allow for more elegant theoretical modeling compared to discrete-time models and many results proven in probability theory can be applied to obtain a simple evaluation method.

In 1900, the father of modern option pricing theory, Louis Bachelier, proposed using *Brownian motion* for modeling stock market prices. There are several reasons why Brownian motion is a popular process. First, Brownian motion is the milestone of the theory of stochastic processes.

However, more realistic general processes that are better suited for financial modeling such as Lévy, additive, or self-similar processes have been developed only since the mid 1990s (see Samorodnitsky and Taqqu (1994), Sato (1999), and Embrechts and Maejima (2002)). Most of

Dr. Bianchi acknowledges that the views expressed in this entry are his own and do not necessarily reflect those of the Bank of Italy.

the practical problems in mathematical finance can be solved by taking into consideration these new processes. For example, the concept of stochastic integral with respect to Brownian motion was introduced in 1933 and only in the 1990s has the general theory of stochastic integration with respect to semi-martingale appeared. From a practical point of view, the second reason for the popularity of Brownian motion is that the normal distribution allows one to solve real-world pricing problems such as option prices as estimations and simulations in a few seconds, and most of the problems have a closed-form solution which can be easily used. See Øksendal (2003) or Karatzas and Shreve (1991) for a complete theoretical treatment of financial applications of continuous-time stochastic processes driven by Brownian motion.

The two basic classes of continuous-time stochastic processes are Brownian motion and the *Poisson process*. The name of the former is due to the botanist Robert Brown who in 1827 described the movement of pollen suspended in water. The theory of Brownian motion was founded by the work of Norbert Wiener who was the first to prove its existence and, as a result, Brownian motion is sometimes also referred to as a Wiener process. The Poisson process generated by the Poisson distribution is the building block of pure *jump processes*. Both processes are fundamentally different with respect to their path properties and they belong to the larger class of *Lévy processes* (for more details about Lévy processes see Sato [1999]). Schoutens (2003), Cont and Tankov (2004), and Rachev et al. (2011) provide details of Lévy processes with applications to option pricing.

Infinitely divisible distributions, including α -stable and tempered stable distributions, can be considered to define continuous-time stochastic processes. In order to model the behavior of a financial asset's returns and prices, one can consider (1) a Brownian motion, (2) a process defined as the sum of a Brownian motion and a Poisson process, or (3) a pure jump Lévy process.

In this entry, we will discuss continuous-time stochastic processes. We will first consider processes consisting of jumps and then we will discuss continuous processes without jumps. We then turn our focus to processes having random time instead of physical time. Finally, we will discuss a general process that contains all of these processes.

SOME PRELIMINARIES

Before we continue with the discussion and the construction of processes, we will briefly define terms that will be used in this entry.

- A *stochastic process* $X = (X_t)_{t \geq 0}$ is a family of \mathbb{R} -valued random variables X_t with parameter $t \geq 0$, defined on the sample space Ω . For every outcome $\omega \in \Omega$, the function $t \mapsto X_t(\omega)$ is called a sample path of the process X .
- Let X be a stochastic process. Given $0 < t_1 < t_2 < \dots < t_n$, if the random variables $X_{t_1} - X_0, X_{t_2} - X_{t_1}, \dots, X_{t_n} - X_{t_{n-1}}$ are independent, we say that X has independent increments. Moreover, for $t \geq 0$, if the distribution of $X_{t+h} - X_t$ does not depend on $t \geq 0$, we say that X has stationary increments. Loosely speaking, one could say that the distribution of the future changes does not depend on past realizations.
- A process X is said to be non-decreasing, if $Y_t \geq 0$ almost surely (a.s.) for $t \geq 0$, and $Y_t \geq Y_s$ a.s. for $0 \leq s \leq t$. Usually, a non-decreasing process is called a *subordinator*. A process X is said to be non-increasing if $Y_t \leq 0$ a.s. for $t \geq 0$, and $Y_t \leq Y_s$ a.s. for $0 \leq s \leq t$.
- We say that a process X has finite (infinite) variation if its sample paths are of finite (infinite) variation, that is, the variation

$$V(X(\omega))_t = \lim_{n \rightarrow \infty} \sum_{k=1}^n |X_{tk/n}(\omega) - X_{t(k-1)/n}(\omega)|, \quad \forall t > 0$$

is finite (infinite) for almost every $\omega \in \Omega$.

- The characteristic function of the stochastic process $X = (X_t)_{t \geq 0}$ on \mathbb{R} is defined as the

function $\phi : \mathbb{R} \rightarrow \mathbb{C}$

$$\phi_{X_t}(u) = E[e^{iuX_t}]$$

POISSON PROCESS

Consider a process $N = (N_t)_{t \geq 0}$ derived by a Poisson distribution with parameter λ as follows:

1. $N_0 = 0$
2. N has independent increments and stationary increments.
3. For any real numbers $t \geq 0$ and $h \geq 0$, the variable $(N_{t+h} - N_t)$ is a Poisson distributed random variable with parameter λh , that is,

$$\mathbb{P}(N_{t+h} - N_t = n) = e^{-\lambda h} \frac{(\lambda h)^n}{n!}, \quad n = 0, 1, 2, \dots$$

The process N is referred to as the *Poisson process* with intensity λ .

If $(\tau_j)_{j \in \mathbb{N}}$ are independent exponential random variables with parameter λ and the random variable N_t is given by

$$N_t = \inf \left\{ n \geq 1 : \sum_{j=1}^n \tau_j > t \right\}$$

then it can be proven that the process $(N_t)_{t \geq 0}$ is the Poisson process with intensity λ .

The Poisson process is a fundamental example of a stochastic process with discontinuous trajectories, and a building block for constructing more complex jump processes.

Compounded Poisson Process

The process $X = (X_t)_{t \geq 0}$ is referred to as a compounded Poisson process if X is defined by

$$X_t = \sum_{k=1}^{N_t} Y_k$$

where

- Y_1, Y_2, \dots are independent and identically distributed (IID) random variables, and f is the probability density function of Y_1 .
- $(N_t)_{t \geq 0}$ is a Poisson process with intensity λ .
- N_t and Y_k are independent for all $t \geq 0$ and $k = 1, 2, \dots$.

The characteristic function of X_t is equal to

$$\phi_{X_t}(u) = \exp \left(\lambda t \int_{-\infty}^{\infty} (e^{iux} - 1) f(x) dx \right)$$

Moreover, if f is given by the probability density function of the normal distribution, then X is referred to as a jump diffusion process.

PURE JUMP PROCESS

Consider a process $X^x = (X_t^x)_{t \geq 0}$ for a given real number x such that

$$X_t^x = x N_t^{\lambda(x)}$$

where $(N_t^{\lambda(x)})_{t \geq 0}$ is the Poisson process with intensity $\lambda(x)$. The number x represents the jump size and the intensity $\lambda(x)$ is the expected number of jumps with size x in the unit time.

Let $S = \{x_j \in \mathbb{R} : x_j \neq 0, j = 1, 2, \dots\}$ be a discrete subset of jump sizes, $\lambda(x_j) > 0$ for all $x_j \in S$, and $Y = (Y_t)_{t \geq 0}$ be a process defined by

$$Y_t = \gamma t + \sum_{j=1}^{\infty} X_t^{x_j}$$

If S consists of positive real numbers and $\gamma > 0$, then the process Y is non-decreasing. Conversely, if S consists of negative real numbers and $\gamma < 0$, Y is non-increasing.

Since the characteristic function of X_t^x is equal to

$$\phi_{X_t^x}(u) = \exp(\lambda(x)t(e^{-iux} - 1))$$

the characteristic function of Y_t is obtained by

$$\phi_{Y_t} = \exp \left(i\gamma ut + t \sum_{j=1}^{\infty} \lambda(x_j)(e^{iux_j} - 1) \right)$$

For the process Y , the function ν defined by $\nu(A) = \sum_{x_j \in A} \lambda(x_j)$ represents the expected number of jumps with size $x \in A$ in the unit time interval, where A is a subset of S . For example, the expected number of jumps whose sizes are in $\{x_1, x_2, \dots, x_n\}$ is equal to $\nu(\{x_1, x_2, \dots, x_n\}) = \sum_{j=1}^n \lambda(x_j)$.

Now, we extend the set of jump size S to the real number set \mathbb{R} . Then the expected number of jumps is defined by a map ν from a subset of \mathbb{R} to a positive number. The map ν is a jump

measure, that is, the expected number of jumps whose sizes are in a real interval $[a, b]$ is represented by $\nu([a, b])$. Using ν , we can obtain an extended process Y such that the characteristic function of Y_t is given by

$$\phi_{Y_t} = \exp\left(i\gamma ut + t \int_{-\infty}^{\infty} (e^{iux} - 1)\nu(dx)\right) \quad (1)$$

where $\gamma \in \mathbb{R}$. Jump sizes of process Y can be defined continuously. In this case, the measure ν is referred to as a Lévy measure, that is, a Borel measure on \mathbb{R} satisfying $\nu(0) = 0$ and

$$\int_{-\infty}^{\infty} \min\{1, x^2\}\nu(dx) < \infty$$

The class of jump processes satisfying (1) cannot contain infinite variation processes. To include infinite variation processes in the class of jump processes we will be using, we need a more general definition. Consider a process $Z = (Z_t)_{t \geq 0}$ such that the characteristic function of Z_t is given by

$$\phi_{Z_t} = \exp\left(i\gamma ut + t \int_{-\infty}^{\infty} (e^{iux} - 1 - iux1_{|x| \leq 1})\nu(dx)\right) \quad (2)$$

The process Z is referred to as the pure jump process. If

$$\int_{-1}^1 |x|\nu(dx) = \infty$$

then the characteristic function (1) is not defined, but the function (2) is well defined. The details can be found in Sato (1999) and Cont and Tankov (2004). The path behavior of the pure jump process is determined by the Lévy measure ν and real number γ .

- $\gamma > 0$ and $\nu(A) = 0$ for all $A \subset (-\infty, 0)$, then Z is non-decreasing.
- $\gamma < 0$ and $\nu(A) = 0$ for all $A \subset (0, \infty)$, then Z is non-increasing.
- If $\nu(\mathbb{R}) < \infty$ (i.e., the expected number of jumps on the unit time is finite), then we say that Z has a finite activity.
- If $\nu(\mathbb{R}) = \infty$ (i.e., the expected number of jumps on the unit time is infinite), then we say that Z has an infinite activity.

- If $\int_{-1}^1 |x|\nu(dx) < \infty$, the process Z has finite variation.
- If $\int_{-1}^1 |x|\nu(dx) = \infty$, the process Z has infinite variation.

The building block of the pure jump process Z is the Poisson process. Hence, Z has the following properties:

- $Z_0 = 0$.
- Z has independent and stationary increments; that is, the random variable $(Z_t - Z_s)$ is independent of the random variable $(Z_v - Z_u)$ for all real number s, t, u , and v with $0 \leq s < t < u < v$.
- $Z_{s+t} - Z_s \stackrel{d}{=} Z_t$ for $s \geq 0$ and $t > 0$. Moreover, we have

$$\log \phi_{z_t}(u) = t \log \phi_{z_1}(u) \quad (3)$$

where $\phi_{z_t}(u)$ is the characteristic function of Z_t for $t > 0$.

If $t = 1$, then we obtain the purely non-Gaussian infinitely divisible random variable. In fact, there is a one-to-one correspondence between a purely non-Gaussian infinitely divisible random variable and a pure jump process.

Gamma Process

Consider the gamma distribution with parameter (c, λ) . Since the gamma distribution is a purely non-Gaussian infinitely divisible distribution, we can define a pure jump process $G = (G_t)_{t \geq 0}$ such that $G_1 \sim \text{Gamma}(c, \lambda)$. By equation (3), the characteristic function ϕ_{G_t} of G_t is given by

$$\phi_{G_t} = \left(\frac{\lambda}{\lambda - iu}\right)^{ct} \quad (4)$$

In this case, the process G is referred to as the *gamma process* with parameter (λ, c) . The sample path of the gamma process is non-decreasing, since the gamma distribution is supported only on the positive real line. When we take $c = 1$ of the gamma process, the process is referred to as an *exponential process*.

Inverse Gaussian Process

Consider the inverse Gaussian distribution with parameter (c, λ) . Since the inverse Gaussian distribution is also a purely non-Gaussian infinitely divisible distribution, we can define a pure jump process $X = (X_t)_{t \geq 0}$ such that $X_1 \sim IG(c, \lambda)$. By equation (3), the characteristic function ϕ_{X_t} of X_t is given by

$$\phi_{X_t} = \exp\left(-ct(\sqrt{\lambda^2 - 2iu} - \lambda)\right) \tag{5}$$

In this case, the process X is referred to as the *inverse Gaussian (IG) process* with parameter (c, λ) . The sample path of the gamma process is nondecreasing, since the inverse Gaussian distribution is supported only on the positive real line.

Variance Gamma Process

The variance gamma process is an infinitely divisible distribution. Thus we can define pure jump processes $X = (X_t)_{t \geq 0}$ such that $X_1 \sim VG(C, \lambda_+, \lambda_-)$. By equation (3), the characteristic function ϕ_{X_t} of X_t is given by

$$\phi_{X_t} = \left(\frac{\lambda_+ \lambda_-}{(\lambda_+ - iu)(\lambda_- + iu)}\right)^{Ct} \tag{6}$$

In this case, the process X is referred to as the *variance gamma (VG) process* with parameter $(C, \lambda_+, \lambda_-)$.

α -Stable Process

The pure jump process $X = (X_t)_{t \geq 0}$ is referred to as the *α -stable process* with parameters $(\alpha, \sigma, \beta, \mu)$ if X_1 is an α -stable random variable, that is, $X_1 \sim S_\alpha(\sigma, \beta, \mu)$. By equation (3), the characteristic function ϕ_{X_t} of X_t is given by

$$\phi_{X_t}(u) = \begin{cases} \left(\exp(i\mu ut - t|\sigma u|^\alpha \times \left(1 - i\beta(\text{sign } u) \tan \frac{\pi\alpha}{2}\right))\right), & \alpha \neq 1 \\ \left(\exp(i\mu ut - t\sigma|u| \times \left(1 + i\beta \frac{2}{\pi}(\text{sign } u) \ln|u|\right))\right), & \alpha = 1 \end{cases}$$

Recall the Lévy measure of the α -stable process can be written as

$$\nu(dx) = \left(\frac{C_+}{x^{1+\alpha}} 1_{x>0} + \frac{C_-}{|x|^{1+\alpha}} 1_{x<0}\right) dx$$

where C_+ and C_- are positive constants. Then we can prove that

$$\nu(\mathbb{R}) = \int_{-\infty}^{\infty} \nu(dx) = \infty$$

and hence the α -stable process is an infinite activity process. On the other hand, since we have

$$\int_{-1}^1 |x| \nu(dx) = \begin{cases} \frac{C_+ + C_-}{1-\alpha}, & \alpha < 1 \\ \infty, & \alpha \geq 1 \end{cases}$$

we conclude that the α -stable process has finite variation if $\alpha < 1$ and the infinite variation if $\alpha \geq 1$.

Tempered Stable Process

The pure jump process $X = (X_t)_{t \geq 0}$ is referred to as the *tempered stable process* if X_1 is the tempered stable random variable.

- The process X is referred to as the classical tempered stable (CTS) process with parameters $(\alpha, C, \lambda_+, \lambda_-, m)$ if $X_1 \sim \text{CTS}(\alpha, C, \lambda_+, \lambda_-, m)$. The process X is referred to as the *standard CTS process* with parameters $(\alpha, \lambda_+, \lambda_-)$ if $X_1 \sim \text{stdCTS}(\alpha, \lambda_+, \lambda_-)$.
- The process X is referred to as the *generalized tempered stable (GTS) process* with parameters $(\alpha_+, \alpha_-, C_+, C_-, \lambda_+, \lambda_-, m)$ if $X_1 \sim \text{GTS}(\alpha_+, \alpha_-, C_+, C_-, \lambda_+, \lambda_-, m)$. The process X is referred to as the *standard GTS process* with parameters $(\alpha_+, \alpha_-, \lambda_+, \lambda_-, p)$ if $X_1 \sim \text{stdGTS}(\alpha_+, \alpha_-, \lambda_+, \lambda_-, p)$.
- The process X is referred to as the *modified tempered stable (MTS) process* with parameters $(\alpha, C, \lambda_+, \lambda_-, m)$ if $X_1 \sim \text{MTS}(\alpha, C, \lambda_+, \lambda_-, m)$. The process X is referred to as the *standard MTS process* with parameters $(\alpha, \lambda_+, \lambda_-)$ if $X_1 \sim \text{stdMTS}(\alpha, \lambda_+, \lambda_-)$.
- The process X is referred to as the *normal tempered stable (NTS) process* with parameters $(\alpha, C, \lambda, \beta, m)$ if $X_1 \sim \text{NTS}(\alpha, C, \lambda, \beta, m)$. The process X is referred to as the *standard*

NTS process with parameters (α, λ, β) if $X_1 \sim \text{stdNTS}(\alpha, \lambda, \beta)$.

- Moreover, the process X is referred to as the *normal inverse Gaussian (NIG) process* with parameters (c, λ, β, m) if $X_1 \sim \text{NIG}(c, \lambda, \beta, m)$. The process X is referred to as the *standard NIG process* with parameters (λ, β) if $X_1 \sim \text{stdNIG}(\lambda, \beta)$.
- The process X is referred to as the *Kim-Rachev tempered stable (KRTS) process* with parameters $(\alpha, k_+, k_-, r_+, r_-, p_+, p_-, m)$ if $X_1 \sim \text{KRTS}(\alpha, k_+, k_-, r_+, r_-, p_+, p_-, m)$. The process X is referred to as the *standard KRTS process* with parameters $(\alpha, r_+, r_-, p_+, p_-)$ if $X_1 \sim \text{stdKRTS}(\alpha, r_+, r_-, p_+, p_-)$.
- The process X is referred to as the *rapidly decreasing tempered stable (RDTS) process* with parameters $(\alpha, C, \lambda_+, \lambda_-, m)$ if $X_1 \sim \text{RDTS}(\alpha, C,$

$\lambda_+, \lambda_-, m)$. The process X is referred to as the *standard RDTS process* with parameters $(\alpha, \lambda_+, \lambda_-)$ if $X_1 \sim \text{stdRDTS}(\alpha, \lambda_+, \lambda_-)$.

The characteristic function ϕ_{X_t} of X_t is obtained by equation (3). For example, if X is the CTS process with parameters $(\alpha, C, \lambda_+, \lambda_-, m)$, then

$$\begin{aligned} \phi_{X_t}(u) &= \exp(t \log(\phi_{\text{CTS}}(u; \alpha, C, \lambda_+, \lambda_-, m))) \\ &= \exp(iumt - iutC\Gamma(1 - \alpha)(\lambda_+^{\alpha-1} - \lambda_-^{\alpha-1}) \\ &\quad + tC\Gamma(-\alpha)((\lambda_+ - iu)^\alpha - \lambda_+^\alpha \\ &\quad + (\lambda_- + iu)^\alpha - \lambda_-^\alpha)) \end{aligned}$$

Characteristic exponents of tempered stable processes are presented in Table 1.

Let $\nu(dx)$ be the Lévy measure of the tempered stable process. Then we can prove that $\nu(\mathbb{R}) = \infty$, $\int_{-1}^1 |x|\nu(dx) < \infty$ if $\alpha < 1$, and $\int_{-1}^1 |x|\nu(dx) = \infty$ if $\alpha \geq 1$. Consequently, the

Table 1 Characteristic Exponents of Tempered Stable Processes

Process	$\psi_{X_t}(u) = \log \phi_{X_t}(u)$
CTS	$iumt - iutC\Gamma(1 - \alpha)(\lambda_+^{\alpha-1} - \lambda_-^{\alpha-1}) + tC\Gamma(-\alpha)((\lambda_+ - iu)^\alpha - \lambda_+^\alpha + (\lambda_- + iu)^\alpha - \lambda_-^\alpha)$
GTS	$iumt - iut\Gamma(1 - \alpha)(C_+\lambda_+^{\alpha-1} - C_-\lambda_-^{\alpha-1}) + tC_+\Gamma(-\alpha_+)((\lambda_+ - iu)^{\alpha_+} - \lambda_+^{\alpha_+}) + tC_-\Gamma(-\alpha_-)((\lambda_- + iu)^{\alpha_-} - \lambda_-^{\alpha_-})$
MTS	$iumt + tC(G_R(u; \alpha, \lambda_+) + G_R(u; \alpha, \lambda_-)) + iutC(G_I(u; \alpha, \lambda_+) - G_I(u; \alpha, \lambda_-))$ where $G_R(x; \alpha, \lambda) = 2^{-\frac{\alpha+3}{2}} \sqrt{\pi} \Gamma\left(-\frac{\alpha}{2}\right) ((\lambda^2 + x^2)^{\frac{\alpha}{2}} - \lambda^\alpha)$ and $G_I(x; \alpha, \lambda) = 2^{-\frac{\alpha+1}{2}} \Gamma\left(\frac{1-\alpha}{2}\right) \lambda^{\alpha-1} \left[{}_2F_1\left(1, \frac{1-\alpha}{2}; \frac{3}{2}; -\frac{x^2}{\lambda^2}\right) - 1 \right]$
NTS	$iumt - iut2^{-\frac{\alpha-1}{2}} C \sqrt{\pi} \Gamma\left(1 - \frac{\alpha}{2}\right) \beta (\lambda^2 - \beta^2)^{\frac{\alpha}{2}-1} + t2^{-\frac{\alpha+1}{2}} C \sqrt{\pi} \Gamma\left(-\frac{\alpha}{2}\right) ((\lambda^2 - (\beta + iu)^2)^{\frac{\alpha}{2}} - (\lambda^2 - \beta^2)^{\frac{\alpha}{2}})$
NIG	$iumt - \frac{iutc\beta}{\sqrt{\lambda^2 - \beta^2}} - tc \left(\sqrt{\lambda^2 - (\beta + iu)^2} - \sqrt{\lambda^2 - \beta^2} \right)$
KRTS	$iumt - iut\Gamma(1 - \alpha) \left(\frac{k_+r_+}{p_+ + 1} - \frac{k_-r_-}{p_- + 1} \right) + tk_+H(iu; \alpha, r_+, p_+) + tk_-H(-iu; \alpha, r_-, p_-)$ where $H(x; \alpha, r, p) = \frac{\Gamma(-\alpha)}{p} ({}_2F_1(p, -\alpha; 1 + p; rx) - 1)$
RDTS	$iumt + tC(G(iu; \alpha, \lambda_+) + G(-iu; \alpha, \lambda_-))$ where $G(x; \alpha, \lambda) = 2^{-\frac{\alpha}{2}-1} \lambda^\alpha \Gamma\left(-\frac{\alpha}{2}\right) \left(M\left(-\frac{\alpha}{2}, \frac{1}{2}; \frac{x^2}{2\lambda^2}\right) - 1 \right) + 2^{-\frac{\alpha}{2}-\frac{1}{2}} \lambda^{\alpha-1} x \Gamma\left(\frac{1-\alpha}{2}\right) \left(M\left(\frac{1-\alpha}{2}, \frac{3}{2}; \frac{x^2}{2\lambda^2}\right) - 1 \right)$

tempered stable process has infinite activity, and has finite variation if $\alpha < 1$ and infinite variation if $\alpha \geq 1$, as explained in Carr et al. (2002), Kim (2005), and Kim et al. (2008 and 2010).

BROWNIAN MOTION

In this section, we will discuss *Brownian motion* by means of an example. We begin with a short summary of the most important and defining properties of a standard Brownian motion $W = (W_t)_{t \geq 0}$

1. $W_0 = 0$
2. W has independent increments and stationary increments.
3. For any real numbers $t \geq 0$ and $h \geq 0$, the variable $(W_{t+h} - W_t)$ is a normally distributed random variable with mean zero and variance h .
4. The paths of $W = (W_t)_{t \geq 0}$ are continuous.

Every process fulfilling the above four properties is referred to as the standard Brownian motion. From the second and third conditions it can be deduced that Brownian motion W_t at time t (which equals the increment from time 0 to time t) is normally distributed with mean zero and variance t .

The paths of Brownian motion are highly irregular and nowhere differentiable. In order to draw a true path, one would have to calculate the value of the process for every real number, which is clearly not feasible. Due to its characteristic path property, it is impossible to draw a real path of Brownian motion. The process can only be evaluated for a discrete set of points. Figure 1 illustrates possible paths of Brownian motion. Strictly speaking, the plotted paths are only discrete approximations to the true paths.

From the above definition of the process, it may not be clear how one can envision a Brownian motion or how one could construct it. Therefore, we will present a constructive method demonstrating how one can generate a Brownian motion as the limit of very simple processes. We restrict the presentation to

the unit interval (i.e., we assume $0 \leq t \leq 1$) but the generalization to the abstract case should be obvious. The procedure is iterative, which means that on the k th step of the iteration we define a process $(X_t^{(k)})_{0 \leq t \leq 1}$, which will serve as an approximation for a standard Brownian motion.

Let random variables I_1, I_2, I_3, \dots be IID with

$$I_j = \begin{cases} 1 & \text{with probability } p = 0.5 \\ -1 & \text{with probability } 1 - p = 0.5 \end{cases}, \\ j = 1, 2, \dots$$

Define $X_t^{(k)} = \frac{1}{\sqrt{k}} \sum_{j=1}^n I_j$ where $t = n/k$ and $n = 0, 1, \dots, k$. If the value t is on the interval $(\frac{n}{k}, \frac{n+1}{k})$, then we take a value obtained by a linear interpolation as

$$X_t^{(k)} = (kt - n)X_{n/k}^{(k)} + (kt - n - 1)X_{(n+1)/k}^{(k)}$$

By doing so, we get a stochastic process with continuous paths.

Let's start with $k = 1$. Then we have

$$X_0^{(1)} = 0, \\ X_1^{(1)} = \begin{cases} 1 & \text{with probability } p = 0.5 \\ -1 & \text{with probability } 1 - p = 0.5 \end{cases}$$

At any time t the random variable $X_1^{(1)}$ can take only two possible values, namely $-t$ and t . At any time, the process has zero mean and the variance at time $t = 1$ equals

$$V(X_1^{(1)}) = 1^2 \cdot 0.5 + (-1)^2 \cdot 0.5 = 1$$

That is not so bad for the first step, but obviously the distribution of $X_1^{(1)}$ is far from being normal.

What we do in the next step, $k = 2$, is allow for two different values until time $t = \frac{1}{2}$ and three different values for $\frac{1}{2} \leq t \leq 1$. We do so by defining:

$$X_0^{(2)} = 0, \\ X_{0.5}^{(2)} = \begin{cases} \frac{1}{\sqrt{2}} & \text{with probability } p = 0.5 \\ -\frac{1}{\sqrt{2}} & \text{with probability } 1 - p = 0.5 \end{cases} \\ X_1^{(2)} = \begin{cases} \sqrt{2} & \text{with probability } p^2 = 0.25 \\ 0 & \text{with probability } p(1 - p) = 0.5 \\ -\sqrt{2} & \text{with probability } (1 - p)^2 = 0.25 \end{cases}$$

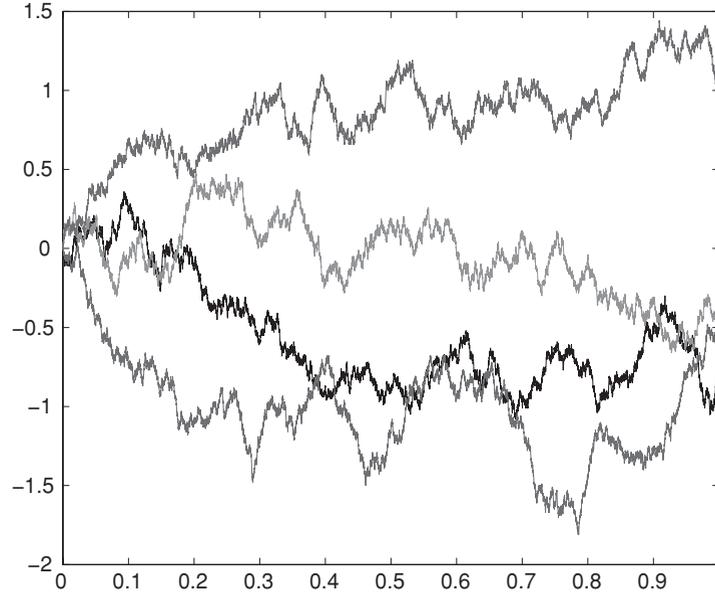


Figure 1 Possible Paths of a Standard Brownian Motion (Every Path Consists of 10,000 Equally Spaced Observations)

The process $X_t^{(2)}$ now has four possible paths. The mean of $X_t^{(2)}$ is zero and the variance of $X_t^{(2)}$ equals

$$V\left(X_{0.5}^{(2)}\right) = \left(\frac{1}{\sqrt{2}}\right)^2 \cdot 0.5 + \left(-\frac{1}{\sqrt{2}}\right)^2 \cdot 0.5 = 0.5$$

$$V\left(X_1^{(2)}\right) = \sqrt{2}^2 \cdot 0.25 + (-\sqrt{2})^2 \cdot 0.25 = 1$$

but still the distribution of $X_t^{(2)}$ is far from being normal.

By iterating the stated procedure, the probability of $X_t^{(k)}$ is given by

$$\mathbf{P}\left(X_t^{(k)} = \frac{n-2m}{\sqrt{k}}\right) = \binom{n}{m} \left(\frac{1}{2}\right)^n$$

if $m \in \{0, 1, 2, \dots, n\}$, $t = n/k$, $n \in \{0, 1, 2, \dots, k\}$. The mean and variance can be obtained as follows:

$$E\left[X_t^{(k)}\right] = \frac{1}{\sqrt{k}} \sum_{j=1}^n E[I_j] = 0$$

$$V\left(X_t^{(k)}\right) = \frac{1}{k} \sum_{j=1}^n E[I_j^2] = \frac{n}{k}$$

where $t = n/k$, $n = 1, 2, \dots, k$. Since $X_{n/k}^{(k)}$ is defined by the sum of IID random variables, it has

- *Independent increments:* $X_{n_1/k}^{(k)}$ and $X_{n_2/k}^{(k)} - X_{n_1/k}^{(k)}$ are independent, for all $n_1, n_2 \in \{0, 1, \dots, k\}$ with $n_1 < n_2$.
- *Stationary increments:* $X_{n_2/k}^{(k)} - X_{n_1/k}^{(k)} \stackrel{d}{=} X_{(n_2-n_1)/k}^{(k)}$ for all $n_1, n_2 \in \{0, 1, \dots, k\}$ with $n_1 < n_2$.

Moreover, the distribution of $X_t^{(k)}$ will approach the normal distribution due to the central limit theorem. Consequently, we have found all the defining properties of a Brownian motion in this simple approximating process, that is, the process $(X_t^{(k)})_{0 \leq t \leq 1}$ converges in distribution to the standard Brownian motion $(W_t)_{0 \leq t \leq 1}$.

In the context of financial applications, there are two main variants of the standard Brownian motion which have to be mentioned: the arithmetic and the geometric Brownian motion. Both are obtained as a function of the standard Brownian motion.

Arithmetic Brownian Motion

Given a Brownian motion $(W_t)_{t \geq 0}$ and two real constants μ and σ , the arithmetic Brownian motion $(X_t)_{t \geq 0}$ is obtained as:

$$X_t = \mu t + \sigma W_t$$

The process $(X_t)_{t \geq 0}$ consists of the sum of a purely deterministic linear trend function μt and a rescaled Brownian motion σW_t . The latter has the property that at time t , σW_t is normally distributed with mean 0 and variance $\sigma^2 t$. The paths will therefore randomly jitter around the deterministic trend with a variance proportional to the point in time t under consideration. The arithmetic Brownian motion is a simple but popular model for financial asset returns.

Geometric Brownian Motion

Given a Brownian motion $(W_t)_{t \geq 0}$, two real constants μ and σ , and a starting value $S_0 > 0$, the geometric Brownian motion $(S_t)_{t \geq 0}$ is obtained as:

$$S_t = S_0 e^{\mu t + \sigma W_t}$$

The process $(S_t)_{t \geq 0}$ is just the exponential of an arithmetic Brownian motion multiplied by a factor. Therefore $\log(S_t/S_0)$ is normally distributed and

$$E[S_t/S_0] = e^{\mu t + \frac{1}{2}\sigma^2 t}$$

TIME-CHANGED BROWNIAN MOTION

If a pure jump process $T = (T_t)_{t \geq 0}$ is non-decreasing, that is, $T_t \geq 0$ a.s. for $t > 0$, and $T_t \geq T_s$ a.s. for $s \leq t$, then the process T is referred to as the subordinator or intrinsic time process. Intuitively, it can be thought of as the cumulative trading volume process for a financial asset which measures the cumulative volume of all the transitions up to physical time t (Rachev and Mitnik, 2000).

The Poisson, gamma, and inverse Gaussian processes are non-decreasing, and hence they are subordinators. Moreover, for the case where

$0 < \alpha < 1$, the support of the α -stable distribution $S_\alpha(\alpha, 1, 0)$ is the positive real line. Hence, the α -stable process with parameters $(\frac{\alpha}{2}, \sigma, 1, 0)$ and $0 < \alpha < 2$ is a subordinator and referred to as α -stable subordinator. If we can consider an additional assumption that $E[T_t] = t$, this would mean that the expected intrinsic time is the same as physical time.

If we take an arithmetic Brownian motion and change the physical time to a subordinator, then we obtain the *time-changed Brownian motion*. That is, take an arithmetic Brownian motion with drift μ and volatility σ as follows:

$$\mu t + \sigma W_t$$

and consider a subordinator $T = (T_t)_{t \geq 0}$ independent to the standard Brownian motion $(W_t)_{t \geq 0}$. Then substituting $t = T_t$ in the arithmetic Brownian motion, we have a new process $X = (X_t)_{t \geq 0}$ with

$$X_t = \mu T_t + \sigma W_{T_t}$$

which is the time-changed Brownian motion.

If T_t is fixed, then the conditional probability of X_t with a fixed variable T_t follows a normal distribution, that is

$$\begin{aligned} \mathbf{P}(X_t < y | T_t) &= \mathbf{P}(\mu T_t + \sigma W_{T_t} < y | T_t) \\ &= \frac{1}{\sqrt{2\pi\sigma^2 T_t}} \int_{-\infty}^y e^{-\frac{(x-\mu T_t)^2}{2\sigma^2 T_t}} dx \end{aligned}$$

Using properties of the conditional probability and independence between W_t and T_t , the distribution function F_{X_t} and the probability density function f_{X_t} of X_t are obtained by

$$\begin{aligned} F_{X_t}(y) &= \mathbf{P}(X_t < y) \\ &= \int_{-\infty}^y \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2 s}} e^{-\frac{(y-\mu s)^2}{2\sigma^2 s}} f_{T_t}(s) ds dx \end{aligned}$$

and

$$f_{X_t}(y) = \frac{d}{dy} F_{X_t}(y) = \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2 s}} e^{-\frac{(y-\mu s)^2}{2\sigma^2 s}} f_{T_t}(s) ds$$

respectively, where f_{T_t} is the probability density function of T_t . Moreover, we can derive the characteristic function ϕ_{X_t} as follows:

$$\phi_{X_t}(u) = \phi_{T_t} \left(\mu u + \frac{i u^2 \sigma^2}{2} \right) \tag{7}$$

where ϕ_{T_t} is the characteristic function of T_t . Using the time-changed Brownian motion, we can define various processes.

The time-changed Brownian motion construction is well known from the theory of stochastic processes and is referred to as the Skorokhod embedding problem. Theoretically, every Lévy process can be defined as the time-changed Brownian motion. More in general, a process can be embedded in a Brownian motion if and only if it is a local semimartingale, as proved by Monroe (1978).

Although the representation via Brownian subordination is a nice property, we do not know a general constructive method to find the process T_t such that $X_t = \mu T_t + \sigma W_{T_t}$. This means that given a semimartingale X_t , the time process T_t is not always of known form. Thus, this approach can be applied only for some particular Lévy processes.

Variance Gamma Process

By considering the gamma process as the subordinator of the Brownian motion, we obtain the VG process. That is, the VG process is defined by $X = (X_t)_{t \geq 0}$ with

$$X_t = \mu G_t + \sigma W_{G_t}$$

where $G = (G_t)_{t \geq 0}$ is the gamma process with parameter (c, λ) . In order to reduce the number of parameters, we consider the assumption $E[G_t] = t$. Since we have $E[G_t] = \frac{ct}{\lambda}$, the assumption is satisfied if $c = \lambda$. Then the characteristic function of X_t is equal to

$$\phi_{X_t}(u) = \left(\frac{c}{c - i\mu u + \frac{u^2 \sigma^2}{2}} \right)^{ct} = \left(\frac{\frac{2c}{\sigma^2}}{\frac{2c}{\sigma^2} - \frac{2\mu}{\sigma^2} ui + u^2} \right)^{ct} \quad (8)$$

by (7) and the characteristic function of G_t given in (4) with $c = \lambda$. Inserting into (8) the parametrization

$$\begin{aligned} \lambda_- - \lambda_+ &= \frac{2\mu}{\sigma^2} \\ \lambda_+ \lambda_- &= \frac{2c}{\sigma^2} \\ C &= c \end{aligned}$$

we obtain the form given by (6).

Normal Inverse Gaussian Process

By considering the inverse Gaussian process as the subordinator of the Brownian motion, we obtain the NIG process.

Define a process $X = (X_t)_{t \geq 0}$ with

$$X_t = \mu T_t + \sigma W_{T_t}$$

where $T = (T_t)_{t \geq 0}$ is the inverse Gaussian process with parameter (c, λ) , satisfying $E[T_t] = t$. The condition $E[T_t] = t$ holds if $c = \lambda$. Then the characteristic function of X_t is equal to

$$\begin{aligned} \phi_{X_t}(u) &= \exp \left(-kt(\sqrt{k^2 - 2i\mu u + 2\sigma^2 u^2} - k) \right) \\ &= \exp \left(-\sqrt{2}k\sigma t \left(\sqrt{\frac{k^2}{2\sigma^2} - \frac{\mu}{\sigma^2} iu + u^2} - \sqrt{\frac{k^2}{2\sigma^2}} \right) \right) \quad (9) \end{aligned}$$

by (7) and the characteristic function of T_t given in (5) with $k := c = \lambda$. Inserting into (9) the parametrization

$$\begin{aligned} \lambda^2 - \beta^2 &= \frac{k^2}{2\sigma^2} \\ \beta &= \frac{\mu}{2\sigma^2} \\ c &= \sqrt{2}k\sigma \end{aligned}$$

we obtain the NIG process with parameter $(c, \lambda, \beta, \frac{c\beta}{\sqrt{\lambda^2 - \beta^2}})$.

Normal Tempered Stable Process

Assume Lévy measure ν is equal to

$$\nu(dx) = \frac{ce^{-\theta x}}{x^{\alpha/2+1}} 1_{x>0} dx \quad (10)$$

where $\alpha \in (0, 2)$, $c > 0$, and $\theta > 0$, and consider the pure jump process $T = (T_t)_{t \geq 0}$ defined by ν and γ , where

$$\gamma = \int_0^1 x\nu(dx)$$

Since $\nu(A) = 0$ for all $A \subset (-\infty, 0)$ and $\mu \geq 0$, the process T is a nondecreasing process. Hence it is a subordinator and referred to as the tempered stable subordinator with parameters

(α, c, θ) . Using equation (2), the characteristic function ϕ_{T_t} of T_t is equal to

$$\phi_{T_t}(u) = \exp \left(tc \int_0^\infty (e^{iux} - 1) \frac{e^{-\theta x}}{x^{\alpha/2+1}} dx \right)$$

Solving the integration in the last equation, we can obtain the following formula,

$$\phi_{T_t}(u) = \exp \left(tc \Gamma \left(-\frac{\alpha}{2} \right) ((\theta - iu)^{\frac{\alpha}{2}} - \theta^{\frac{\alpha}{2}}) \right) \quad (11)$$

The mean of T_t is computed by the first cumulant, that is,

$$E[T_t] = \frac{1}{i} \frac{\partial}{\partial u} \log \phi_{T_t}(u)|_{u=0} = tc \Gamma \left(1 - \frac{\alpha}{2} \right) \theta^{\frac{\alpha}{2}-1}$$

Hence, the condition $E[T_t] = t$ holds if $c = (\Gamma(1 - \frac{\alpha}{2}) \theta^{\frac{\alpha}{2}-1})^{-1}$.

By considering the tempered stable subordinator as the subordinator of the Brownian motion, we obtain the NTS process. That is, define a process $X = (X_t)_{t \geq 0}$ with

$$X_t = \mu T_t + \sigma W_{T_t}$$

where $T = (T_t)_{t \geq 0}$ is the tempered stable subordinator with parameter $(\alpha, (\Gamma(1 - \frac{\alpha}{2}) \theta^{\frac{\alpha}{2}-1})^{-1}, \theta)$. The characteristic function of X_t is equal to

$$\begin{aligned} \phi_{X_t}(u) &= \exp \left(\frac{t \Gamma(-\frac{\alpha}{2})}{\Gamma(1 - \frac{\alpha}{2}) \theta^{\frac{\alpha}{2}-1}} \left(\left(\theta - i \left(\mu u + \frac{i \sigma^2 u^2}{2} \right) \right)^{\frac{\alpha}{2}} - \theta^{\frac{\alpha}{2}} \right) \right) \\ &= \exp \left(\frac{-2t}{\alpha \theta^{\frac{\alpha}{2}-1}} \left(\left(\theta - i \left(\mu u + \frac{i \sigma^2 u^2}{2} \right) \right)^{\frac{\alpha}{2}} - \theta^{\frac{\alpha}{2}} \right) \right) \quad (12) \end{aligned}$$

by (7) and (11) with $c = (\Gamma(1 - \frac{\alpha}{2}) \theta^{\frac{\alpha}{2}-1})^{-1}$. The last equation can be changed to the following expression:

$$\begin{aligned} \phi_{X_t}(u) &= \exp \left(\frac{t \Gamma(-\frac{\alpha}{2}) \left(\frac{\sigma^2}{2} \right)^{\frac{\alpha}{2}}}{\Gamma(1 - \frac{\alpha}{2}) \theta^{\frac{\alpha}{2}-1}} \left(\left(\frac{2\theta}{\sigma^2} + \left(\frac{\mu}{\sigma^2} \right)^2 \right. \right. \right. \\ &\quad \left. \left. \left. - \left(\frac{\mu}{\sigma^2} + iu \right)^2 \right)^{\frac{\alpha}{2}} - \left(\frac{2\theta}{\sigma^2} \right)^{\frac{\alpha}{2}} \right) \right) \quad (13) \end{aligned}$$

Inserting into (13) the parametrization

$$\begin{aligned} \lambda &= \sqrt{\frac{2\theta}{\sigma^2} + \left(\frac{\mu}{\sigma^2} \right)^2} \\ \beta &= \frac{\mu}{\sigma^2} \\ C &= \frac{\sqrt{2\sigma^\alpha}}{\sqrt{\pi} \Gamma \left(1 - \frac{\alpha}{2} \right) \theta^{\frac{\alpha}{2}-1}} \end{aligned}$$

we obtain the NTS process with parameter $(\alpha, C, \lambda, \beta, m)$ where

$$m = -2^{-\frac{\alpha+1}{2}} C \sqrt{\pi} \Gamma \left(\frac{\alpha}{2} \right) \beta (\lambda^2 - \beta^2)^{\frac{\alpha}{2}-1}$$

LÉVY PROCESS

A stochastic process $X = (X_t)_{t \geq 0}$ is called a *Lévy process* if the following five conditions are satisfied :

1. $X_0 = 0$ a.s.
2. X has independent increments.
3. X has stationary increment.
4. X is stochastically continuous that is, $\forall t \geq 0$ and $a > 0$,

$$\lim_{s \rightarrow t} \mathbf{P}[|X_s - X_t| > a] = 0$$

5. X is right continuous and has left limits (cadlag).

The standard Brownian motion, arithmetic Brownian motions, and pure jump processes are all Lévy processes. Moreover, a Lévy process can be decomposed by a Brownian motion and a pure jump process $(Z_t)_{t \geq 0}$ independent to the Brownian motion, that is

$$X_t = \sigma W_t + Z_t$$

Hence we obtain the characteristic function of X_t as follows:

$$\begin{aligned} \phi_{X_t}(u) &= \phi_{\sigma W_t}(u) \phi_{Z_t}(u) \\ &= \exp \left(-\frac{t}{2} \sigma^2 u^2 \right) \exp \left(i \gamma u t + t \int_{-\infty}^\infty (e^{iux} - 1 - iux \mathbf{1}_{|x| \leq 1}) \nu(dx) \right) \\ &= \exp \left(i \gamma u t - \frac{t}{2} \sigma^2 u^2 + t \int_{-\infty}^\infty (e^{iux} - 1 - iux \mathbf{1}_{|x| \leq 1}) \nu(dx) \right) \end{aligned}$$

where $\phi_{\sigma W_t}(u)$ is the characteristic function of $N(0, \sigma^2 t)$, and $\phi_{Z_t}(u)$ given by (2). Therefore,

if $X = (X_t)_{t \geq 0}$ is a Lévy process, then for any $t \geq 0$, X_t is an infinitely divisible random variable. Conversely, if Y is an infinitely divisible random variable, then there exists uniquely a Lévy process $(X_t)_{t \geq 0}$ such that $X_1 = Y$, as proved by Sato (1999, p. 38).

KEY POINTS

- Continuous-time stochastic processes are the building block of financial modeling and they are usually used to explain the uncertain behavior of financial assets. Some results of probability theory can be usefully applied to financial derivatives pricing and risk management.
- Given any infinitely divisible random variable X_1 , it is possible to define a stochastic process with independent and stationary increments such that for all $t > s$, the increment $X_t - X_s$ has characteristic function $\exp((t - s) \log \phi_{X_1}(u))$. These processes are known as Lévy processes.
- Brownian motion and Poisson processes are Lévy processes. All Lévy processes can be constructed by changing the deterministic time t of the Brownian motion W_t with a stochastic time T_t . This construction is called Brownian subordination and the increasing process T_t is a subordinator.
- There are two main variants of the standard Brownian motion used in financial applications: the arithmetic and the geometric Brownian motion.
- The Poisson process is a fundamental example of a stochastic process with discontinuous trajectories, and a building block for constructing more complex jump processes.
- Pure jump processes include also the gamma process, the inverse Gaussian process, the variance gamma process, the α -stable process, and the tempered stable process.

REFERENCES

- Carr, P., Geman, H., Madan, D., and Yor, M. (2002). The fine structure of asset returns: An empirical investigation. *Journal of Business* 75, 2: 305–332.
- Cont, R., and Tankov, P. (2004). *Financial Modelling with Jump Processes*. Boca Raton: CRC Press.
- Embrechts, P., and Maejima, M. (2002). *Selfsimilar Processes*. New Jersey: Princeton University Press.
- Karatzas, I., and Shreve, S. (1991). *Brownian Motion and Stochastic Calculus*. New York: Springer.
- Kim, Y. S., Rachev, S. T., Bianchi, M. L., and Fabozzi, F. J. (2008). A new tempered stable distribution and its application to finance. In G. Bol, S. T. Rachev, and R. Wuerth (Eds.), *Risk Assessment: Decisions in Banking and Finance*. Heidelberg: Springer Verlag, 77–109.
- Kim, Y. S., Rachev, S. T., Bianchi, M. L., and Fabozzi, F. (2010). Tempered stable and tempered infinitely divisible GARCH models. *Journal of Banking and Finance* 34, 9: 2096–2109.
- Kim, Y. S. (2005). *The Modified Tempered Stable Processes with Application to Finance*. Ph.D thesis, Sogang University.
- Monroe, I. (1978). Processes that can be embedded in Brownian motion. *The Annals of Probability* 6, 1: 42–56.
- Øksendal, B. (2003). *Stochastic Differential Equations: An Introduction with Applications, 5th ed.* New York: Springer.
- Rachev, S. T., Kim, Y. S., Bianchi, M. L., and Fabozzi, F. J. (2011). *Financial Models with Lévy Processes and Volatility Clustering*. Hoboken, NJ: Wiley.
- Rachev, S. T., and Mittnik, S. (2000). *Stable Parettian Models in Finance*. Chichester: John Wiley & Sons.
- Samorodnitsky, G., and Taqqu, M. (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. New York: CRC Press.
- Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge: Cambridge University Press.
- Schoutens, W. (2003). *Lévy Processes in Finance: Pricing Financial Derivatives*. Hoboken, NJ: Wiley.

Conditional Expectation and Change of Measure

SVETLOZAR T. RACHEV, PhD, Dr Sci

Frey Family Foundation Chair-Professor, Department of Applied Mathematics and Statistics, Stony Brook University, and Chief Scientist, FinAnalytica

YOUNG SHIN KIM, PhD

Research Assistant Professor, School of Economics and Business Engineering, Karlsruhe Institute of Technology, Germany

MICHELE LEONARDO BIANCHI, PhD

Research Analyst, Specialized Intermediaries Supervision Department, Bank of Italy

FRANK J. FABOZZI, PhD, CFA, CPA, CFA

Professor of Finance, EDHEC Business School

Abstract: The current price of an option is obtained by the conditional expectation of the payoff function under the risk-neutral measure. The risk-neutral measure is the measure equivalent to the real market measure under which the discounted price process of the underlying stock becomes a martingale. In the Black-Scholes model, the risk-neutral measure can be obtained by the Girsanov theorem. The Esscher transform has been used to find the risk-neutral measure for the continuous Lévy process models. The general theory of the Esscher transform is applied to find the risk-neutral measure under tempered stable Lévy process models.

In this entry, we present some issues in stochastic processes. We begin by defining events of a probability space mathematically, and then discuss the concept of conditional expectation. We then explain two important notions for stochastic processes: martingale properties and Markov properties. The former relates to the fair price in a market and

the latter describes the efficiency of a market. Finally, “change of measures” for processes are discussed. Change of measures for tempered stable processes are important for determining no-arbitrage pricing for assets. Further details about no-arbitrage pricing with the change of measure is discussed in Rachev et al. (2011).

Dr. Bianchi acknowledges that the views expressed in this entry are his own and do not necessarily reflect those of the Bank of Italy.

EVENTS, σ -FIELDS, AND FILTRATION

A set of possible outcomes in a given sample space Ω is called an event. An event is mathematically defined as a subset of Ω . If we have one event A , then the set of outcomes that are not included in A is also an event. For example, if we consider an event that the return of the stock of Disney tomorrow will be positive, then the set of outcomes that Disney's return tomorrow will be negative is also an event. Moreover, if we have two events A and B , then a set of outcomes included in both A and B is also an event. For instance, consider two events, the first event being that Disney's stock return tomorrow will be positive, and the other event that IBM's stock return tomorrow will be positive. Then a set of outcomes that both stock returns will be positive tomorrow is an event.

The class of events is described mathematically by the σ -field. The σ -field, denoted by \mathcal{F} , is the class of the subsets of Ω that satisfy the following properties:

Property 1. $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$.

Property 2. If $A \in \mathcal{F}$, then $A^c = \{x \in \Omega \mid x \notin A\} \in \mathcal{F}$.

Property 3. If $A_1, A_2, A_3, \dots \in \mathcal{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Let \mathcal{G} denote a class of subsets contained in Ω . Then the smallest σ -field containing \mathcal{G} is referred to as the σ -field generated by \mathcal{G} , and is denoted by $\sigma(\mathcal{G})$. For a given random variable X , consider the class $\mathcal{G} = \{A \subseteq \Omega : A = X^{-1}(I), \text{ for all open interval } I \text{ in } \mathbb{R}\}$, where X^{-1} is the inverse image of X . Then the σ -field generated by \mathcal{G} is referred to as the σ -field generated by X , and denoted by $\sigma(X)$. If there is a σ -field \mathcal{F} such that $\sigma(X) \subseteq \mathcal{F}$, then we say that X is \mathcal{F} -measurable.

The probability \mathbf{P} is a map from a given σ -field \mathcal{F} to the unit interval $[0, 1]$. If $A \subseteq N \in \mathcal{F}$ and $P(N) = 0$, then the set A is referred to as a null set with respect to $(\Omega, \mathcal{F}, \mathbf{P})$. Let \mathcal{N} be the class of all null sets with respect to $(\Omega, \mathcal{F}, \mathbf{P})$. The space $(\Omega, \tilde{\mathcal{F}}, \tilde{\mathbf{P}})$ is referred to as a comple-

tion of $(\Omega, \mathcal{F}, \mathbf{P})$ if $\tilde{\mathcal{F}} = \sigma(\mathcal{F} \cup \mathcal{N})$ and $\tilde{\mathbf{P}}(A \cup N) = \mathbf{P}(A)$ for all $A \in \mathcal{F}$ and $N \in \mathcal{N}$. All probability spaces in this entry are assumed to be completions of spaces, that is, all null sets are contained in given σ -fields, and probabilities are defined on completed σ -fields.

Let $(\mathcal{F}_t)_{t \geq 0}$ be a sequence of σ -field with continuous index $t \geq 0$ (or discrete index $t = 0, 1, 2, \dots$). If $\mathcal{F}_s \subseteq \mathcal{F}_t$ for all $0 \leq s \leq t$, then $(\mathcal{F}_t)_{t \geq 0}$ is referred to as a filtration. \mathcal{F}_t can be interpreted as the "information" available to all market agents at time t . The filtration describes increasing information for time t .

Consider a stochastic process $X = (X_t)_{t \geq 0}$. If X_t is \mathcal{F}_t -measurable for all $t \geq 0$, then X is referred to as a $(\mathcal{F}_t)_{t \geq 0}$ -adapted process. If X_t is \mathcal{F}_{t-1} -measurable for all discrete index $t = 0, 1, 2, \dots$, then X is referred to as a $(\mathcal{F}_t)_{t \geq 0}$ -predictable process.

For a given process $X = (X_t)_{t \geq 0}$, we can generate a filtration $(\mathcal{F}_t)_{t \geq 0}$ by

$$\mathcal{F}_t = \sigma(X_s; 0 \leq s \leq t)$$

where $\sigma(X_s; 0 \leq s \leq t)$ is the smallest σ -field containing all $\sigma(X_s)$ with $0 \leq s \leq t$. Then the process X is $(\mathcal{F}_t)_{t \geq 0}$ -adapted and this filtration is referred to as a filtration generated by X .

CONDITIONAL EXPECTATION

The *conditional expectation* is a value of the expectation of a random variable under some restricted events. Let g be a Borel function, X be a random variable on a space (Ω, \mathbf{P}) with $E[g(X)] < \infty$, and A be an event. The conditional expectation $E[g(X)|A]$ is defined by

$$E[g(X)|A] = \frac{E[g(X)1_A]}{P(A)}$$

where

$$1_A(\omega) = \begin{cases} 0 & \text{if } \omega \notin A \\ 1 & \text{if } \omega \in A \end{cases}$$

Consider a Borel function g , a stochastic process $X = (X_t)_{t \geq 0}$ adapted to filtration $(\mathcal{F}_t)_{t \geq 0}$. We can define the conditional expectation on \mathcal{F}_t as

a random variable. That is, the conditional expectation $E[g(X_T)|\mathcal{F}_t]$ for $t \leq T$ is a random variable, such that

$$E[g(X_T)|\mathcal{F}_t](\omega) = E[g(X_T)|A_\omega], \quad \omega \in \Omega$$

where A_ω is the smallest event in \mathcal{F}_t with $\omega \in A_\omega$, or $A_\omega = \cap_{\omega \in B_\omega \in \mathcal{F}_t} B_\omega$. Moreover, if g and h are Borel functions, and $0 \leq s \leq t \leq T \leq T^*$, then we have the following properties:

- $E[g(X_t)|\mathcal{F}_0] = E[g(X_t)]$ where $\mathcal{F}_0 = \{\emptyset, \Omega\}$.
- $E[E[g(X_T)|\mathcal{F}_t]|\mathcal{F}_s] = E[g(X_T)|\mathcal{F}_s]$.
- $E[g(X_t)h(X_T)|\mathcal{F}_t] = g(X_t)E[h(X_T)|\mathcal{F}_t]$.
- $E[ag(X_T) + bh(X_{T^*})|\mathcal{F}_t] = aE[g(X_T)|\mathcal{F}_t] + bE[h(X_{T^*})|\mathcal{F}_t]$, for $a, b \in \mathbb{R}$.

We write $E[g(X_T)|X_t]$ instead of $E[g(X_T)|\mathcal{F}_t]$ when $\mathcal{F}_t = \sigma(X_t)$. Hence we have:

- $E[E[g(X_T)|X_t]|X_s] = E[g(X_T)|X_s]$.
- $E[g(X_t)h(X_T)|X_t] = g(X_t)E[h(X_T)|X_t]$.
- $E[ag(X_T) + bh(X_{T^*})|X_t] = aE[g(X_T)|X_t] + bE[h(X_{T^*})|X_t]$, for $a, b \in \mathbb{R}$.

If a (\mathcal{F}_t) -adapted process $X = (X_t)_{t \geq 0}$ satisfies the condition

$$E[g(X_T)|\mathcal{F}_t] = E[g(X_T)|X_t]$$

for all $0 \leq t \leq T$ and Borel function g , then the process X is referred to as a Markov process.

In finance, a Markov process is used to explain the efficient market hypothesis. Suppose X is a price process of an asset, and consider a forward contract on the asset with maturity T . The σ -field \mathcal{F}_t contains all market information until time t . Hence, $F_t = E[X_T|\mathcal{F}_t]$ is the expected price of the forward contract based on the information up to t . If the market is efficient, all information until t is impounded into the current price X_t . Hence, the expected price of the forward contract can be obtained by $F_t = E[X_T|X_t]$.

If a (\mathcal{F}_t) -adapted process $X = (X_t)_{t \geq 0}$ satisfies the condition

$$X_t = E[X_T|\mathcal{F}_t]$$

for all $0 \leq t \leq T$, then the process X is referred to as a martingale process. The process $X = (X_t)_{t \geq 0}$

with $X_t = \sigma W_t$ is a martingale process, where $\sigma > 0$ and $(W_t)_{t \geq 0}$ is the standard Brownian motion. Since X_t is \mathcal{F}_t -measurable, we have

$$\begin{aligned} E[X_T|\mathcal{F}_t] &= E[X_T - X_t + X_t|\mathcal{F}_t] \\ &= E[X_T - X_t|\mathcal{F}_t] + X_t \end{aligned}$$

Since X has stationary and independent increments,

$$\begin{aligned} E[X_T - X_t|\mathcal{F}_t] &= E[X_T - X_t] = E[X_{T-t}] \\ &= E[\sigma W_{T-t}] = 0 \end{aligned}$$

Hence the process X is a martingale.

In finance, a martingale process describes the fair price or no-arbitrage price for an asset. For example, consider one share of a stock and a forward contract that required delivery of one share of that stock to the forward contract holder at the maturity date. Suppose $(S_t)_{t \geq 0}$ is a stock price process and $(F_t)_{0 \leq t \leq T}$ is the price process for the forward contract with maturity T . The forward price at time $t < T$ is given by the conditional expectation of S_T based on the information until time t , that is, $F_t = E[S_T|\mathcal{F}_t]$. Moreover, we can see that $F_t = S_t$ for all t with $0 \leq t \leq T$ by the following argument. Suppose $F_t > S_t$. Then we obtain the difference $F_t - S_t > 0$ at time t by purchasing one share of the stock at price S_t and selling the forward contract at price F_t . We invest the proceeds in a money market account with interest rate r . At time T , by delivering the stock to the holder of the forward contract, we will then have $e^{r(T-t)}(F_t - S_t)$, which is an arbitrage profit. If $F_t > S_t$, then another arbitrage opportunity can be found by selling (i.e., shorting) one share of the stock and purchasing the forward contract. Therefore, to eliminate arbitrage opportunities, F_t should be equal to S_t ; that is, the stock price process should be a martingale.

CHANGE OF MEASURES

In this section, we will present *change of measure* for random variables and Lévy processes. Change of measure is an important method

to determine no-arbitrage prices of assets and derivatives.

Equivalent Probability Measure

Consider two probability measures \mathbf{P} and \mathbf{Q} on a sample space Ω and σ -field \mathcal{F} . If they satisfy the condition

$$\mathbf{Q}(A) = 0 \Rightarrow \mathbf{P}(A) = 0,$$

then we say that \mathbf{P} is absolutely continuous with respect to \mathbf{Q} , and denote $\mathbf{P} \ll \mathbf{Q}$. Moreover, if $\mathbf{P} \ll \mathbf{Q}$ and $\mathbf{Q} \ll \mathbf{P}$, that is,

$$\mathbf{Q}(A) = 0 \Leftrightarrow \mathbf{P}(A) = 0,$$

then we say that \mathbf{P} and \mathbf{Q} are equivalent.

If $\mathbf{Q} \ll \mathbf{P}$, then there exists a positive random variable ξ with $\int_{\Omega} \xi d\mathbf{P} = 1$ and

$$\mathbf{Q}(A) = \int_A \xi d\mathbf{P} \quad (1)$$

for any $A \in \mathcal{F}$. In this case, ξ is referred to as the Radon-Nikodym derivative, and denotes

$$\xi = \frac{d\mathbf{Q}}{d\mathbf{P}}$$

Conversely, if there is a positive random variable ξ with $\int_{\Omega} \xi d\mathbf{P} = 1$ and \mathbf{Q} is defined by equation (1), then \mathbf{Q} is also a probability measure and $\mathbf{Q} \ll \mathbf{P}$.

Let X be a random variable on a probability measure \mathbf{P} , and $f(x) = \frac{\partial}{\partial x} \mathbf{P}(X \leq x)$ be the probability density function (p.d.f.) of X . Suppose \mathbf{Q} is a probability measure and the probability density function of X on \mathbf{Q} is given by $g(x) = \frac{\partial}{\partial x} \mathbf{Q}(X \leq x)$. If \mathbf{P} and \mathbf{Q} are equivalent, then the Radon-Nikodym derivative is equal to

$$\frac{d\mathbf{Q}}{d\mathbf{P}} = \frac{g(X)}{f(X)}$$

For example, $X \sim N(0, 1)$ is normally distributed on \mathbf{P} . If we take the Radon-Nikodym derivative by

$$\xi_1 = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}} / \sqrt{2\pi\sigma^2}}{e^{-\frac{x^2}{2}} / \sqrt{2\pi}},$$

then the measure \mathbf{Q}_1 defined by $\mathbf{Q}_1(A) = \int_A \xi_1 d\mathbf{P}$ for $A \in \mathcal{F}$ is equivalent to \mathbf{P} and $X \sim N(\mu, \sigma^2)$ on the measure \mathbf{Q}_1 . On the other hand, if we take the Radon-Nikodym derivative by

$$\xi_2 = \frac{h(X)}{e^{-\frac{x^2}{2}} / \sqrt{2\pi}}$$

where

$$h(x) = \frac{\sigma}{\pi((x-\mu)^2 + \sigma^2)}$$

which is the probability density function of the Cauchy distribution, then the measure \mathbf{Q}_2 , defined by $\mathbf{Q}_2(A) = \int_A \xi_2 d\mathbf{P}$ for $A \in \mathcal{F}$ is equivalent to \mathbf{P} and $X \sim S_1(\sigma, 0, \mu)$ on the measure \mathbf{Q}_2 .

Consider a finite discrete process $(X_t)_{t \in \{1, 2, \dots, T\}}$ of independent and identically distributed (IID) real random variables on both probability measures \mathbf{P} and \mathbf{Q} , where T is a positive integer. By the independent property of the process on \mathbf{P} , we have

$$\begin{aligned} & \mathbf{P}[X_1 \in \mathbb{R}, \dots, X_{t-1} \in \mathbb{R}, X_t < x, \\ & \quad X_{t+1} \in \mathbb{R}, \dots, X_T \in \mathbb{R}] \\ &= \mathbf{P}[X_1 \in \mathbb{R}] \cdots \mathbf{P}[X_{t-1} \in \mathbb{R}] \cdot \mathbf{P}[X_t < x] \\ & \quad \cdot \mathbf{P}[X_{t+1} \in \mathbb{R}] \cdots \mathbf{P}[X_T \in \mathbb{R}] \\ &= \mathbf{P}[X_t < x] \end{aligned}$$

By the same argument, we have

$$\begin{aligned} & \mathbf{Q}[X_1 \in \mathbb{R}, \dots, X_{t-1} \in \mathbb{R}, X_t < x, \\ & \quad X_{t+1} \in \mathbb{R}, \dots, X_T \in \mathbb{R}] = \mathbf{Q}[X_t < x] \end{aligned}$$

Since X_t 's are identically distributed on \mathbf{P} and \mathbf{Q} , respectively, we have $\mathbf{P}[X_t < x] = \mathbf{P}[X_s < x]$ and $\mathbf{Q}[X_t < x] = \mathbf{Q}[X_s < x]$ for all $t, s \in \{1, 2, \dots, T\}$. Suppose that for all $t \in \{1, 2, \dots, T\}$ the probability density functions of X_t are given by $f(x)$ and $g(x)$ on probability measures \mathbf{P} and \mathbf{Q} , respectively. That is

$$f(x) = \frac{\partial}{\partial x} \mathbf{P}[X_t < x]$$

and

$$g(x) = \frac{\partial}{\partial x} \mathbf{Q}[X_t < x]$$

If the domain of the function f is the same as the domain of the function g , then \mathbf{P} and \mathbf{Q} are

equivalent and the Radon-Nikodym derivative is equal to

$$\frac{d\mathbf{Q}}{d\mathbf{P}} = \frac{g(X_1)g(X_2)\cdots g(X_T)}{f(X_1)f(X_2)\cdots f(X_T)}$$

However, that method cannot be used for either continuous-time processes or infinite-discrete processes. In the next section, we discuss the change of measure for continuous-time processes using Girsanov's theorem and the extended Girsanov's theorem.

Change of Measure for Continuous-Time Processes

A continuous-time process is a function from the sample space to the set of appropriate functions. Hence, the change of measure for processes is more complex than the change of measure for a random variable.

Brownian motion is a function from the sample space to the set of continuous functions. For Brownian motion, we can find an equivalent measure using the following theorem, which is referred to as *Girsanov's theorem*.¹

Theorem 1. Let $W = (W_t)_{t \geq 0}$ be a standard Brownian motion under measure \mathbf{P} and $(\mathcal{F}_t)_{t \geq 0}$ be a filtration generated by W . Consider a process $(\xi_t)_{t \geq 0}$ defined by

$$\xi_t = e^{-\theta W_t - \frac{\theta^2}{2}t}$$

Then the probability measure \mathbf{Q} given by

$$\mathbf{Q}(A) |_{\mathcal{F}_t} = \int_A \xi_t d\mathbf{P}, \quad A \in \mathcal{F}_t$$

is equivalent to $\mathbf{P}|_{\mathcal{F}_t}$ for all $t \geq 0$, and the process $\tilde{W} = (\tilde{W}_t)_{t \geq 0}$ with $\tilde{W}_t = \theta t + W_t$ is a standard Brownian motion under the measure \mathbf{Q} .

Girsanov's theorem shows how stochastic processes change under the change of measure. For example, let a process $X = (X_t)_{t \geq 0}$ be an arithmetic Brownian motion under measure \mathbf{P} such that

$$X_t = \mu t + \sigma W_t$$

where $(W_t)_{t \geq 0}$ is the standard Brownian motion. The process X is not martingale on the measure \mathbf{P} , but we can obtain a measure where X is a martingale by Girsanov's theorem. Indeed, we define a measure \mathbf{Q} equivalent to \mathbf{P} such that

$$\mathbf{Q}(A) |_{\mathcal{F}_t} = \int_A e^{-\frac{\mu W_t}{\sigma} - \frac{\mu^2}{2\sigma^2}t} d\mathbf{P}, \quad A \in \mathcal{F}_t$$

Then the process X becomes $X_t = \sigma \tilde{W}_t$ with $\tilde{W}_t = \frac{\mu t}{\sigma} + W_t$ and the process $(\tilde{W}_t)_{t \geq 0}$ is a standard Brownian motion on the measure \mathbf{Q} . Therefore, the process X is a martingale on the measure \mathbf{Q} .

A Lévy process is a function from the sample space to the set of right continuous functions with left limits at any point of the domain.² Girsanov's theorem can be extended for Lévy processes by the following theorem:

Theorem 2. Suppose a process $X = (X_t)_{t \geq 0}$ is a Lévy process with Lévy triplets (σ^2, ν, γ) under measure \mathbf{P} . If there is a real number θ satisfying $\int_{|x| \geq 1} e^{\theta x} \nu(dx) < \infty$, then we can find the equivalent measure \mathbf{Q} whose Radon-Nikodym derivative is given by

$$\frac{d\mathbf{Q}}{d\mathbf{P}} |_{\mathcal{F}_t} = \xi_t = \frac{e^{\theta X_t}}{E_{\mathbf{P}}[e^{\theta X_t}]} = e^{\theta X_t - l(\theta)t}$$

where $l(\theta) = \log E_{\mathbf{P}}[e^{\theta X_1}]$. That is,

$$\mathbf{Q}(A) |_{\mathcal{F}_t} = \int_A \xi_t d\mathbf{P}, \quad A \in \mathcal{F}_t$$

is equivalent to $\mathbf{P}|_{\mathcal{F}_t}$ for all $t \geq 0$. Moreover, the process X is a Lévy process with Lévy triplets $(\sigma^2, \tilde{\nu}, \tilde{\gamma})$ under the measure \mathbf{Q} , where $\tilde{\nu}(dx) = e^{\theta x} \nu(dx)$ and $\tilde{\gamma} = \gamma + \int_{|x| \leq 1} x(e^{\theta x} - 1) \nu(dx)$.

The change of measure using Theorem 2 is referred to as the *Esscher transform*. The most general theorem of change of measure for Lévy processes is given by the following theorem (see Sato, 1999):

Theorem 3. Suppose a process $X = (X_t)_{t \geq 0}$ has Lévy triplets (σ^2, ν, γ) and $(\tilde{\sigma}^2, \tilde{\nu}, \tilde{\gamma})$ under measures \mathbf{P} and \mathbf{Q} , respectively.

1. In the case where $\sigma^2 \neq 0$ and $\tilde{\sigma}^2 \neq 0$, $\mathbf{P}|_{\mathcal{F}_t}$ and $\mathbf{Q}|_{\mathcal{F}_t}$ are equivalent for all $t > 0$ if and only if

the Lévy triplets satisfy

$$\sigma^2 = \tilde{\sigma}^2 > 0 \tag{2}$$

and

$$\int_{-\infty}^{\infty} (e^{\psi(x)/2} - 1)^2 \nu(dx) < \infty \tag{3}$$

where $\psi(x) = \ln \left(\frac{d\tilde{\nu}}{d\nu} \right)$.

2. In the case where $\sigma^2 = \tilde{\sigma}^2 = 0$, $\mathbf{P}|_{\mathcal{F}_t}$ and $\mathbf{Q}|_{\mathcal{F}_t}$ are equivalent for all $t \geq 0$ if and only if the Lévy triplets satisfy (3) and

$$\tilde{\gamma} - \gamma = \int_{|x| \leq 1} x(\tilde{\nu} - \nu)(dx) \tag{4}$$

When \mathbf{P} and \mathbf{Q} are equivalent, the Radon-Nikodym derivative is

$$\frac{d\mathbf{Q}}{d\mathbf{P}} \Big|_{\mathcal{F}_t} = e^{\xi t}$$

where $\xi = (\xi_t)_{t \geq 0}$ is a Lévy process with Lévy triplet $(\sigma_\xi^2, \nu_\xi, \gamma_\xi)$ given by

$$\begin{cases} \sigma_\xi^2 = \sigma^2 \eta^2 \\ \nu_\xi = \nu \circ \psi^{-1} \\ \gamma_\xi = -\frac{\sigma^2 \eta^2}{2} - \int_{-\infty}^{\infty} (e^y - 1 - y1_{|y| \leq 1}) \nu_\xi(dy) \end{cases} \tag{5}$$

and η is such that

$$\tilde{\gamma} - \gamma - \int_{|x| \leq 1} x(\tilde{\nu} - \nu)(dx) = \begin{cases} \sigma^2 \eta & \text{if } \sigma > 0 \\ 0 & \text{if } \sigma = 0 \end{cases}$$

Change of Measure in Tempered Stable Processes

In this section, we present the change of measure for six tempered stable processes: the classical tempered stable (CTS) process, Kim-Rachev tempered stable (KRTS) process, modified tempered stable (MTS) process, normal tempered stable (NTS) process, and rapidly decreasing tempered stable (RDTS) process. The six processes are defined as follows:

- Let $\alpha \in (0, 2)$, $C, \lambda_+, \lambda_- > 0$, and $m \in \mathbb{R}$. A Lévy process $(X_t)_{t \geq 0}$ is referred to as the classical tempered stable (CTS) process³ if the charac-

teristic function of X_t is given by

$$\begin{aligned} \phi_{X_t}(u) = \exp(iumt - iutC\Gamma(1 - \alpha)(\lambda_+^{\alpha-1} - \lambda_-^{\alpha-1}) \\ + tC\Gamma(-\alpha)((\lambda_+ - iu)^\alpha - \lambda_+^\alpha \\ + (\lambda_- + iu)^\alpha - \lambda_-^\alpha)) \end{aligned}$$

If we take a special parameter C defined by

$$C = (\Gamma(2 - \alpha)(\lambda_+^{\alpha-2} + 2\lambda_-^{\alpha-2}))^{-1} \tag{6}$$

and $m = 0$ then $E[X_t] = 0$ and $V(X_t) = t$. In this case, X is called the standard CTS process with parameters $(\alpha, \lambda_+, \lambda_-)$.

- A Lévy process $(X_t)_{t \geq 0}$ is referred to as the generalized tempered stable (GTS) process if the characteristic function of X_t is given by

$$\begin{aligned} \phi_{X_t}(u) = \exp(iumt - iut\Gamma(1 - \alpha)(C_+\lambda_+^{\alpha-1} \\ - C_-\lambda_-^{\alpha-1}) \\ + tC_+\Gamma(-\alpha_+)((\lambda_+ - iu)^{\alpha_+} - \lambda_+^{\alpha_+}) \\ + tC_-\Gamma(-\alpha_-)((\lambda_- + iu)^{\alpha_-} - \lambda_-^{\alpha_-})), \end{aligned} \tag{7}$$

where $\alpha_+, \alpha_- \in (0, 1) \cup (1, 2)$, $C_+, C_-, \lambda_+, \lambda_- > 0$, and $m \in \mathbb{R}$. If we substitute

$$C_+ = \frac{p\lambda_+^{2-\alpha_+}}{\Gamma(2 - \alpha_+)}, \quad C_- = \frac{(1 - p)\lambda_-^{2-\alpha_-}}{\Gamma(2 - \alpha_-)} \tag{8}$$

where $p \in (0, 1)$, and $m = 0$ then $E[X_t] = 0$ and $V(X_t) = t$. In this case, X is called the standard GTS process with parameters $(\alpha_+, \alpha_-, \lambda_+, \lambda_-, p)$.

- Let $\alpha \in (0, 2) \setminus \{1\}$, $k_+, k_-, r_+, r_- > 0$, $p_+, p_- \in \{p > -\alpha \mid p \neq -1, p \neq 0\}$, and $m \in \mathbb{R}$. A Lévy process $(X_t)_{t \geq 0}$ is referred to as the Kim-Rachev (KR) process⁴ if the characteristic function of X_t is given by

$$\begin{aligned} \phi_{X_t}(u) = \exp(iumt - iut\Gamma(1 - \alpha) \\ \times \left(\frac{k_+r_+}{p_+ + 1} - \frac{k_-r_-}{p_- + 1} \right) \\ + tk_+H(iu; \alpha, r_+, p_+) \\ + tk_-H(-iu; \alpha, r_-, p_-)) \end{aligned}$$

where

$$H(x; \alpha, r, p) = \frac{\Gamma(-\alpha)}{p} ({}_2F_1(p, -\alpha; 1 + p; rx) - 1)$$

where ${}_2F_1$ is the hypergeometric function.⁵ If p_+ and p_- approach to the infinite, then the

KR process converges to the CTS process. If we substitute

$$k_+ = C \frac{\alpha + p_+}{r_+^\alpha}$$

$$k_- = C \frac{\alpha + p_-}{r_-^\alpha}$$

where

$$C = \frac{1}{\Gamma(2-\alpha)} \left(\frac{\alpha + p_+}{2 + p_+} r_+^{2-\alpha} + \frac{\alpha + p_-}{2 + p_-} r_-^{2-\alpha} \right)^{-1} \quad (9)$$

and $m = 0$ then $E[X_t] = 0$ and $V(X_t) = t$. In this case, X is called the standard KRTS process with parameters $(\alpha, r_+, r_-, p_+, p_-)$.

- Let $\alpha \in (0, 2) \setminus \{1\}$, $C, \lambda_+, \lambda_- > 0$, and $m \in \mathbb{R}$. A Lévy process $(X_t)_{t \geq 0}$ is referred to as the modified tempered stable (MTS) process⁶ if the characteristic function of X_t is given by

$$\begin{aligned} \phi_{X_t}(u) = & \exp(iumt + tC(G_R(u; \alpha, C, \lambda_+) \\ & + G_R(u; \alpha, C, \lambda_-)) \\ & + iutC(G_I(u; \alpha, \lambda_+) - G_I(u; \alpha, \lambda_-))) \end{aligned}$$

where for $u \in \mathbb{R}$,

$$\begin{aligned} G_R(x; \alpha, \lambda) = & 2^{-\frac{\alpha+3}{2}} \sqrt{\pi} \Gamma\left(-\frac{\alpha}{2}\right) \\ & \times ((\lambda^2 + x^2)^{\frac{\alpha}{2}} - \lambda^\alpha) \end{aligned}$$

and

$$\begin{aligned} G_I(x; \alpha, \lambda) = & 2^{-\frac{\alpha+1}{2}} \Gamma\left(\frac{1-\alpha}{2}\right) \lambda^{\alpha-1} \\ & \times \left[{}_2F_1\left(1, \frac{1-\alpha}{2}; \frac{3}{2}; -\frac{x^2}{\lambda^2}\right) - 1 \right] \end{aligned}$$

If we substitute

$$C = 2^{\frac{\alpha+1}{2}} \left(\sqrt{\pi} \Gamma\left(1 - \frac{\alpha}{2}\right) (\lambda_+^{\alpha-2} + \lambda_-^{\alpha-2}) \right)^{-1} \quad (10)$$

and $m = 0$ then $E[X_t] = 0$ and $V(X_t) = t$. In this case, X is called the standard MTS process with parameters $(\alpha, \lambda_+, \lambda_-)$.

- Let $\alpha \in (0, 2)$, $C, \lambda > 0$, $|\beta| < \lambda$, and $m \in \mathbb{R}$. A Lévy process $(X_t)_{t \geq 0}$ is referred to as the normal tempered stable (NTS) process⁷ if the

characteristic function of X_t is given by

$$\begin{aligned} \phi_{X_t}(u) = & \exp\left(iumt + iut2^{-\frac{\alpha+1}{2}} C \sqrt{\pi} \Gamma\left(-\frac{\alpha}{2}\right) \right. \\ & \times \alpha \beta (\lambda^2 - \beta^2)^{\frac{\alpha}{2}-1} + 2^{-\frac{\alpha+1}{2}} t C \sqrt{\pi} \Gamma\left(-\frac{\alpha}{2}\right) \\ & \left. \times ((\lambda^2 - (\beta + iu)^2)^{\frac{\alpha}{2}} - (\lambda^2 - \beta^2)^{\frac{\alpha}{2}}) \right) \end{aligned}$$

If we substitute

$$C = 2^{\frac{\alpha+1}{2}} \left(\sqrt{\pi} \Gamma\left(-\frac{\alpha}{2}\right) \alpha (\lambda^2 - \beta^2)^{\frac{\alpha}{2}-2} \right. \\ \left. \times (\alpha \beta^2 - \lambda^2 - \beta^2) \right)^{-1} \quad (11)$$

and $m = 0$ then $E[X_t] = 0$ and $V(X_t) = t$. In this case, X is called the standard NTS process with parameters (α, λ, β) .

- Let $\alpha \in (0, 2) \setminus \{1\}$, $C, \lambda_+, \lambda_- > 0$, and $m \in \mathbb{R}$. A Lévy process $(X_t)_{t \geq 0}$ is referred to as the rapidly decreasing tempered stable (RDTS) process⁸ if the characteristic function of X_t is given by

$$\begin{aligned} \phi_{X_t}(u) = & \exp(iumt + tC(G(iu; \alpha, \lambda_+) \\ & + G(-iu; \alpha, \lambda_-))) \end{aligned}$$

where

$$\begin{aligned} G(x; \alpha, \lambda) = & 2^{-\frac{\alpha}{2}-1} \lambda^\alpha \Gamma\left(-\frac{\alpha}{2}\right) \left(M\left(-\frac{\alpha}{2}, \frac{1}{2}; \frac{x^2}{2\lambda^2}\right) - 1 \right) \\ & + 2^{-\frac{\alpha}{2}-\frac{1}{2}} \lambda^{\alpha-1} x \Gamma\left(\frac{1-\alpha}{2}\right) \\ & \times \left(M\left(\frac{1-\alpha}{2}, \frac{3}{2}; \frac{x^2}{2\lambda^2}\right) - 1 \right) \end{aligned}$$

and M is the confluent hypergeometric function. See Andrews (1998). If we take a special parameter C defined by

$$C = 2^{\frac{\alpha}{2}} \left(\Gamma\left(1 - \frac{\alpha}{2}\right) (\lambda_+^{\alpha-2} + \lambda_-^{\alpha-2}) \right)^{-1} \quad (12)$$

and $m = 0$ then $E[X_t] = 0$ and $V(X_t) = t$. In this case, X is called the standard RDTS process with parameters $(\alpha, \lambda_+, \lambda_-)$.

The six tempered stable processes are pure jump Lévy processes with Lévy triplet $(0, \nu, \gamma)$, where $\gamma = m - \int_{|x|>1} x \nu(dx)$ and Lévy measures are presented in Table 1.

Table 1 Lévy Measures for Tempered Stable Processes

	Lévy Measure $\nu(dx)$
CTS	$C \left(\frac{e^{-\lambda_+ x}}{x^{1+\alpha}} 1_{x>0} + \frac{e^{-\lambda_- x }}{ x ^{1+\alpha}} 1_{x<0} \right) dx$
GTS	$\left(\frac{C_+ e^{-\lambda_+ x}}{x^{1+\alpha_+}} 1_{x>0} + \frac{C_- e^{-\lambda_- x }}{ x ^{1+\alpha_-}} 1_{x<0} \right) dx$
MTS	$\left(\frac{C_+ (\lambda_+ x)^{\frac{\alpha+1}{2}} K_{\frac{\alpha+1}{2}}(\lambda_+ x)}{x^{1+\alpha}} 1_{x>0} + \frac{C_- (\lambda_- x)^{\frac{\alpha+1}{2}} K_{\frac{\alpha+1}{2}}(\lambda_- x)}{ x ^{1+\alpha}} 1_{x<0} \right) dx$
NTS	$\frac{C e^{\beta x} (\lambda x)^{\frac{\alpha+1}{2}} K_{\frac{\alpha+1}{2}}(\lambda x)}{ x ^{1+\alpha}} dx$
KRTS	$\left(\frac{k_+ r_+^{-p_+}}{x^{1+\alpha}} \int_0^{r_+} e^{-x/s} s^{\alpha+p_+-1} ds 1_{x>0} + \frac{k_- r_-^{-p_-}}{ x ^{1+\alpha}} \int_0^{r_-} e^{- x /s} s^{\alpha+p_- -1} ds 1_{x<0} \right) dx$
RDTs	$\left(\frac{C_+ e^{-\frac{\lambda_+ x^2}{2}}}{x^{1+\alpha}} 1_{x>0} + \frac{C_- e^{-\frac{\lambda_- x ^2}{2}}}{ x ^{1+\alpha}} 1_{x<0} \right) dx$

Let $X = (X_t)_{t \geq 0}$ be one tempered stable process among the six tempered stable processes. Then $E[X_t] = mt$ and X has stationary and independent increments. Therefore, we have

$$\begin{aligned} E[X_T | \mathcal{F}_t] &= E[X_T - X_t | \mathcal{F}_t] + X_t \\ &= E[X_{T-t}] + X_t = m(T-t) + X_t \end{aligned}$$

and hence X is a martingale when $m = 0$.

The properties of tempered stable processes change under the change of measure using the Esscher transform. For example, let a process $X = (X_t)_{t \geq 0}$ be a symmetric CTS process under measure \mathbf{P} (that is $\lambda_+ = \lambda_- = \lambda$). Then the Lévy measure $\nu(dx)$ of X is given by

$$\nu(dx) = C \left(\frac{e^{-\lambda x}}{x^{1+\alpha}} 1_{x>0} + \frac{e^{-\lambda |x|}}{|x|^{1+\alpha}} 1_{x<0} \right) dx$$

Since we have $\int_{|x| \geq 1} e^{\theta x} \nu(dx) < \infty$ for some real number θ with $-\lambda \leq \theta \leq \lambda$, we can define a measure \mathbf{Q} equivalent to \mathbf{P} such that

$$\mathbf{Q}(A) |_{\mathcal{F}_t} = \int_A e^{\theta X_t - I(\theta)t} d\mathbf{P}, \quad A \in \mathcal{F}_t$$

where

$$\begin{aligned} I(\theta) &= \log E_{\mathbf{P}}[e^{\theta X_1}] = C \Gamma(-\alpha) ((\lambda - \theta)^\alpha \\ &\quad + (\lambda + \theta)^\alpha - 2\lambda^\alpha) \end{aligned}$$

Moreover, the Lévy measure $\tilde{\nu}(dx)$ of X under \mathbf{Q} is given by

$$\begin{aligned} \tilde{\nu}(dx) &= e^{\theta x} \nu(dx) \\ &= C \left(\frac{e^{-(\lambda-\theta)x}}{x^{1+\alpha}} 1_{x>0} + \frac{e^{-(\lambda+\theta)|x|}}{|x|^{1+\alpha}} 1_{x<0} \right) dx \end{aligned}$$

By the same argument, we discuss the relation between the symmetric MTS and NTS process. That is, let a process $X = (X_t)_{t \geq 0}$ be a symmetric MTS process under measure \mathbf{P} . Then the Lévy measure $\nu(dx)$ of X is given by

$$\nu(dx) = C (\lambda |x|)^{\frac{\alpha+1}{2}} K_{\frac{\alpha+1}{2}}(\lambda |x|) dx$$

Since we have $\int_{|x| \geq 1} e^{-\beta x} \nu(dx) < \infty$ for some real number β with $-\lambda \leq \beta \leq \lambda$, we can define a measure \mathbf{Q} equivalent to \mathbf{P} such that

$$\mathbf{Q}(A) |_{\mathcal{F}_t} = \int_A e^{-\beta X_t - I(-\beta)t} d\mathbf{P}, \quad A \in \mathcal{F}_t,$$

where

$$\begin{aligned} I(x) &= \log E_{\mathbf{P}}[e^{x X_1}] \\ &= C 2^{-\frac{\alpha+1}{2}} \sqrt{\pi} \Gamma\left(-\frac{\alpha}{2}\right) ((\lambda^2 + x^2)^{\frac{\alpha}{2}} - \lambda^\alpha) \end{aligned}$$

Moreover, the Lévy measure $\tilde{\nu}(dx)$ of X under \mathbf{Q} is given by

$$\tilde{\nu}(dx) = e^{-\beta x} \nu(dx) = C e^{-\beta x} (\lambda |x|)^{\frac{\alpha+1}{2}} K_{\frac{\alpha+1}{2}}(\lambda |x|) dx$$

which is the Lévy measure for the NTS process.

Table 2 Condition for Equivalent between **P** and **Q**

$(X_t)_{t \geq 0}$	Parameters under Measure P	Parameters under Measure Q	Equivalent Condition
CTS process	$(\alpha, C, \lambda_+, \lambda_-, m)$	$(\tilde{\alpha}, \tilde{C}, \tilde{\lambda}_+, \tilde{\lambda}_-, \tilde{m})$	$C = \tilde{C}, \alpha = \tilde{\alpha}$, and $\tilde{m} - m = C\Gamma(1 - \alpha)(\tilde{\lambda}_+^{\alpha-1} - \tilde{\lambda}_-^{\alpha-1} - \lambda_+^{\alpha-1} + \lambda_-^{\alpha-1})$
GTS process	$\left(\begin{matrix} \alpha_+, \alpha_-, C_+, C_- \\ \lambda_+, \lambda_-, m \end{matrix} \right)$	$\left(\begin{matrix} \tilde{\alpha}_+, \tilde{\alpha}_-, \tilde{C}_+, \tilde{C}_- \\ \tilde{\lambda}_+, \tilde{\lambda}_-, \tilde{m} \end{matrix} \right)$	$\alpha_+ = \tilde{\alpha}_+, \alpha_- = \tilde{\alpha}_-, C_+ = \tilde{C}_+, C_- = \tilde{C}_-$, and $\tilde{m} - m = C_+\Gamma(1 - \alpha_+)(\tilde{\lambda}_+^{\alpha_+-1} - \tilde{\lambda}_-^{\alpha_+-1}) - C_-\Gamma(1 - \alpha_-)(\lambda_+^{\alpha_+-1} + \lambda_-^{\alpha_+-1})$
MTS process	$(\alpha, C, \lambda_+, \lambda_-, m)$	$(\tilde{\alpha}, \tilde{C}, \tilde{\lambda}_+, \tilde{\lambda}_-, \tilde{m})$	$C = \tilde{C}, \alpha = \tilde{\alpha}$, and $\tilde{m} - m = 2^{-\frac{\alpha+1}{2}}C\Gamma\left(\frac{1-\alpha}{2}\right)(\tilde{\lambda}_+^{\alpha-1} - \tilde{\lambda}_-^{\alpha-1} - \lambda_+^{\alpha-1} + \lambda_-^{\alpha-1})$
NTS process	$(\alpha, C, \lambda, \beta, m)$	$(\tilde{\alpha}, \tilde{C}, \tilde{\lambda}, \tilde{\beta}, \tilde{m})$	$C = \tilde{C}, \alpha = \tilde{\alpha}$, and $\tilde{m} - m = \kappa(\tilde{\beta}(\tilde{\lambda}^2 - \tilde{\beta}^2)^{\frac{\alpha}{2}-1} - \beta(\lambda^2 - \beta^2)^{\frac{\alpha}{2}-1})$, where $\kappa = 2^{-\frac{\alpha-1}{2}}\sqrt{\pi}C\Gamma\left(1 - \frac{\alpha}{2}\right)$
KRTS process	$\left(\begin{matrix} \alpha_1, k_{1,+}, k_{1,-}, r_{1,+} \\ r_{1,-}, p_{1,+}, p_{1,-}, m_1 \end{matrix} \right)$	$\left(\begin{matrix} \alpha_2, k_{2,+}, k_{2,-}, r_{2,+} \\ r_{2,-}, p_{2,+}, p_{2,-}, m_2 \end{matrix} \right)$	$\begin{cases} p_{j,\pm} > 1/2 - \alpha_j \text{ and } p_{j,\pm} \neq 0, & \alpha_j \in (0, 1) \\ p_{j,\pm} > 1 - \alpha_j \text{ and } p_{j,\pm} \neq 0, & \alpha_j \in (1, 2) \end{cases}$, for $j = 1, 2$ $\alpha := \alpha_1 = \alpha_2$ $\frac{k_{1,+} + r_{1,+}^\alpha}{\alpha + p_{1,+}} = \frac{k_{2,+} + r_{2,+}^\alpha}{\alpha + p_{2,+}}, \frac{k_{1,-} + r_{1,-}^\alpha}{\alpha + p_{1,-}} = \frac{k_{2,-} + r_{2,-}^\alpha}{\alpha + p_{2,-}}$, and $m_2 - m_1 = \Gamma(1 - \alpha) \sum_{j=1,2} (-1)^j \left(\frac{k_{j,+} + r_{j,+}}{p_{j,+} + 1} - \frac{k_{j,-} + r_{j,-}}{p_{j,-} + 1} \right)$
RDTs process	$(\alpha, C, \lambda_+, \lambda_-, m)$	$(\tilde{\alpha}, \tilde{C}, \tilde{\lambda}_+, \tilde{\lambda}_-, \tilde{m})$	$C = \tilde{C}, \alpha = \tilde{\alpha}$, and $\tilde{m} - m = 2^{-\frac{\alpha+1}{2}}C\Gamma\left(\frac{1-\alpha}{2}\right)(\tilde{\lambda}_+^{\alpha-1} - \tilde{\lambda}_-^{\alpha-1} - \lambda_+^{\alpha-1} + \lambda_-^{\alpha-1})$

We can apply Theorem 3 to tempered stable processes. For example, let $X = (X_t)_{t \geq 0}$ be a CTS process with parameters $(\alpha, C, \lambda_+, \lambda_-, m)$ on measure **P** and a CTS process with parameters $(\tilde{\alpha}, \tilde{C}, \tilde{\lambda}_+, \tilde{\lambda}_-, \tilde{m})$ on measure **Q**. Then **P** and **Q** are equivalent if and only if $C = \tilde{C}, \alpha = \tilde{\alpha}$, and

$$\tilde{m} - m = C\Gamma(1 - \alpha)(\tilde{\lambda}_+^{\alpha-1} - \tilde{\lambda}_-^{\alpha-1} - \lambda_+^{\alpha-1} + \lambda_-^{\alpha-1}) \tag{13}$$

When **P** and **Q** are equivalent, the Radon-Nikodym derivative is $\left. \frac{d\mathbf{Q}}{d\mathbf{P}} \right|_{\mathcal{F}_t} = e^{U_t}$ where $U = (U_t)_{t \geq 0}$ is a Lévy process with Lévy triplet $(\sigma_U^2, \nu_U, \gamma_U)$ given by

$$\begin{aligned} \sigma_U^2 &= 0, \quad \nu_U = \nu \circ \psi^{-1}, \\ \gamma_U &= -\int_{-\infty}^{\infty} t(e^y - 1 - y1_{|y| \leq 1})(\nu \circ \psi^{-1})(dy) \end{aligned} \tag{14}$$

In equation (14), ν is the CTS Lévy measure given by

$$\nu(dx) = C \left(\frac{e^{-\lambda_+x}}{x^{1+\alpha}} 1_{x>0} + \frac{e^{-\lambda_-|x|}}{|x|^{1+\alpha}} 1_{x<0} \right) dx$$

and $\psi(x) = (\lambda_+ - \tilde{\lambda}_+)x1_{x>0} - (\lambda_- - \tilde{\lambda}_-)x1_{x<0}$. Proofs can be obtained by Theorem 3, but we will not discuss the proofs here.⁹

We can apply the same argument to the GTS, MTS, NTS, KRTS, and RDTs processes.¹⁰ The necessary and sufficient equivalent condition for change of measures for the six tempered stable distributions are presented in Table 2. Radon-Nikodym derivatives are omitted in the table.

By applying change of measures, we can obtain a martingale process from a CTS process. Let a process $X^0 = (X_t^0)_{t \geq 0}$ be a CTS process with

Table 3 Change of Measures for Standard TS Processes: $Y_t = \mu t + X_t$

$(X_t)_{t \geq 0}$ under Measure P	$(Y_t)_{t \geq 0}$ under Measure Q	Relations of Parameters
Standard CTS process with parameters $(\alpha, \lambda_+, \lambda_-)$	Standard CTS process with parameters $(\alpha, \tilde{\lambda}_+, \tilde{\lambda}_-)$	$\lambda_+^{\alpha-2} + \lambda_-^{\alpha-2} = \tilde{\lambda}_+^{\alpha-2} + \tilde{\lambda}_-^{\alpha-2},$ $\mu = \frac{\lambda_+^{\alpha-1} - \lambda_-^{\alpha-1} - \tilde{\lambda}_+^{\alpha-1} + \tilde{\lambda}_-^{\alpha-1}}{(1-\alpha)(\lambda_+^{\alpha-2} + \lambda_-^{\alpha-2})}$
Standard GTS process with parameters $(\alpha_+, \alpha_-, \lambda_+, \lambda_-, p)$	Standard GTS process with parameters $(\alpha_+, \alpha_-, \tilde{\lambda}_+, \tilde{\lambda}_-, \tilde{p})$	$p\lambda_+^{2-\alpha_+} = \tilde{p}\tilde{\lambda}_+^{2-\alpha_+}$ $(1-p)\lambda_-^{\alpha_- - 2} = (1-\tilde{p})\tilde{\lambda}_-^{\alpha_- - 2}$ $\mu = p \frac{\lambda_+^{\alpha_+ - 1} - \tilde{\lambda}_+^{\alpha_+ - 1}}{(1-\alpha_+)\lambda_+^{\alpha_+ - 1}} + (1-p) \frac{\tilde{\lambda}_-^{\alpha_- - 1} - \lambda_-^{\alpha_- - 1}}{(1-\alpha_-)\lambda_-^{\alpha_- - 1}}$
Standard MTS process with parameters $(\alpha, \lambda_+, \lambda_-)$	Standard MTS process with parameters $(\alpha, \tilde{\lambda}_+, \tilde{\lambda}_-)$	$\tilde{\lambda}_+^{\alpha-2} + \tilde{\lambda}_-^{\alpha-2} = \lambda_+^{\alpha-2} + \lambda_-^{\alpha-2}$ $\mu = \frac{\Gamma(\frac{1-\alpha}{2})(\lambda_+^{\alpha-1} - \lambda_-^{\alpha-1} - \tilde{\lambda}_+^{\alpha-1} + \tilde{\lambda}_-^{\alpha-1})}{\sqrt{\pi}\Gamma(1-\frac{\alpha}{2})(\tilde{\lambda}_+^{\alpha-2} + \tilde{\lambda}_-^{\alpha-2})}$
Standard NTS process with parameters (α, λ, β)	Standard NTS process with parameters $(\alpha, \tilde{\lambda}, \tilde{\beta})$	$\frac{\alpha\beta^2 - \lambda^2 - \beta^2}{(\lambda^2 - \beta^2)^{2-\frac{\alpha}{2}}} = \frac{\alpha\tilde{\beta}^2 - \tilde{\lambda}^2 - \tilde{\beta}^2}{(\tilde{\lambda}^2 - \tilde{\beta}^2)^{2-\frac{\alpha}{2}}}$ $\mu = \frac{\beta(\lambda^2 - \beta^2)^{\frac{\alpha}{2}-1} - \tilde{\beta}(\tilde{\lambda}^2 - \tilde{\beta}^2)^{\frac{\alpha}{2}-1}}{(\lambda^2 - \beta^2)^{\frac{\alpha}{2}-2}(\alpha\beta^2 - \lambda^2 - \beta^2)}$
Standard KRTS process with parameters $(\alpha, r_{1,+}, r_{1,-}, p_{1,+}, p_{1,-})$	Standard KRTS process with parameters $(\alpha, r_{2,+}, r_{2,-}, p_{2,+}, p_{2,-})$	$r_{2,+}, r_{2,-} > 0$ $\frac{\alpha + p_{1,+}}{2 + p_{1,+}} r_{1,+}^{2-\alpha} + \frac{\alpha + p_{1,-}}{2 + p_{1,-}} r_{1,-}^{2-\alpha} + \frac{\alpha + p_{2,+}}{2 + p_{2,+}} r_{2,+}^{2-\alpha} + \frac{\alpha + p_{2,-}}{2 + p_{2,-}} r_{2,-}^{2-\alpha}$ $\mu = \sum_{j=1,2} (-1)^j c_j \left(\frac{p_{j,+} + \alpha}{p_{j,+} + 1} r_{j,+}^{1-\alpha} - \frac{p_{j,-} + \alpha}{p_{j,-} + 1} r_{j,-}^{1-\alpha} \right)$
Standard RDTS process with parameters $(\alpha, \lambda_+, \lambda_-)$	Standard RDTS process with parameters $(\alpha, \tilde{\lambda}_+, \tilde{\lambda}_-)$	$\lambda_+^{\alpha-2} + \lambda_-^{\alpha-2} = \tilde{\lambda}_+^{\alpha-2} + \tilde{\lambda}_-^{\alpha-2},$ $\mu = \frac{\Gamma(\frac{1-\alpha}{2})(\lambda_+^{\alpha-1} - \lambda_-^{\alpha-1} - \tilde{\lambda}_+^{\alpha-1} + \tilde{\lambda}_-^{\alpha-1})}{\sqrt{2}\Gamma(1-\frac{\alpha}{2})(\tilde{\lambda}_+^{\alpha-2} + \tilde{\lambda}_-^{\alpha-2})}$

parameters $(\alpha, C, \lambda_+, \lambda_-, 0)$ on measure **P** and let $X = (X_t)_{t \geq 0}$ be a process with $X_t = mt + X_t^0$. Then X becomes the CTS process with parameters $(\alpha, C, \lambda_+, \lambda_-, m)$ on the measure **P**. The process X is not a martingale on the measure **P**, but we can obtain a measure where X is a martingale by the change of measures for CTS processes. We assume that $\tilde{\lambda}_+$ and $\tilde{\lambda}_-$ are positive real numbers such that

$$0 - m = C\Gamma(1-\alpha)(\tilde{\lambda}_+^{\alpha-1} - \tilde{\lambda}_-^{\alpha-1} - \lambda_+^{\alpha-1} + \lambda_-^{\alpha-1})$$

and we define a measure **Q** equivalent to **P** such that

$$\mathbf{Q}(A) |_{\mathcal{F}_t} = \int_A e^{U_t} d\mathbf{P}, \quad A \in \mathcal{F}_t$$

where $(U_t)_{t \geq 0}$ is the Lévy process with Lévy triplet $(\sigma_U^2, \nu_U, \gamma_U)$ given by equation (14). Then

the process X becomes the CTS process with parameters $(\alpha, C, \tilde{\lambda}_+, \tilde{\lambda}_-, 0)$ on the measure **Q**. Therefore, the process X is a martingale on measure **Q**.

Furthermore, by applying change of measures to the standard CTS process, we obtain the following result. Let $(X_t)_{t \geq 0}$ be a standard CTS process with parameters $(\alpha, \lambda_+, \lambda_-)$ under a measure **P**, and $\tilde{\lambda}_+, \tilde{\lambda}_- > 0$ and real number μ satisfy the following:

$$\begin{cases} \lambda_+^{\alpha-2} + \lambda_-^{\alpha-2} = \tilde{\lambda}_+^{\alpha-2} + \tilde{\lambda}_-^{\alpha-2} \\ \mu = \frac{\lambda_+^{\alpha-1} - \lambda_-^{\alpha-1} - \tilde{\lambda}_+^{\alpha-1} + \tilde{\lambda}_-^{\alpha-1}}{(1-\alpha)(\lambda_+^{\alpha-2} + \lambda_-^{\alpha-2})} \end{cases} \quad (15)$$

Then we can find a measure **Q** equivalent to **P** such that a process $(Y_t)_{t \geq 0}$ with $Y_t = \mu t + X_t$ is a standard CTS process with parameters $(\alpha, \tilde{\lambda}_+, \tilde{\lambda}_-)$ under a measure **Q**.

We apply the same argument to the standard GTS, standard MTS, standard NTS, standard KRIS, and standard RDIS processes. The relations of parameters between standard tempered stable process $(X_t)_{t \geq 0}$ under \mathbf{P} and standard tempered stable process $(Y_t)_{t \geq 0}$ with $Y_t = \mu t + X_t$ under \mathbf{Q} are presented in Table 3.

KEY POINTS

- The information available to all market agents at one time interprets the filtration.
- Conditional expectation is the best approximation of the price of assets, portfolios, and derivatives under information until the current time.
- Markov processes are used to explain the efficient market hypothesis in finance.
- Martingale processes describe the fair price or no-arbitrage price for an asset in finance.
- Change of measure on the Brownian motion process is achieved by Girsanov's theorem, while change of measure on the Lévy process is achieved by the Esscher transform or the generalized Girsanov theorem.
- Using the generalized Girsanov theorem, the tempered stable process becomes a martingale process.

NOTES

1. The general form of the Girsanov's theorem is presented in many articles including Karatzas and Shreve (1991), Oksendal (2000), and Klebaner (2005). The Black-Scholes option pricing formula is derived by applying Girsanov's theorem in Harrison and Pliska (1981).
2. We refer to such functions as cadlag functions.
3. See Koponen (1995), Boyarchenko and Levendorskii (2000), and Carr et al. (2002).
4. See Kim et al. (2008c, 2007).
5. See Andrews (1998).
6. See Kim et al. (2009).

7. See Barndorff-Nielsen and Levendorskii (2001).
8. See Bianchi et al. (2010) and Kim et al. (2010).
9. See Kim and Lee (2006) for more details.
10. See Kim et al. (2008a, 2008b, 2009, 2010) and Bianchi et al. (2010) for more details.

REFERENCES

- Andrews, L. D. (1998). *Special Functions of Mathematics for Engineers*. New York, Oxford University Press.
- Barndorff-Nielsen, O. E., and Levendorskii, S. (2001). Feller processes of normal inverse Gaussian type. *Quantitative Finance* 1.
- Bianchi, M. L., Rachev, S. T., Kim, Y. S., and Fabozzi, F. J. (2010). Tempered infinitely divisible distributions and processes. *Theory of Probability and Its Applications (TVP), Society for Industrial and Applied Mathematics (SIAM)* 55, 1: 58–86.
- Boyarchenko, S. I., and Levendorskii, S. Z. (2000). Option pricing for truncated Lévy processes. *International Journal of Theoretical and Applied Finance* 3.
- Carr, P., Geman, H., Madan, D., and Yor, M. (2002). The fine structure of asset returns: An empirical investigation. *Journal of Business* 75, 2: 305–332.
- Harrison, J. M., and Pliska, S. R. (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and Their Applications* 11, 3: 215–260.
- Karatzas, I., and Shreve, S. (1991). *Brownian Motion and Stochastic Calculus, 2nd ed.* New York, Springer.
- Kim, Y. S., and Lee, J. H. (2006). The relative entropy in CGMY processes and its applications to finance. *Mathematical Methods of Operations Research* 66, 2: 327–338.
- Kim, Y. S., Rachev, S. T., Bianchi, M. L., and Fabozzi, F. J. (2008a). Financial market models with Lévy processes and time-varying volatility. *Journal of Banking and Finance* 32, 7: 1363–1378.
- Kim, Y. S., Rachev, S. T., Bianchi, M. L., and Fabozzi, F. J. (2008b). A new tempered stable distribution and its application to finance. In G. Bol, S. T. Rachev, and R. Wuerth (Eds.), *Risk Assessment: Decisions in Banking and Finance*. Heidelberg, Springer.

- Kim, Y. S., Rachev, S. T., Bianchi, M. L., and Fabozzi, F. J. (2010). Tempered stable and tempered infinitely divisible GARCH models. *Journal of Banking and Finance* 34: 2096–2109.
- Kim, Y. S., Rachev, S. T., Bianchi, M. L., and Fabozzi, F. J. (2007). A new tempered stable distribution and its application to finance. In G. Bol, S. T. Rachev, and R. Wuerth (Eds.), *Risk Assessment: Decisions in Banking and Finance*, pp. 77–110. Heidelberg, Physika Verlag, Springer.
- Kim, Y. S., Rachev, S. T., Bianchi, M. L., and Fabozzi, F. J. (2008c). Financial market models with Lévy processes and time-varying volatility. *Journal of Banking and Finance*. 32, 7:1363–1378.
- Kim, Y. S., Rachev, S. T., Chung, D. M., and Bianchi, M. L. (2009). The modified tempered stable distribution, GARCH-models and option pricing. *Probability and Mathematical Statistics* 29, 1: 91–117.
- Klebaner, F. C. (2005). *Introduction to Stochastic Calculus with Applications*, 2nd ed. London, Imperial College Press.
- Koponen, I. (1995). Analytic approach to the problem of convergence of truncated Lévy flights towards the Gaussian stochastic process. *Physical Review E* 52.
- Oksendal, B. (2000). *Stochastic Differential Equations: An Introduction with Applications*, 5th ed. New York: Springer.
- Rachev, S. T., Kim, Y. S., Bianchi, M. L., and Fabozzi, F. J. (2011). *Financial Models with Lévy Processes and Volatility Clustering*. Hoboken, NJ: John Wiley & Sons.
- Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. New York: Cambridge University Press.

Change of Time Methods

ANATOLIY SWISHCHUK, PhD

Professor of Mathematics and Statistics, University of Calgary

Abstract: Change of time can be used in financial modeling to introduce stochastic volatility or solve many stochastic differential equations. The main idea of the change of time method is to change time from t to a nonnegative process $T(t)$ with nondecreasing sample paths (e.g., subordinator). Many Lévy processes may be written as time-changed Brownian motion. Lévy processes can also be used as a time change for other Lévy processes (subordinators). Using change of time, we can get an option pricing formula for an asset following geometric Brownian motion (e.g., Black-Scholes formula) and obtain an explicit option pricing formula for an asset following the mean-reverting process (e.g., continuous-time GARCH process).

In this entry, we provide an overview on *change of time* methods (CTM), and show how to solve many *stochastic differential equations* (SDEs) in finance (geometric Brownian motion [GBM], Ornstein-Uhlenbeck [OU], Vasiček, continuous-time GARCH, etc.) using the change of time method. As applications of CTM we present two different models: geometric Brownian motion (GBM) and *mean-reverting models*. The solutions of these two models are different. But the nice thing is that they can be solved by CTM like many other models mentioned in this entry. And moreover, we can use these solutions to find easy option pricing formulas: One is classic-Black-Scholes and another one is new for a mean-reverting asset. These formulas can be used in practice (for example, in the energy market) because they all are explicit.¹

This includes:

- CTM in *martingale* and *semimartingale* setting
- CTM in SDEs setting
- Subordination as a change of time

We present two applications of CTM:

- Black-Scholes formula
- *Explicit option pricing formula* for a mean-reverting asset

CHANGE OF TIME METHOD

The main idea of the change of time method is to change time from t to a nonnegative process $T(t)$ with nondecreasing sample paths. One example is subordinator: If $X(t)$ and $T(t) > 0$ are some processes, then $X(T(t))$ is subordinated to

$X(t); T(t)$ is a change of time. Another example is time-changed Brownian motion: $M(t) = B(T(t))$, where $B(t)$ is a Brownian motion and $T(t)$ is a subordinator (e.g., variance-gamma process² $V(t) = B(T(t))$, where $T(t)$ is a gamma process).

Bochner (1949) introduced the notion of change of time (time-changed Brownian motion). Clark (1973) introduced Bochner's change of time into financial economics. Feller (1966) introduced subordinated process $X(T(t))$ with Markov process $X(t)$ and $T(t)$ as a process with independent increments ($T(t)$ was called randomized operational time). Johnson (1979) introduced the time-changed *stochastic volatility* model (SVM) in continuous time. Johnson and Shanno (1987) studied the pricing of options using the time-changed stochastic volatility (SV) model. Ikeda and Watanabe (1981) introduced and studied change of time for the solution of SDEs. Barndorff-Nielsen, Nicolato, and Shephard (2003) studied the relationship between subordination and SVM using change of time ($T(t)$ -chronometer). Carr, Geman, Madan, and Yor (2003) used subordinated processes to construct SV for Lévy processes ($T(t)$ -business time).

The change of time method is closely associated with the embedding problem: To embed a process $X(t)$ in Brownian motion is to find a Wiener process $W(t)$ and an increasing family of stopping times $T(t)$ such that $W(T(t))$ has the same joint distribution as $X(t)$. Skorokhod (1965) first treated the embedding problem, showing that the sum of any sequence of independent random variables (r.v.) with mean zero and finite variation could be embedded in Brownian motion using stopping times. Dambis (1965) and Dubins and Schwartz (1965) independently showed that every continuous martingale could be embedded in Brownian motion. Knight (1971) discovered the multivariate extension of Dambis (1965) and Dubins and Schwartz's (1965) result. Huff (1969) showed that every process of pathwise bounded variation could be embedded in Brownian motion. Monroe (1972) proved that every right continu-

ous martingale could be embedded in a Brownian motion. Monroe (1978) proved that a process can be embedded in Brownian motion if and only if this process is a local semimartingale. Meyer (1971) and Papangelou (1972) independently discovered Knight's (1971) result for point processes.

Rosiński and Woyczyński (1986) considered time changes for integrals over stable Lévy processes. Kallenberg (1992) considered time change representations for stable integrals.

Lévy processes can also be used as a time change for other Lévy processes (subordinators). Madan and Seneta (1990) introduced the variance gamma (VG) process (Brownian motion with drift time changed by a gamma process). Geman, Madan, and Yor (2001) considered time changes for Lévy processes (business time). Carr, Geman, Madan, and Yor (2003) used change of time to introduce stochastic volatility into a Lévy model to achieve leverage effect and a long-term skew. Kallsen and Shiryaev (2001) showed that the Rosiński-Woyczyński-Kallenberg statement cannot be extended to any other Lévy process but symmetric α -stable. Swishchuk (2004, 2007) applied change of time method for options and swaps pricing for Gaussian models.³

The General Theory of Time Changes

The general theory of change of time for martingale and semimartingale theories⁴ is well known. In this entry we give a brief description of the change of time method in the following settings: martingales and stochastic differential equations.

Martingale and Semimartingale Settings of Change of Time

Let (Ω, \mathcal{F}, P) be a given probability space with a right continuous filtration $(\mathcal{F}_t)_{t \geq 0}$. Suppose M_t is a square integrable local continuous martingale such that $\lim_{t \rightarrow +\infty} \langle M \rangle(t) = +\infty$

almost sure (a.s.), where $\tau_t := \inf\{\mu : \langle M \rangle(u) > t\}$ and $\tilde{\mathcal{F}}_t = \mathcal{F}_{\tau_t}$. Then the time-changed process $B(t) := M(\tau_t)$ is an $\tilde{\mathcal{F}}_t$ -Brownian motion. Also, $M(t) = B(\langle M \rangle(t))$. Here, $\langle \cdot \rangle$ defines predictable quadratic variation.

If ϕ_t is a change of time process (i.e., any continuous \mathcal{F}_t -adapted process such that $\phi_0 = 0$, $t \rightarrow \phi_t$ is strictly increasing and $\lim_{t \rightarrow +\infty} \phi_t = +\infty$ a.s.) and if X_t is an \mathcal{F}_t -adapted semimartingale, then the process $\tilde{X}_t := X_{\tau_t}$ is an $\tilde{\mathcal{F}}_t$ -adapted semimartingale, where $\tau_t := \inf\{u : \phi_u > t\}$, and $\tilde{\mathcal{F}}_t := \mathcal{F}_{\tau_t}$. \tilde{X}_t is called the time change of X_t by ϕ_t .

Geman, Madan, and Yor (2001) consider pure jump Lévy processes (which are semimartingales) of finite variation with an infinite arrival rate of jumps as models for the logarithm of asset prices. These processes also may be written as time-changed Brownian motion. Their paper exhibits the explicit time change for each of a wide class of Lévy processes and shows that the time change is a weighted price move measure of time.

Stochastic Differential Equations Setting of Change of Time

The change of time method is used to solve the following SDE:

$$dX_t = \alpha(t, X_t)dB(t)$$

with $B(t)$ being a Brownian motion and $\alpha(t, x)$ being a "good" function of $t \geq 0$ and $x \in R$. Having solved the equation we can also solve the general SDE

$$dX_t = \beta(t, X_t)dt + \gamma(t, X_t)dB(t)$$

with drift $\beta(t, X_t)$ using the method of transformation of drift (the Girsanov transformation).⁵

Subordinators as Time Changes

Subordinators

Feller (1966) introduced a subordinated process X_{τ_t} for a Markov process X_t and τ_t a process with independent increments. τ_t was called a

randomized operational time. Increasing Lévy processes can also be used as a time change for other Lévy processes.⁶ Lévy processes of this kind are called subordinators. They are very important ingredients for building Lévy-based models in finance.⁷ If S_t is a subordinator, then its trajectories are almost surely increasing, and S_t can be interpreted as a "time deformation" and used to "time change" other Lévy processes. Roughly, if $(X_t)_{t \geq 0}$ is a Lévy process and $(S_t)_{t \geq 0}$ is a subordinator independent of X_t , then the process $(Y_t)_{t \geq 0}$ defined by $Y_t := X_{S_t}$ is a Lévy process.⁸ This time scale has the financial interpretation of business time,⁹ that is, the integrated rate of information arrival.

Subordinators and Stochastic Volatility

The time change method was used to introduce stochastic volatility into a Lévy model to achieve the leverage effect and a long-term skew.¹⁰ In the Bates (1996) model the leverage effect and long-term skew were achieved using correlated sources of randomness in the price process and the instantaneous volatility. The sources of randomness are thus required to be Brownian motions. In the Barndorff-Nielsen et al. (2001, 2002) model the leverage effect and long-term skew are generated using the same jumps in the price and volatility without a requirement for the sources of randomness to be Brownian motions. Another way to achieve the leverage effect and long-term skew is to make the volatility govern the time scale of the Lévy process driving jumps in the price. Carr et al. (2003) suggested the introduction of stochastic volatility into an exponential-Lévy model via a time change. The generic model here is $S_t = \exp(X_t) = \exp(Y_{\nu_t})$, where $\nu_t := \int_0^t \sigma_s^2 ds$. The volatility process should be positive and mean-reverting (i.e., an Ornstein-Uhlenbeck or Cox-Ingersoll-Ross process). Barndorff-Nielsen et al. (2003) reviewed and placed in context some of their recent work on stochastic volatility models including the relationship between subordination and stochastic volatility.

The main difference between the change of time method and the subordinator method is that in the former case the change of time process ϕ_t depends on the process X_t , but in the latter case, the subordinator S_t and Lévy process X_t are independent.

APPLICATIONS OF CHANGE OF TIME METHOD

The change of time method may be applied to get Black-Scholes formula for GBM, explicit option pricing formula for a mean-reverting asset, and to price swaps in financial models with stochastic volatility.

Black-Scholes by Change of Time Method

In the early 1970s, Black et al. (1973) made a major breakthrough by deriving a pricing formula for vanilla option written on a stock. Their model and its extensions assume that the probability distribution of the underlying cash flow at any given future time is lognormal. There are many proofs of their result, including partial differential equation and the martingale approach.¹¹

One of the aims of this entry is to give an idea of how to get the Black-Scholes result by the change of time method.

An Option Pricing Formula for a Mean-Reverting Asset Model Using a Change of Time Method

Some commodity prices, like oil and gas, exhibit mean reversion. This means that they tend over time to return to some long-term mean. This mean-reverting model is a one-factor version of the two-factor model made popular in the context of energy modeling by Pilipovic (1997). Black's model (1976) and Schwartz's model (1997) have become standard tools to price options on commodities. These models have the advantage that they give rise to closed-

form solutions for some types of option.¹² We note that the recent book by Geman (2005) discusses hard and soft commodities (that is, energy, agriculture, and metals) and also presents an analysis of economic and geopolitical issues in commodities markets. Here, we show how to get an explicit option pricing formula for a continuous-time GARCH asset price model using change of time.

One of the aims of this entry is to get an explicit option pricing formula for a mean-reverting asset using change of time method.

Swaps by Change of Time Method: Heston Model

One of the applications of change of time method is to value variance, volatility, covariance, and correlation swaps for Heston's (1993) model. Change of time method for pricing of different types of swaps for Heston's model and pricing of options has been considered in Swishchuk (2004, 2007, 2008c). Applications of change of time method to Lévy-based stochastic volatility models, interest rates, and energy derivatives have been considered in Swishchuk (2008a, 2008b, 2010a, 2010b).

In this section, we apply the change of time method to get the Black-Scholes formula and to obtain an explicit option pricing formula for a mean-reverting asset.

Change of Time Method

In this section we give a brief description of the *change of time method* for the martingales and stochastic differential equations. Throughout this entry we consider $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ to be a probability space with a right continuous filtration $(\mathcal{F}_t)_{t \geq 0}$

Change of Time Method in Martingale Setting

In this section, we describe the change of time method for a martingale $M(t) \in \mathcal{M}_2^{c,loc}$, the space of local square integrable continuous martingales.¹³

If $M(t) \in \mathcal{M}_2^{c,loc}$, $\lim_{t \rightarrow +\infty} \langle M \rangle (t) = +\infty$ a.s., $\tau_t := \inf\{u : \langle M \rangle (u) > t\}$ and $\tilde{\mathcal{F}}_t := \mathcal{F}_{\tau_t}$, then the following process with changed time

$$W(t) := M(\tau_t)$$

is an $\tilde{\mathcal{F}}_t$ -Brownian motion (or standard Wiener process).

Consequently, we can express a local martingale $M(t)$ using an $\tilde{\mathcal{F}}_t$ -Brownian motion $W(t)$ and an $\tilde{\mathcal{F}}_t$ -stopping time. (since $\{\langle M \rangle (t) \leq u\} = \{\tau_u \geq t\} \in \mathcal{F}_{\tau_u} = \tilde{\mathcal{F}}_u$)

$$M(t) = W(\langle M \rangle (t))$$

Change of Time Method in a Stochastic Differential Equation Setting

We consider the following generalization of the previous results to an SDE of the following form (without a drift)

$$dX(t) = \alpha(t, X(t))dW(t)$$

where $W(t)$ is a Brownian motion and $\alpha(t, X)$ is a continuous and measurable by t and X function on $[0, +\infty) \times R$.

The reason we consider this equation is if we solve the equation, then we can solve a more general equation with a drift $\beta(t, X)$ using the Girsanov transformation.¹⁴ The following result is used frequently to find a solution of an SDE using change of time method. The following theorem is due to Ikeda and Watanabe (1981).¹⁵

Let $\tilde{W}(t)$ be a one-dimensional \mathcal{F}_t -Wiener process with $\tilde{W}(0) = 0$, given on a probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t>0}, P)$ and let $X(0)$ be an \mathcal{F}_0 -adopted random variable. Define a continuous process $V = V(t)$ by

$$V(t) = X(0) + \tilde{W}(t)$$

Let ϕ_t be the change of time process:

$$\phi_t = \int_0^t \alpha^{-2}(\phi_s, X(0) + \tilde{W}(s))ds$$

If

$$X(t) := V(\phi_t^{-1}) = X(0) + \tilde{W}(\phi_t^{-1})$$

and $\tilde{\mathcal{F}}_t := \mathcal{F}_{\phi_t^{-1}}$, then there exists $\tilde{\mathcal{F}}_t$ -adopted Wiener process $W = W(t)$ such that $(X(t), W(t))$ is a solution of the initial equation on probability space $(\Omega, \mathcal{F}, \tilde{\mathcal{F}}_t, P)$.¹⁶

We note that the solution of the following SDE

$$dX(t) = a(X(t))dW(t)$$

may be presented in the following form (which follows from the previous theorem)

$$X(t) = X(0) + \tilde{W}(\phi_t^{-1})$$

where $a(X)$ is a continuous measurable function, $\tilde{W}(t)$ is an n -dimensional \mathcal{F}_t -Wiener process with $\tilde{W}(0) = 0$, given on a probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, P)$ and $X(0)$ is an \mathcal{F}_0 -adopted random variable. In this case¹⁷

$$\phi_t = \int_0^t a^{-2}(X(0) + \tilde{W}(s))ds$$

and

$$\phi_t^{-1} = \int_0^t a^2(X(0) + \tilde{W}(\phi_s^{-1}))ds$$

Examples: Solutions of Some SDEs¹⁸

1. *Solution for Ornstein-Uhlenbeck (OU) Process Using Change of Time.*

Let S_t satisfy the following SDE:

$$dS_t = -\alpha S_t dt + \sigma dW_t$$

Then S_t may be presented in the following form using the change of time method:

$$S_t = e^{-\alpha t}[S_0 + \tilde{W}(\phi_t^{-1})]$$

where ϕ_t^{-1} satisfies

$$\phi_t^{-1} = \sigma^2 \int_0^t (e^{\alpha s}(S_0 + \tilde{W}(\phi_s^{-1})))^2 ds$$

2. *Solution for Vasićek Process Using Change of Time.*

Let S_t satisfy the following SDE:

$$dS_t = \alpha(b - S_t)dt + \sigma dW_t$$

Then S_t may be presented in the following form using the change of time method

$$S_t = e^{-\alpha t}[S_0 - b + \tilde{W}(\phi_t^{-1})]$$

where ϕ_t^{-1} satisfies

$$\phi_t^{-1} = \sigma^2 \int_0^t (e^{\alpha s} (S_0 - b + \tilde{W}(\phi_s^{-1})) + b)^2 ds$$

The above theorem may also be applied to solve the Cox-Ingersoll-Ross (1985) equation, mean-reversion equation for commodity price (Pilipovic, 1997) and geometric Brownian motion equation (Black-Scholes, 1973).¹⁹

Black-Scholes Formula by Change of Time Method

Let $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ be a probability space with a sample space Ω , σ -algebra of Borel sets \mathcal{F} and probability P . The filtration $\mathcal{F}_t, t \in [0, T]$ is the natural filtration of a standard Brownian motion $W_t, t \in [0, T]$, and $\mathcal{F}_T = \mathcal{F}$.

Black-Scholes Formula

The well-known Black-Scholes (1973) formula states if we have (B, S) -security market consisting of riskless asset $B(t)$ with interest rate r as a constant

$$dB(t) = rB(t)dt, \quad B(0) > 0, \quad r > 0 \quad (1)$$

and risky asset (stock) $S(t)$

$$dS(t) = \mu S(t)dt + \sigma S(t)dW(t), \quad S(0) > 0 \quad (2)$$

where $\mu \in R$ is an appreciation rate, $\sigma > 0$ is a volatility, then the option price formula for European call option with pay-off function $f(T) = \max(S(T) - K, 0)$ ($K > 0$ is a strike price) has the following look

$$C(T) = S(0)\Phi(y_+) - e^{-rT}K\Phi(y_-) \quad (3)$$

where

$$y_{\pm} := \frac{\ln\left(\frac{S(0)}{K}\right) + \left(r \pm \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}} \quad (4)$$

and

$$\Phi(y) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y e^{-\frac{x^2}{2}} dx \quad (5)$$

Solution of SDE for Geometric Brownian Motion using Change of Time Method

The solution of equation (2) has the following look:

$$S(t) = e^{\mu t} (S(0) + \tilde{W}(\phi_s^{-1})) \quad (6)$$

where $\tilde{W}(t)$ is a one-dimensional Wiener process,

$$\phi_t^{-1} = \sigma^2 \int_0^t [S(0) + \tilde{W}(\phi_s^{-1})]^2 ds$$

and

$$\phi_t = \sigma^{-2} \int_0^t [S(0) + \tilde{W}(s)]^{-2} ds$$

Black-Scholes Formula by Change of Time Method

In a risk-neutral world the dynamic of stock price $S(t)$ has the following look:

$$dS(t) = rS(t)dt + \sigma S(t)dW^*(t) \quad (7)$$

where

$$W^*(t) := W(t) + \frac{\mu - r}{\sigma} t \quad (8)$$

From (6) we have the solution of equation (7)

$$S(t) = e^{rt} [S(0) + \tilde{W}^*(\phi_t^{-1})] \quad (9)$$

where

$$\tilde{W}^*(\phi_t^{-1}) = S(0)(e^{\sigma W^*(t) - \frac{\sigma^2 t}{2}} - 1) \quad (10)$$

and $W^*(t)$ is defined in (8).

Let E_{P^*} be an expectation under risk-neutral measure (or martingale measure) P^* (i.e., process $e^{-rT}S(t)$ is a martingale under the measure P^*).

Then the option pricing formula for European call option with payoff function

$$f(T) = \max[S(T) - K, 0]$$

has the following look

$$\begin{aligned} C(T) &= e^{-rT} E_{P^*}[f(T)] \\ &= e^{-rT} E_{P^*}[\max(S(T) - K, 0)] \end{aligned} \quad (11)$$

Using change of time method we have the following representation for the process $S(t)$ (see (9))

$$S(t) = e^{rt}[S(0) + \tilde{W}^*(\phi_t^{-1})]$$

where $\tilde{W}^*(\phi_t^{-1})$ is defined in (10). From (7)–(11), after substitution $\tilde{W}^*(\phi_t^{-1})$ into (9) and $S(T)$ into (11), it follows that

$$\begin{aligned} C(T) &= e^{-rT} E_{P^*}[\max(S(T) - K, 0)] \\ &= e^{-rT} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \max[S(0)e^{\sigma u\sqrt{T} + (r - \frac{\sigma^2}{2})T} \\ &\quad - K, 0] e^{-\frac{u^2}{2}} du \end{aligned} \quad (12)$$

Let y_0 be a solution of the following equation

$$S(0)e^{\sigma y\sqrt{T} + (r - \sigma^2/2)T} = K$$

namely,

$$y_0 = \frac{\ln\left(\frac{K}{S(0)}\right) - (r - \sigma^2/2)T}{\sigma\sqrt{T}}.$$

Then (12) may be presented in the following form

$$\begin{aligned} C(T) &= e^{-rT} \frac{1}{\sqrt{2\pi}} \int_{y_0}^{+\infty} (S(0)e^{\sigma u\sqrt{T} + (r - \frac{\sigma^2}{2})T} - K) \\ &\quad \times e^{-\frac{u^2}{2}} du \end{aligned} \quad (13)$$

Finally, straightforward calculation of the integral in the right-hand side of (13) gives us the

Black-Scholes result:²⁰

$$\begin{aligned} C(T) &= \frac{1}{\sqrt{2\pi}} \int_{y_0}^{+\infty} S(0)e^{\sigma u\sqrt{T} - \frac{\sigma^2 T}{2}} e^{-u^2/2} du \\ &\quad - Ke^{-rT}[1 - \Phi(y_0)] \\ &= \frac{S(0)}{\sqrt{2\pi}} \int_{y_0 - \sigma\sqrt{T}}^{+\infty} e^{-u^2/2} du - Ke^{-rT}[1 - \Phi(y_0)] \\ &= S(0)[1 - \Phi(y_0 - \sigma\sqrt{T})] - Ke^{-rT}[1 - \Phi(y_0)] \\ &= S(0)\Phi(y_+) - Ke^{-rT}\Phi(y_-) \end{aligned} \quad (14)$$

where y_{\pm} and $\Phi(y)$ are defined in (4) and (5).

Explicit Option Pricing Formula for Mean-Reverting Asset Model (MRAM) by Change of Time Method

In this section, we consider a risky asset S_t following the mean-reverting stochastic process given by the following stochastic differential equation

$$dS_t = a(L - S_t)dt + \sigma S_t dW_t \quad (15)$$

where W_t is an \mathcal{F}_t -measurable one-dimensional standard Wiener process, $\sigma > 0$ is the volatility, constant L is called the long-term mean of the process, to which it reverts over time, and $a > 0$ measures the “strength” of mean reversion. We find explicit solution of the equation (15) using the change of time method, give some properties of the mean-reverting asset S_t , and present an explicit option pricing formula for the European call option for this mean-reverting asset model of commodity price.

Explicit Solution of SDE for MRAM Equation

$$dS_t = a(L - S_t)dt + \sigma S_t dW_t$$

in (15) has the following solution

$$S_t = e^{-at}[S_0 - L + \tilde{W}(\phi_t^{-1})] + L$$

where $\tilde{W}(\phi_t^{-1})$ is a one-dimensional Wiener process and

$$\phi_t^{-1} = \sigma^2 \int_0^t (S_0 - L + \tilde{W}(\phi_s^{-1}) + e^{as}L)^2 ds$$

which follows from the substitution

$$V_t := e^{at}(S_t - L)$$

and theorem above.

Explicit Option Pricing Formula for European Call Option under Risk-Neutral Measure

In this section, we are going to obtain an explicit option pricing formula for a European call option under risk-neutral measure P^* using the change of time method.

Mean-Reverting Risk-Neutral Asset Model

Consider the model given by (15)

$$dS_t = a(L - S_t)dt + \sigma S_t dW_t$$

We want to find a probability P^* equivalent to P , under which the process $e^{-rt}S_t$ is a martingale, where $r > 0$ is a constant interest rate.

In a risk-neutral world the model in (15) takes the following look:

$$dS_t = a^*(L^* - S_t)dt + \sigma S_t dW_t^* \quad (16)$$

where

$$a^* := a + \lambda\sigma, \quad L^* := \frac{aL}{a + \lambda\sigma} \quad (17)$$

$$W_t^* = W_t + \lambda \int_0^t S(u)du \quad (18)$$

and $\lambda \in R$ is a market price of risk, which follows from the Girsanov theorem.²¹

Now, we are going to apply our method of changing of time to the model (16) to obtain the explicit option pricing formula.

Explicit Solution for Mean-Reverting Risk-Neutral Asset Model

Applying the above results to our model (16) we obtain the explicit solution (19) for our risk-neutral model (16). The explicit solution for the risk-neutral model given by (16) has the following look

$$S_t = e^{-a^*t} [S_0 - L^* + \tilde{W}^*((\phi_t^*)^{-1})] + L \quad (19)$$

where $\tilde{W}^*(t)$ is an \mathcal{F}_t -measurable standard one-dimensional Wiener process in (18) under measure P^* , $(\phi_t^*)^{-1}$ is an inverse function to ϕ_t^* :

$$\phi_t^* = \sigma^{-2} \int_0^t (S_0 - L^* + \tilde{W}^*(s) + e^{a^*s} L^*)^{-2} ds \quad (20)$$

We note that

$$(\phi_t^*)^{-1} = \sigma^2 \int_0^t (S_0 - L^* + \tilde{W}^*((\phi_t^*)^{-1}) + e^{a^*s} L^*)^2 ds$$

where a^* and L^* are defined in (17).

Explicit Option Pricing Formula for European Call Option under Risk-Neutral Measure

The payoff function f_T for the European call option equals

$$f_T = (S_T - K)^+ := \max(S_T - K, 0)$$

where S_T is an asset price, T is an expiration time (maturity), and K is a strike price.

In this way (see (19)),

$$\begin{aligned} f_T &= [e^{-a^*T}(S_0 - L^* + \tilde{W}^*(\phi_T^{-1})) + L^* - K]^+ \\ &= [S(0)e^{-a^*T} e^{\sigma W^*(T) - \frac{\sigma^2 T}{2}} \\ &\quad + a^* L^* e^{-a^*T} e^{\sigma W^*(T) - \frac{\sigma^2 T}{2}} \int_0^T e^{a^*s} e^{-\sigma W^*(s) \frac{\sigma^2 s}{2}} \\ &\quad \times ds - K]^+ \end{aligned} \quad (21)$$

The explicit option pricing formula for the European call option under a risk-neutral measure for mean-reverting asset $S(t)$ in (21) has the following look:

$$\begin{aligned} C_T^* &= e^{-(r+a^*)T} S(0) \Phi(y_+) - e^{-rT} K \Phi(y_-) \\ &\quad + L^* e^{-(r+a^*)T} [(e^{a^*T} - 1) - \int_0^{y_0} z F_T^*(dz)] \end{aligned} \quad (22)$$

where y_0 is the solution of the following equation

$$\begin{aligned} y_0 &= \frac{\ln\left(\frac{K}{S(0)}\right) + \left(\frac{\sigma^2}{2} + a^*\right)T}{\sigma\sqrt{T}} \\ &\quad - \frac{\ln\left(1 + \frac{a^*L^*}{S(0)} \int_0^T e^{a^*s} e^{-\sigma y_0 \sqrt{s} + \frac{\sigma^2 s}{2}} ds\right)}{\sigma\sqrt{T}} \end{aligned} \quad (23)$$

$$y^+ := \sigma\sqrt{T} - y_0 \quad \text{and} \quad y_- := -y_0,$$

$$a^* := a + \lambda\sigma, \quad L^* := \frac{aL}{a + \lambda\sigma} \quad (24)$$

λ is the market price of risk and $F_T^*(dz)$ is the distribution with characteristic function

$$\phi_\lambda^*(T) = e^{i\lambda(e^{a^*T}-1)}, \quad i := \sqrt{-1}, \quad \lambda \in \mathbb{C}$$

This result can be obtained from the following expression

$$\begin{aligned} C_T &:= e^{-rT} E_{P^*} f_T \\ &= e^{-rT} E_{P^*} [e^{-a^*T} (S_0 - L + \tilde{W}^*(\phi_T^{-1})) + L^* - K]^+ \\ &= \frac{1}{\sqrt{2\pi}} e^{-rT} \int_{-\infty}^{+\infty} \max[S(0)e^{-a^*T} e^{\sigma y \sqrt{T} - \frac{\sigma^2 T}{2}} \\ &\quad + aLe^{-a^*T} e^{\sigma y \sqrt{T} - \frac{\sigma^2 T}{2}} \int_0^T e^{a^*s} e^{-\sigma y \sqrt{s} + \frac{\sigma^2 s}{2}} ds \\ &\quad - K, 0] e^{-\frac{y^2}{2}} dy \end{aligned}$$

and above-mentioned results.

Connection with Black-Scholes Result: $L^* = 0$ and $a^* = -r$ and Black-Scholes formula follows!

If $L^* = 0$ and $a^* = -r$ then we obtain from (22)

$$C_T = S(0)\Phi(y_+) - e^{-rT}K\Phi(y_-) \quad (25)$$

where

$$y_+ := \sigma\sqrt{T} - y_0 \quad \text{and} \quad y_- := -y_0 \quad (26)$$

and y_0 is the solution of the following equation (see (23))

$$S(0)e^{-rT}e^{\sigma y_0 \sqrt{T} - \frac{\sigma^2 T}{2}} = K$$

or

$$y_0 = \frac{\ln\left(\frac{K}{S(0)}\right) + \left(\frac{\sigma^2}{2} - r\right)T}{\sigma\sqrt{T}} \quad (27)$$

and

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{y^2}{2}} dy$$

But (25)–(27) is exactly the well-known Black-Scholes result!

In this way, we can see that the option pricing formula in (22) for the mean-reverting asset $S(t)$ consists of a Black-Scholes part and an additional part due to mean reversion.

The results of this section may be also used to model and price variance and volatility swaps in energy and commodity markets for assets with stochastic volatility that are described

by a continuous-time mean-reverting GARCH model; see Swishchuk (2010a).

KEY POINTS

- The main idea of the change of time method is to change time from t to a nonnegative process $T(t)$ with nondecreasing sample paths (e.g., subordinator).
- Many Lévy processes may be written as time-changed Brownian motion.
- Lévy processes can also be used as a time change for other Lévy processes (subordinators).
- Change of time can be used to introduce stochastic volatility or solve many stochastic differential equations.
- Using change of time, we can get an option pricing formula for an asset following geometric Brownian motion such as the Black-Scholes formula.
- Using change of time, we can get an explicit option pricing formula for an asset following the mean-reverting process, such as continuous-time GARCH process.

NOTES

1. Swishchuk (2007) and Swishchuk (2008c).
2. Madan et al. (1990).
3. Barndorff-Nielsen and Shiryaev (2010) state the main ideas and results of the stochastic theory of change of time and change of measure.
4. Ikeda and Watanabe (1981).
5. Ikeda and Watanabe (1981), Chapter IV, Section 4, p. 176.
6. Applebaum (2004), Barndorff-Nielsen et al. (2001), Barndorff-Nielsen et al. (2003), Bertoin (1996), Cont et al. (2004), and Schoutens (2003).
7. Cont et al. (2004) and Schoutens (2003).
8. Cont et al. (2004).
9. Geman et al. (2001).
10. Carr et al. (2003).
11. Wilmott et al. (1995) and Elliott et al. (1999).
12. Wilmott (2000).

13. Ikeda and Watanabe (1981), Theorem 7.2, Chapter 2.
14. Ikeda and Watanabe (1981), Chapter 4, Section 4.
15. Chapter IV, Theorem 4.3.
16. The proof of this theorem may be found in Ikeda and Watanabe (1981), Chapter IV, Theorem 4.3.
17. Ikeda and Watanabe (1981), Chapter IV, Example 4.2.
18. Swishchuk (2007).
19. Swishchuk (2007).
20. Black and Scholes (1973).
21. Elliott et al. (1999).

REFERENCES

- Applebaum, D. (2003). *Lévy Processes and Stochastic Calculus*. Cambridge: Cambridge University Press.
- Barndorff-Nielsen, E., and Shiryaev, A. N. (2009) *Change of Time and Change of Measures*. New York: World Scientific.
- Barndorff-Nielsen, O. E., Nicolato, E., and Shephard, N. (1996). Some recent development in stochastic volatility modeling. *Quantitative Finance* 2: 11–23.
- Bates, D. (1996). Jumps and stochastic volatility: The exchange rate processes implicit in Deutschemark options. *Review Finance Studies* 9: 69–107.
- Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–654.
- Black, F. (1976). The pricing of commodity contracts. *Journal of Financial Economics* 3: 167–179.
- Bochner, S. (1949). Diffusion equation and stochastic processes. USA: *Proceedings of National Academy of Sciences* 85: 369–370.
- Carr, P., Geman, H., Madan, D., and Yor, M. (2003). Stochastic volatility for Lévy processes. *Mathematical Finance* 13, 3: 345–382.
- Carr, P., and Wu, L. (2003). The finite moment logstable process and option pricing. *The Journal of Finance* 53, 2: 753–777.
- Carr, P., and Wu, L. (2004). Time-changed Lévy processes and option pricing. *Journal of Financial Economics* 71: 113–141.
- Clark, P. (1973). A subordinated stochastic process model with fixed variance for speculative prices. *Econometrica* 41: 135–156.
- Cont, R., and Tankov, P. (2004). *Financial Modeling with Jump Processes*. Princeton, NJ: Chapman & Hall/CRC.
- Cox, J., Ingersoll, J., and Ross, S. (1985). A theory of the term structure of interest rate. *Econometrica* 53: 385–407.
- Dubins, E., and Schwartz, G. E. (1965). On continuous martingales. *Proceedings of the National Academy of Sciences* 53: 913–916.
- Elliott, R., and Swishchuk, A. (2007). Pricing options and variance swaps in Markov-modulated Brownian markets. In *Hidden Markov Models in Finance*. New York: Springer.
- Geman, H. (2005). *Commodities and Commodity Derivatives: Modelling and Pricing for Agricultural, Metals and Energy*. New York: John Wiley & Sons.
- Geman, H., Madan, D., and Yor, M. (2002). Time changes for Lévy processes. *Mathematical Finance* 11: 79–96.
- Heston, S. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies* 6: 327–343.
- Huff, B. (2000). The loose subordination of differential processes to Brownian motion. *Annals of Mathematical Statistics* 40: 1603–1609.
- Ikeda, N., and Watanabe, S. (1981). *Stochastic Differential Equations and Diffusion Processes*. Tokyo: North-Holland/Kodansha Ltd.
- Johnson, H. (1979). Option pricing when the variance rate is changing. *Working paper*, Los Angeles: University of California.
- Johnson, H., and Shanno, D. (1987). Option pricing when the variance is changing. *Journal of Financial Quantitative Analysis* 22: 143–152.
- Kallsen, J., and Shiryaev, A. (2002). Time change representation of stochastic integrals. *Theory Probability and Its Applications* 46, 3: 522–528.
- Knight, F. (1971). A reduction of continuous, square-integrable martingales to Brownian motion. *Martingales*. Berlin: Springer, Lecture Notes in Mathematics 190: 19–31.
- Madan, D., and Seneta, E. (1999). The variance gamma (VG) model for share market returns. *Journal of Business* 63: 511–524.
- Meyer, P. A. (1971). Demonstration simplifiée d'un théorème de Knight. *Seminaire de Probabilités V*. Berlin: Springer, Lecture Notes in Mathematics 191: 191–195.
- Monroe, I. (1972). On embedding right continuous martingales in Brownian motion. *Annals of Mathematical Statistics* 43: 1293–1311.

- Monroe, I. (1978). Processes that can be embedded in Brownian motion. *The Annals of Probability* 6, 1: 42–56.
- Papangelou, F. (1972). Integrability of expected increments of point processes and a related random change of scale. *Transactions of American Mathematical Society* 165: 486–506.
- Pilipović, D. (1997). *Valuing and Managing Energy Derivatives*. New York: McGraw-Hill.
- Schoutens, W. (2003). *Lévy Processes in Finance. Pricing Financial Derivatives*. New York: Wiley & Sons.
- Schwartz, E. (1997). The stochastic behavior of commodity prices: Implications for pricing and hedging. *Journal of Finance* 52: 923–973.
- Shephard, N. (2005). *Stochastic Volatility: Selected Readings*. Oxford: Oxford University Press.
- Skorokhod, A. (1965). *Studies in the Theory of Random Processes*. Reading: Addison-Wesley.
- Swishchuk, A. (2010a). Variance and volatility swaps in energy markets. *Journal of Energy Markets* forthcoming.
- Swishchuk, A. (2010b). Multi-factor Lévy models for pricing financial and energy derivatives. *Canadian Applied Mathematics Quarterly* 17, 4.
- Swishchuk, A. (2008a). Lévy-based interest rate derivatives: Change of time and PIDEs. *Canadian Applied Mathematics Quarterly* 16, 2.
- Swishchuk, A. (2008b). Multi-factor Lévy models: Change of time and pricing of financial and energy derivatives. *Working Paper*, Calgary: University of Calgary.
- Swishchuk, A. (2008c). Explicit option pricing formula for a mean-reverting asset in energy market. *Journal of Numerical Applied Mathematics* 1, 96: 216–233.
- Swishchuk, A. (2007). Change of time method in mathematical finance. *Canadian Applied Mathematics Quarterly* 15, 3: 299–336.
- Swishchuk, A. (2004). Modelling and valuing of variance and volatility swaps for financial markets with stochastic volatilities. *Wilmott Magazine* 2, September: 64–72.
- Wilmott, P., Howison, S., and Dewynne, J. (1995). *Option Pricing: Mathematical Models and Computations*. Oxford: Oxford Financial Press.

Term Structure Modeling

The Concept and Measures of Interest Rate Volatility

ALEXANDER LEVIN, PhD

Director, Financial Engineering, Andrew Davidson & Co., Inc.

Abstract: The knowledge of interest rates and cash flows represents the basis for valuation of fixed income financial instruments. In reality, not only are future interest rates random, but the future cash flows of many securitized investments are also uncertain, as they depend (are “contingent”) on interest rates. Valuation of rate options and embedded option bonds, including MBS and ABS, requires sophisticated models of this randomness.

In this entry, we introduce the concepts of market volatility and discuss how it is measured. The dynamics of rates are subject to market forces, mean reversion, and combinations of diffusions and jumps.

BASIC DEFINITIONS AND FIRST FINDINGS

We can't tell in advance what interest rates will be. Investors may be either enriched or bankrupted from sudden changes in interest rates. Financial institutions devote considerable resources to risk management and hedging. Yet, if future interest rates were deterministic, there would be no need to hedge. Coping with uncertainty is a central feature of investment markets.

The pricing of options and embedded-options instruments utilizes a statistical concept to describe the magnitude of potential interest rates changes. The key notion is the volatility of interest rates. While this term conjures up images

of instability, flares of activity, and unpredictability, it is actually a very specific description of the range of possible outcomes. More precisely, volatility can be defined as the standard deviation of a rate's annualized daily increments. Table 1 provides an example for yields on the 10-year Treasury measured over 10 consecutive business days. As part of the measurement, we will be taking a daily time series and then transforming into “absolute returns” and “relative returns”—much like measuring portfolio performance.

The absolute rate changes are computed by taking the difference between the interest rates on successive days. The relative changes are computed by dividing the absolute change by the starting rate. For example, for the first day the absolute change is $5.00343 - 5.03234 = -0.0289$. The relative increment is $-0.0289 / 5.03234 = -0.0057$. In order to calculate the daily volatility, we just take the standard deviation of the daily absolute and relative change series. In the example above, the standard

Table 1 Example of Volatility Calculations

Date	Rate	Absolute Increments	Relative Increments
03-Jun-02	5.03234		
04-Jun-02	5.00343	-0.0289	-0.0057
05-Jun-02	5.04900	0.0456	0.0091
06-Jun-02	5.01176	-0.0372	-0.0074
07-Jun-02	5.06165	0.0499	0.0100
10-Jun-02	5.03885	-0.0228	-0.0045
11-Jun-02	4.97500	-0.0639	-0.0127
12-Jun-02	4.95004	-0.0250	-0.0050
13-Jun-02	4.90280	-0.0472	-0.0095
14-Jun-02	4.80276	-0.1000	-0.0204

deviations are 0.048 (absolute increments) and 0.00966 (relative increments). The former number is the standard deviation for daily absolute increments; the latter number represents that of the daily relative changes. To compute volatility, we place these daily measures on an annual basis scaling by the number of trading days in the year (approximately 260):

Relative Volatility

$$= \text{Daily Standard Relative Deviation} \times \sqrt{260}$$

$$= 0.00966 \times \sqrt{260} = 0.1557$$

Absolute Volatility

$$= \text{Daily Standard Absolute Deviation} \times \sqrt{260}$$

$$= 0.0480 \times \sqrt{260} = 0.773$$

Thus, in our example of the 10-day yield series, we would calculate the annual volatility as 77.3 basis points (absolute) or 15.57% (relative). The relative volatility times the average yield for the period $0.1557 \times 4.983 = 0.776$ is close to the absolute yield volatility of 0.773—as one would expect.

The second relevant clarification may damage a naïve understanding of volatility as the annual standard deviation. Volatility measures only the pace of uncertainty; this concept does not assume the daily-measured volatility remains constant over time, just as when driving in traffic with starts and stops there is a difference between instantaneous speed and the average velocity. Third, an important assumption for annualizing the daily volatilities is that

the *daily increments* are serially independent. If there is a relationship between rate changes on one day and another day, then we say there is serial correlation. The “square-root rule” will not be an accurate measure of the annual volatility if there is serial correlation in the random process. Figure 1 illustrates that volatility has been volatile.

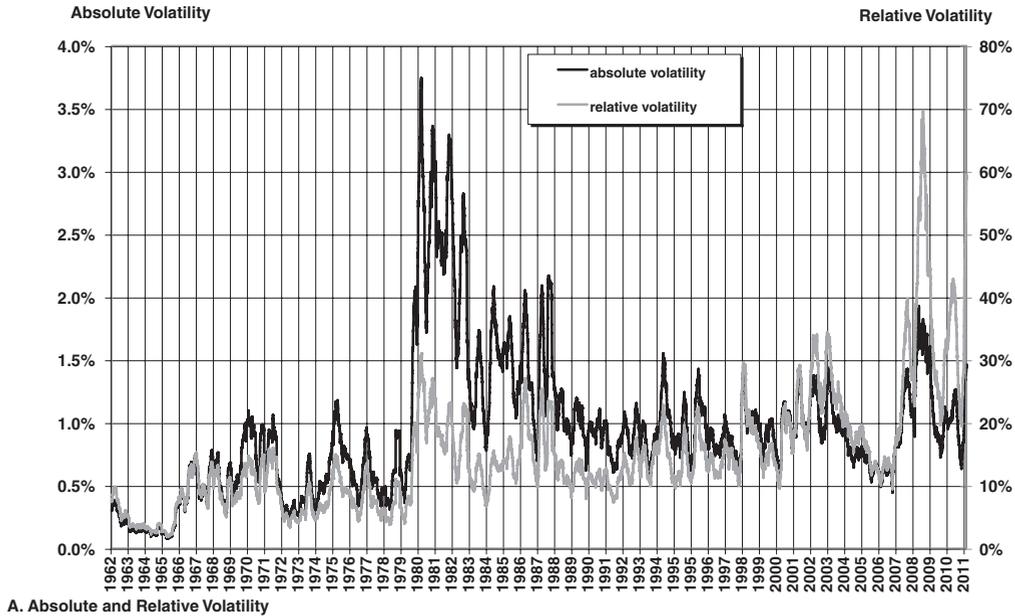
In the 1980s, both volatility measures exhibited instability, although the relative one appeared to be much more stable. However, since the 1990s, the absolute volatility measure has become more stable, oscillating around 1% (100 basis points). Based on these observations, it is not hard to understand why during different time periods, the relative volatility was moving inversely, and the absolute volatility directly, with respect to the rate level. Aside from the explicit level-related effect, both volatility measures seemingly synchronously react to economic disturbances. Pricing in the interest rate options market reflects these important findings.

Different points of the yield curve have differing volatility, too. This observation suggests that not only do the rates have a “term structure,” but their volatility has a term structure as well. A hump shape of such a volatility curve is often observed (see Figure 2). It can be attributed to (1) absence of change in the short rates unless regulators take actions and (2) the dampening force of the mean reversion. We will explain both factors further in the entry.

A DIFFUSIVE MODEL FOR RANDOMNESS

Can we describe the randomness mathematically? It is perhaps simpler than it sounds. In fact, having become acquainted with volatility, we did most of the task. A general diffusive model for an interest rate process that describes how interest rates will vary over time, $r(t)$, will have the following form:

$$dr = (\text{Drift})dt + (\text{Volatility})dz \quad (1)$$



A. Absolute and Relative Volatility



B. Rate Level

Figure 1 History of Volatility for the 10-year Treasury Yield

What does this mean? Notations dr and dt refer to small increments measured over infinitesimally short time. Variable dz represents small changes in $z(t)$ which is called *Brownian motion*, also known as the Wiener process. It is the source of randomness. We cannot control the exact value of this variable. *Drift* and *volatility*

describe how the changes in rates are related to changes in time and the random variable dz . Mathematical model (1) can be thought of in the following way: The change in interest rate over a small time period is the result of a number representing systematic drift times the amount of time change plus a random shock scaled by

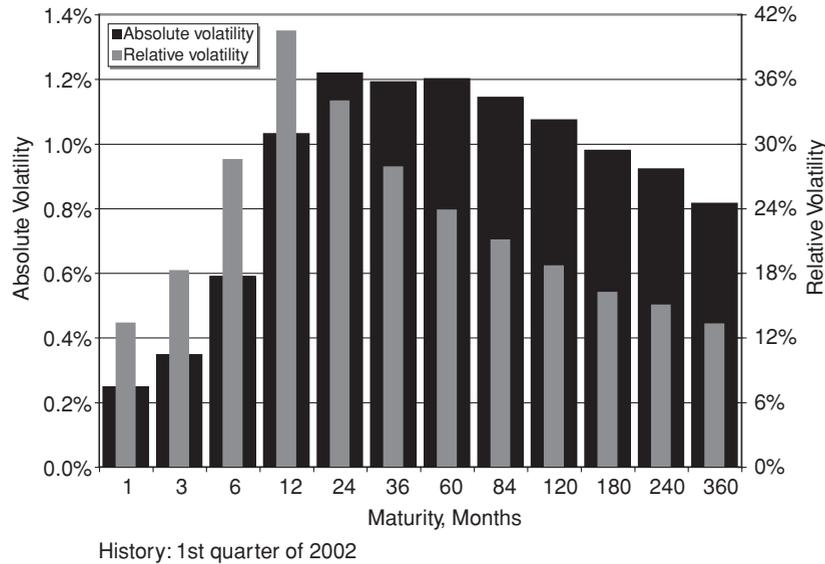


Figure 2 Historical Volatility Term Structure for the Swap Rates

the amount of volatility.

$$\Delta r = (\text{Drift})(\text{Passage of time}) + (\text{Volatility}) \times (\text{Random shock})$$

A Brief Excursion to Brownian Motion

Brownian motion:

- Is continuous
- Is normally distributed
- Has a zero mean (“centered”)
- Has time increments that are serially independent
- Has its own volatility scaled to 1

Therefore, $z(t)$ has a standard deviation of \sqrt{t} (for the same reason that a square root appears in the annualization of daily volatility). Any particular function $z(t)$ is said to be a “realization” or a “sample path” of the Brownian motion. The Brownian motion, therefore, can be thought of as a container of random paths subject to the conditions described above. Figure 3 depicts a sample path and the single- and double-standard deviation zones.

With the use of *volatility* multiples, we can scale the rate process to any volatility level. The *drift* variable simulates a systematic, nonrandom tendency. For example, it can model a central tendency function known as *mean reversion*. Equation (1) is called the *stochastic differential equation*. Both multiples, *drift* and *volatility* do not have to be constants. They can be functions of time, t , and rate, r . Any particular specification of $drift(t,r)$ and $volatility(t,r)$ leads to a specific rate model, but not necessarily a good one. At this stage, it is enough to understand that a good model can be a strong quantitative pricing tool. Although we cannot know what the random variable $z(t)$ is going to do, we, at least, can simulate its behavior with a large number of random scenarios. The Monte Carlo method draws on this idea. On the other hand, we may be able to do some intelligent analytical work making the brute-force simulations unnecessary. We could also make sure that the model is consistent with (“calibrated to”) prices of widely traded interest rate instruments; then we will feel more confident applying it to the exotic options or the market for mortgage-backed securities (MBSs) and asset-backed securities (ABSs).

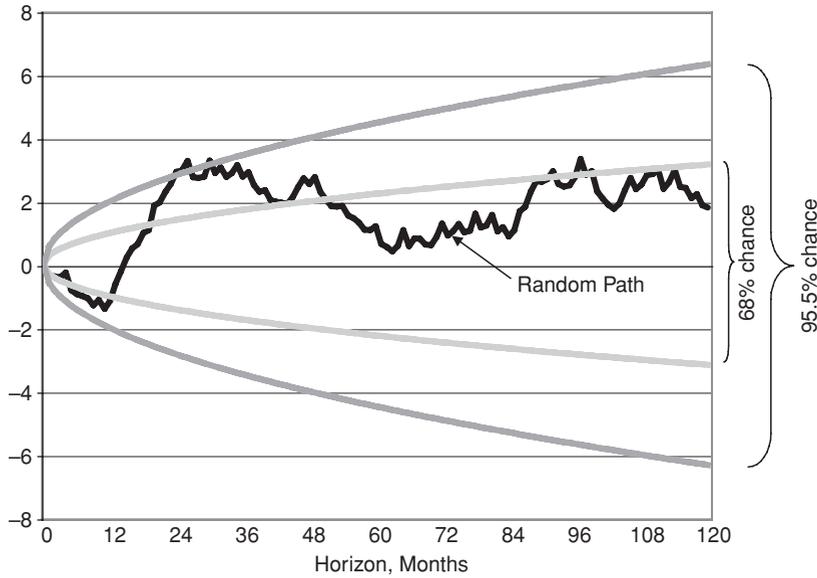


Figure 3 Brownian Motion's Sample Path with Deviation Zones

MEAN REVERSION AND MARKET STABILITY

Consider the following special form of equation (1): $dr = 5r dt + \sigma dz$. Can this equation model an actual interest rate? Since the formula shows that the change in interest rate increases with changes in time, the average rate will continue to grow. Utilizing calculus, we note that the solution to this equation will not only grow with time, but will grow exponentially as it contains an e^{5t} term. Since interest rates cannot increase exponentially forever (at least, they never have), we need to dismiss this formula as inappropriate for the job.

How about $dr = \sigma dz$? The drift is chosen to be zero, and provided that the initial value $r(0)$ is known, the process will randomly evolve around this value, on average. Whether the initial rate is high or low, the model will stay centered around it. The standard deviation, as we already know, will grow as \sqrt{t} . This may not be a very good thing either. A century from now, the magnitude of the standard deviation will be huge, at ten times annual volatility. Figure 1B demonstrated that interest rates tend to stay within a range.

Both models briefly reviewed above suffer with the same disease: They are unstable. Observable objects in economy, finance, engineering, or physics tend to be stable; otherwise, they would not be able to exist long enough to be observed. The feature making financial markets stable is known as mean reversion. It is simply a properly chosen specification of the drift term that would ensure the dampening effect (also known as central tendency). If the rate randomly has grown too high or fallen too low, the drift term will help "return" it back. Here is an example:

$$dr = a(r_{\infty} - r)dt + \sigma dz \tag{2}$$

where mean reversion parameter $a > 0$. This time, the solution will contain e^{-at} , a decaying component that indicates stability. The mean converges to parameter r_{∞} , the long-term equilibrium (now we see the point for this strange notation). The standard deviation will grow with time as

$$\sigma \sqrt{(1 - e^{-2at})/2a}$$

and converges to

$$\sigma/\sqrt{2a}$$

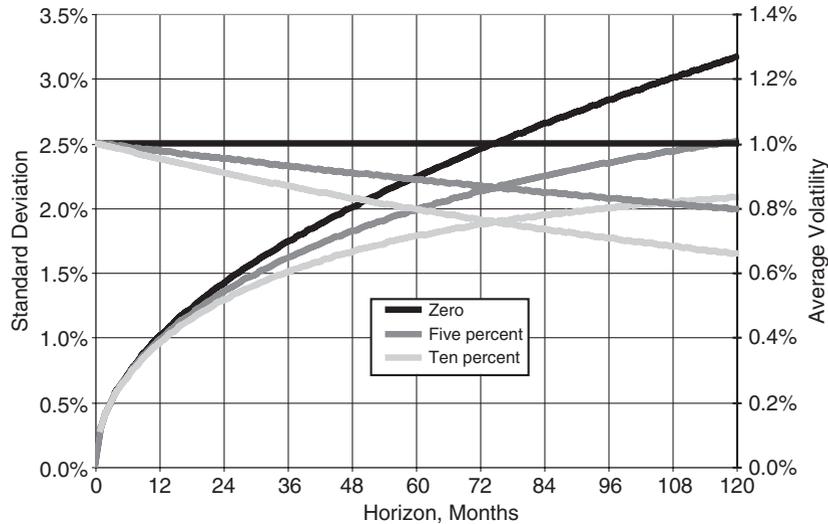


Figure 4 Standard Deviation and Average Volatility for Different Values of a

as the horizon extends. Figure 4 compares the standard deviation (lines launching from the origin) and equivalent (“average”) volatility for different levels of mean reversion including the zero one ($\sigma = 1\%$ was assumed).

Mean reversion, therefore, stabilizes the market. It also explains why volatility is typically measured on a daily basis; in the presence of mean reversion, the average volatility measured over a time horizon generally depends on this horizon. For example, if $a = 10\%$, only 95% of actual volatility is seen in annual increments.

If $r(t)$ in (2) is the short market rate, then every other rate (the 5-year, the 10-year, etc.) can be derived as a function of it. In particular, for mean-reverting models, volatility of *long rates* should eventually fade with maturity of the rate, and it does happen as seen in Figure 2. Mathematically, it is a direct consequence of the mean reversion: The *short rate's* uncertainty gets limited going forward, thereby making long-term bonds less volatile now. Economically, discount rates for very remote cash flows should be almost certain; otherwise their present values would be infinitely risky.

Does it seem that model (2) makes sense? Well, Vasicek (1977) noticed it, as one of the first interest rate models. It is been popular and

important since and was a basis for many of the models that are used today.

THE RATE DISTRIBUTION

Equation (2) is a linear stochastic differential equation disturbed by a Brownian motion. The math tells us that the output of this equation, rate $r(t)$, is going to be normally distributed. Although it makes the model tractable, the negative rates are not precluded, which may or may not be a problem. Arguably, the actual rates should stay positive—at least, they almost always have been. When using process (2), odds of negative rates grow with future time, as the present value falls. In addition, mortgages and related securities are amortizing and may have small balances and cash flows years from now. Levin (2004) provided a quantitative support for the use of normal distribution by showing it does not distort options’ value materially.

The fact that a Brownian motion is normally distributed does not require that the rate process be such. For example, considering exponential transformation $R = \exp(r)$ and using R as the rate, rather than r , we ensure that the rate

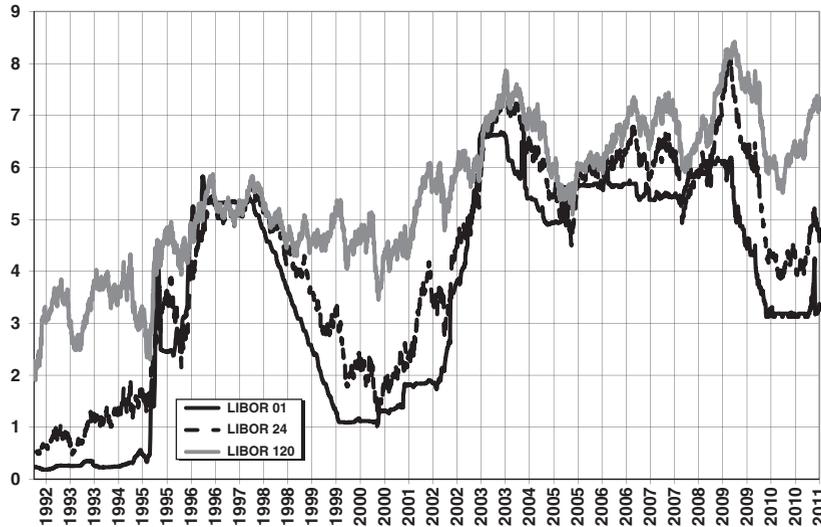


Figure 5 Jumpy and Continuous Interest Rates

remains positive. Such a process is said to have lognormal distribution. The mean and standard deviation for this known distribution is explicitly stated through the mean μ and the standard deviation σ of the original variable, $r(t)$:

$$E(R) = \exp\left(\mu + \frac{1}{2}\sigma^2\right)$$

$$std(R) = \exp\left(\mu + \frac{1}{2}\sigma^2\right)\sqrt{\exp(\sigma^2) - 1}$$

Another popular example is the squared transformation, $R = r^2$, that also guarantees that rate R stays positive; the distribution of such defined rate is known as noncentral χ^2 . For the squared transformation,

$$E(R) = \mu^2 + \sigma^2$$

$$std(R) = \sigma\sqrt{2\sigma^2 + 4\mu^2}$$

INTEREST RATE JUMPS

Stochastic differential equation (1) is not the most general mathematical form of a random process. It applies only to *diffusions*, that is, continuous random processes. Stochastic calculus considers many other forms of randomness; an important one is called random *jumps* or random events. To appreciate this type of random-

ness, let us consider the history of three rates, the 1-month London Interbank Offered Rate (LIBOR) (“LIBOR 01”), the 2-year swap (“LIBOR 24”), and the 10-year swap rate (“LIBOR 120”), depicted in Figure 5.

Both swap rates change almost continuously, day after day, and little by little, at times randomly oscillating, in response to the market forces. The 1-month rate was changing in a suspiciously smooth fashion between sudden jumps featuring apparent and prolonged plateaus that are not seen for the swap rates. For example, in 2002 to 2004, it had been barely changing for a while, and then plunged responding to the Fed’s actions. Furthermore, whereas the visual dynamics for all three rates seem to resemble one another, statistical measurements of correlation between daily increments overwhelmingly reject such a conclusion. For this 18-year-long history (over 4,500 observations) we computed a small 7% correlation between daily increments of the 1-month and the 2-year rates, and an even smaller 4% correlation between 1-month and 10-year rates. What if we measure correlations between increments in monthly averages (216 observations), thereby filtering out the disparity between daily dynamics? Then the 7% goes way up to 46% and

Table 2 Empirical Correlations between Periodic Increments, 1992–2010 History

	Daily	Monthly	Quarterly	Semiannually	Annually
2-year to 1-month	7%	46	65	81	90
10-year to 1-month	4	24	40	51	71
Total observations	4,527	216	71	35	17

the 4% to 24%, and the interrater correlations continue to improve steadily as we extend the averaging period (Table 2).

These objective facts suggest that a stochastic diffusive model suitable for swap rates may be not perfectly appropriate for short rates. A random jump component may be necessary to explain the actual short-rate behavior and associated option pricing. One popular mathematical form of jumps is the *Poisson* process. It is simply a random occurrence of events described by a single parameter λ called frequency or intensity. The average number of events to occur during a time interval of t is equal to λt ; curiously enough the variance of this number is equal to λt too. Probability that we will have exactly j events during this time interval is equal to $e^{-\lambda t}(\lambda t)^j/j!$. It is only the period's length that really matters, not when the period starts—for this reason, the Poisson process can't be used to describe, say, human deaths or bulb failures when the attained age is a strong factor. However, it is plausible to assume that the Poisson jumps describe some events in financial markets.

Aside from the jumps' arrival, the size of jumps can be also random. Merton (1976) introduced an option-pricing model when the underlying process includes Poisson jumps with normally distributed magnitude. Using mathematical notations, we can express the model as

$$dr = (\text{Drift})dt + (\text{Volatility})dz + (\text{Jump Volatility})dN \quad (3)$$

where N is the Poisson-Merton jump variable. When jump occurs, dN is drawn from the standard normal distribution $N[0,1]$; it stays 0 oth-

erwise. In a less strict notations,

$$\begin{aligned} \Delta r = & (\text{Drift})(\text{Passage of time}) \\ & + (\text{Volatility})(\text{Random shock}) \\ & + (\text{Jump volatility})(\text{Random jump}) \end{aligned}$$

The practical difference between random shock and random jump is that, for a small time interval, the former is small, but nonzero, whereas the latter is mostly zero and rarely finite. Hence, equation (3) describes a more general stochastic process combining diffusion and jumps ("jump-diffusion"). Notably, mathematical variance of the Poisson process $N(t)$ is too proportional to the time horizon t . This fact allows aligning interpretations of $\sigma_d \equiv \text{Volatility}$ and $\sigma_j \equiv \text{Jump Volatility}$: for very small t , the standard deviation of $r(t)$ is equal to

$$t\sqrt{\sigma_d^2 + \lambda\sigma_j^2}$$

meaning that the mixed volatility will be simply

$$\sigma = \sqrt{\sigma_d^2 + \lambda\sigma_j^2}$$

Furthermore, if we generalize the linear mean-reverting Vasicek model given by (2) by adding a jump term, then expressions for the mean and the standard deviation of $r(t)$ won't change; it will be enough to replace σ .

At first, it is tempting to interpret a jump-diffusion process as diffusion with another volatility scale. In reality, probability distributions of these two stochastic patterns are different. Inclusion of jumps "fattens" the distribution's tail (see Table 3) and is much more suitable for modeling and pricing rare events like a corporation's defaults or credit downgrade, a financial crisis, reaching a very remote option's strike, or change in the short rate over a fairly short horizon.

Table 3 Comparison of Distribution's Tails for Poisson-Merton Jump Processes

Value of λt	$\text{Prob}(r < \mu - 4\sigma)$	$\text{Prob}(r < \mu - 3\sigma)$	$\text{Prob}(r < \mu - 2\sigma)$	$\text{Prob}(r < \mu - 1\sigma)$
0.2	0.789%	1.777%	3.505%	6.022%
1	0.158	0.753	3.357	12.568
5	0.027	0.303	2.559	14.732
Infinite (normal)	0.003	0.135	2.275	15.866

Models with Poisson-Merton processes converge to normal when the value of λt is large (frequent jumps are similar to diffusion), but may produce significantly different results when it is small.

Stochastic differential equations (1) and (3) can be viewed as building blocks for the interest rate modeling. Some models used today in the financial industry are multifactor with the short rate $r(t)$ defined not as the solution to equations (1) or (3), but as their sum. When modeling LIBOR, neither the jump arrivals have to be Poissonian, nor the magnitude has to be normal. For example, Chan et al. (2003) developed a model with rate jumps timed to periodic Fed meetings, and the magnitude being a random multiple of 25 bps. There exist other modeling views at interest rate dynamics that we don't cover in this entry including continuous randomness with stochastic volatility levels; see James and Webber (2000) for a comprehensive overview.

KEY POINTS

- The most common way of simulating interest rates' uncertainty is employing stochastic differential equations containing drift and volatility terms.
- In older times, absolute volatility was directionally related to the rate's level; this relationship has gradually disappeared.

- The drift term must contain a mean reversion, that is, stabilization force.
- Actual short rates (LIBOR) have been historically jumpy and require adding random jumps to diffusions.

ACKNOWLEDGMENTS

The author wishes to thank Andrew Davidson for his help in shaping this material, Will Searle for his comments, and Nancy Davidson for her editorial work.

REFERENCES

- Chan, Y.-K., Bhattacharjee, R., Russell, R., and Teytel, M. (2003). *A New Term-Structure Model Based on Federal Funds Target*. Citigroup, Mortgage Securities.
- James, J. and Webber, N. (2000). *Interest Rate Modeling: Financial Engineering*. New York: John Wiley & Sons.
- Levin, A. (2004). Interest rate model selection. *Journal of Portfolio Management* 30, 2: 74–86.
- Levin, A. (2001). A tractable skew-and-smile model. The Dime Bancorp, Working Paper.
- Merton, R. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3: 125–144.
- Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics* 5: 177–188.

Short-Rate Term Structure Models

ALEXANDER LEVIN, PhD

Director, Financial Engineering, Andrew Davidson & Co., Inc.

Abstract: Market randomness makes the fair value of a financial instrument an expectation. It also requires a rigorous quantification of the dynamics of interest rates; that is, a well-defined interest rate model. Prices of interest rate options and options embedded in bonds such as corporate or agency callable debts, mortgage-backed securities, and asset-backed securities will firmly depend on this modeling work. Contemporary interest rate models employ the available information about currently observed forward rates and vanilla European options and are “calibrated” to them. The relationships between bond rates should preclude arbitrage. Some analytically tractable models ensure these properties explicitly. Selecting the “best” term structure model is becoming more a conscientious task and less a matter of taste. Measuring “volatility skew” for widely traded swaptions is a simple technique that yields rich results. Another method is computing volatility indexes produced by different models and tracking their stability. Recent trading history confirms normalization of the swaption market making the Hull-White model, the extended Cox-Ingersoll-Ross model, or the squared Gaussian model more attractive than formerly popular lognormal models. Single-factor models cannot value accurately curve options or some exotic derivatives that are exposed to the yield curve shape and require multifactor modeling work. The affine theory offers a systematic method of constructing such models. It also allows for jump-diffusion extensions that may be necessary to explain volatility smile; that is, an excessive convexity of the Black volatility as a function of strike.

This entry introduces a family of models for stochastic behavior of interest rates and the principles of their design widely used by market participants.

THE CONCEPT OF SHORT-RATE MODELING

Why do we call interest rate models term-structure models? Aren't there too many rates for one model? The tree-based valuation ex-

amples found in many books and research papers show us that we can value an any-maturity bond and thereby reconstruct the entire term structure using only dynamics of one-period rate (see, for example, Davidson et al., 2003 [Chapter 12], and Fabozzi, 1994). Interest rate models operating with the short (one-period) rate $r(t)$ as their main object are commonly referred to as “short-rate models.” They are different by construction from so-called “forward rate models,” such as the Heath-Jarrow-Morton model (Heath, Jarrow,

and Morton, 1992) or the Brace-Gatarek-Musiela model (Brace, Gatarek, and Musiela, 1997). Both types of interest rate modeling are designed to solve the same problems and are widely used for valuation of fixed income options and embedded option bonds, but operate with different objects. Unlike the short-rate modeling family, forward rate models employ and randomly evolve the entire forward curve of the short rate, $f(t, T)$, in which the t is time and T is the forward time, to which the short rate applies.

We restrict our attention solely to the short-rate modeling. This term does not assume that any short-rate term structure model is a one-factor model or depends only on the short rate.

The Arbitrage-Free Interrate Relationship

Let us assume that we have a stochastic process, possibly multifactor, describing the short rate dynamics $r(t)$. Let us denote $P_T(t)$ to be the market price observed at time t of a T -maturity zero-coupon bond; that is, a bond paying \$1 at $t + T$. This price is exponential to the yield to maturity ("rate") $r_T(t)$ of this bond: $P_T(t) = \exp[-r_T(t)T]$. However, we can use the arbitrage argument claiming that, once prices of instruments reflect rate expectations and risks, there should exist no advantage or disadvantage in investing in the zero-coupon bond over continuous reinvesting into the short rate. Hence, the same price should be equal to

$$P_T(t) = E \left[\exp \left[- \int_t^{t+T} r(\tau) d\tau \right] \right]$$

where E denotes the arbitrage-free expectation.

Equating these two expressions, we get

$$r_T(t) = -\frac{1}{T} \text{Ln} E \left[\exp \left[- \int_t^{t+T} r(\tau) d\tau \right] \right] \quad (1)$$

Formula (1) allows us to compute any-maturity zero-coupon rates via some expectation involving random behavior of the short rate. Of course, once we establish the entire zero-coupon curve, we can restore a yield for any other bond including a coupon-paying one. To compute the expectation in (1), we must know two things: stochastic equation (or equations) for $r(\tau)$ and initial (time t) conditions. The latter represents public information about the market at time t and includes every factor affecting the short rate. Therefore, it would be correct to state that an any-maturity rate can be recovered using only factors that determine the evolution of the short rate. In particular, if only one Brownian motion drives the short rate dynamics, it will define the entire yield curve as well.

Consistency with the Initial Yield Curve

Let us apply the interrater relationship (1) to the initial point of time, $t = 0$:

$$r_T(0) = -\frac{1}{T} \text{Ln} E \left[\exp \left[- \int_0^T r(\tau) d\tau \right] \right] \quad (2)$$

The left-hand side of this formula is known from today's term structure of interest rates. Hence, the short rate dynamics $r(t)$ must be such as to ensure (2) holds. In practical terms, adjusting a rate process to fit the initial yield curve is part of a more general task often termed "calibration." Without this necessary step, an interest rate model can't be used to value even simple, option-free bonds. Computation of expectation in formulas (1) and (2) can be done numerically or, in some models, analytically.

Consistency with European Option Values

If a term structure model is built to value complex derivative instruments, it must value, at minimum, simple European options. Suppose

we have an option that is exercised at a future point of time t and generates a cash flow that we denote $g[r(t)]$; that is, some nonlinear function of the short rate observed at t . Note that the actual option's exercise may be triggered by a long, rather than the short, rate; nevertheless, it will depend either on $r(t)$ (single-factor models) or all market factors (*multifactor models*) known at t . The value of the option is going to be

$$\text{option} = E \left[g[r(t)] \exp \left[- \int_0^t r(\tau) d\tau \right] \right] \quad (3)$$

where E denotes the same expectation as before.

We may now demand that the short rate process $r(t)$ produces options values (3) that match market prices. Most commonly, term structure models are calibrated to LIBOR caps, or European options on swaps (swaptions), or both. These are standard, widely traded European options. For example, a call option on a T -maturity swap will generate cash flow equal to $g[r(t)] = A_T(t)[K - c_T(t)]^+$ where A denotes annuity, c denotes the swap rate, both measured at t , and superscript “+” indicates that only a positive value is taken. Another standard derivative is the LIBOR cap made of “caplets,” that is, European calls on some relatively short rate. A T -maturity LIBOR caplet ($T = 3$ months for standard caps) expiring at t pays $[r_T(t) - K]^+$ at $t + T$. To recognize the time difference T between the caplet's expiry and the actual pay, we can move the payoff from $t + T$ to t and express it as $g[r(t)] = [r_T(t) - K]^+ / (1 + Tr_T(t))$. We then have to make sure that formula (3) yields correct values for the caplets. Note that the cap market does not usually quote caplets directly; however, their values can be assessed by bootstrapping.

SINGLE-FACTOR SHORT-RATE MODELS

In this section, we describe several different single-factor models, which employ the

short rate as the only factor. We also give some evidence on the relative performance of the models. For each of the models, we emphasize three key aspects: the model's formulation, its arbitrage-free calibration, and the interrate relationship that recovers the entire term structure contingent on the dynamics of the short rate.

The Hull-White/Vasicek Model

The Hull-White (HW) model (Hull and White, 1994) describes the dynamics of the short rate $r(t)$ in the form given by

$$dr = a(t)(\theta(t) - r)dt + \sigma(t)dz \quad (4)$$

Here, $a(t)$ denotes *mean reversion*, $\sigma(t)$ stands for volatility; both can be time-dependent. Function $\theta(t)$ is sometimes referred to as “arbitrage-free” drift. This terminology is caused by the fact that, by selecting proper $\theta(t)$, we can match any observed yield curve. The HW model was preceded by the Vasicek model having $\theta(t) = 0$. The short rate is normally distributed in this model, so the volatility represents absolute rather than relative changes.

This can be seen mathematically as (4) is a linear equation disturbed by the Brownian motion (a normally distributed variable); the short rate is normally distributed as well. Therefore, its integral is normally distributed too, and the expectation found in the right-hand side of formulas (1), (2), and, in some cases, (3) can be computed in a closed form. Without going through the math we provide here the analytical *calibration* results to the observed short forward curve $f(t)$ for the constant-parameter case:

$$\theta(t) = f(t) + \frac{1}{a} \frac{df(t)}{dt} + \frac{\sigma^2}{2a^2} (1 - e^{-2at}) \quad (5)$$

The short rate's expectation is found as

$$E[r(t)] = f(t) + \frac{\sigma^2}{2a^2} (1 - e^{-at})^2 \quad (6)$$

The last term in (6) is called the convexity adjustment; that is, the difference between mathematically expected short rates in the future and

the forward short rates. This adjustment is proportional to volatility squared; for zero mean reversion, it is simply equal to $\frac{1}{2}\sigma^2t$. It is therefore up to financial engineers to make sure the convexity adjustment is properly implemented in a pricing system; it is very volatility sensitive.

The expected value for any long, T -maturity, zero-coupon rate is proven to be in the same form: forward rate + convexity adjustment. This time, the exact formula for this relation is

$$E[r_T(t)] = f_T(t) + \frac{\sigma^2}{4a^3T}(1 - e^{-aT})[2(1 - e^{-at})^2 + (1 - e^{-2at})(1 - e^{-aT})] \quad (7)$$

Any long zero-coupon rate is normally distributed too and proven to be linear in the short rate; deviations from their respective mean levels are related as

$$\frac{\Delta r_T}{\Delta r} \equiv \frac{r_T(t) - E[r_T(t)]}{r(t) - E[r(t)]} = \frac{1 - e^{-aT}}{aT} \equiv B_T \quad (8)$$

The function B_T of maturity T plays an important role in the HW model. It helps, for example, to link the *short-rate volatility* to the long-rate one and explicitly calibrate it to the market. If $a = 0$, this function becomes identical to 1, regardless the maturity T . This important special case allows for a pure parallel change in the entire curve (every point moves by the same amount). This particular specification can be suitable for standardized risk measurement tests.

The HW model is a very tractable arbitrage-free model, which allows for the use of analytical solutions as well as Monte Carlo simulation. The volatility σ and mean reversion a can be analytically calibrated to European options on zero-coupon bonds. Most commonly, the HW model is calibrated to either a set of short-rate options (LIBOR caps) or swaptions. In the later case, very good approximations can be constructed (see Levin, 2001; Musiela and Rutkowski, 2000). The model's chief drawback is that it produces negative interest rates. However, with mean reversion, the effect of negative

rates is reduced. The rate history of the 1990s and 2000s supports this type of formulation of a term structure model.

The Cox-Ingersoll-Ross Model

The Cox-Ingersoll-Ross model (CIR model) is a unique example of a model supported by the general equilibrium arguments (see Cox, Ingersoll, and Ross, 1985). CIR argued that the fixed income investment opportunities should not be dominated by neither expected return (the rate), nor the risk. The latter was associated with the return variance, thus suggesting that volatility-squared should be of the same magnitude as the rate:

$$dr = a(t)(\theta(t) - r)dt + \sigma(t)\sqrt{r} dz \quad (9)$$

Equation (9) is actually a no-arbitrage extension to the "original CIR" that allows fitting the initial rate and volatility curves. Since the volatility term is proportional to the square root of the short rate, the latter is meant to remain positive. The extended CIR model is analytically tractable, but to a lesser extent than the HW model. Perhaps the most important result of CIR is that the long zero-coupon rates are also proven linear in the short rate—in line with (8). However, the slope function has now a quite different form; it depends on both maturity T and time t and is found as $B_T(t) = -b(t, t + T)/T$. Function $b(t, T)$ used in this expression solves a Ricatti-type differential equation, considered for any fixed maturity T :

$$\frac{db(t, T)}{dt} = a(t)b(t, T) - \frac{1}{2}\sigma^2(t)b^2(t, T) + 1 \quad (10)$$

subject to terminal condition $b(T, T) = 0$.

If the mean reversion a and "CIR volatility" σ are constant (the "original CIR"), equation (10) allows for an explicit solution. In this case, $b(t, T)$ is a function of $T - t$ only, and B_T is appeared to be time-independent:

$$B_T = \frac{2(e^{\gamma T} - 1)}{(\gamma T + aT)(e^{\gamma T} - 1) + 2\gamma T} \quad (11)$$

where $\gamma = \sqrt{a^2 + 2\sigma^2}$.

Without a mean reversion, this formula reduces to a more concise

$$B_T = \frac{\tanh(\gamma T/2)}{(\gamma T/2)}$$

Note that this ratio is always less than 1. This means that the long rates are less volatile than the short one, even without a mean reversion. This is in contrast to the HW model where, with $a = 0$, the yield curve would experience a strictly parallel reaction to a short rate shock.

Generally speaking, calibration to the currently observed short forward curve $f(T)$ cannot be done as elegantly and explicitly as in the HW model. Once the $b(t, T)$ function is found, the calibrating function $\theta(t)$ satisfies an integral equation:

$$-f(T) = \int_0^T \frac{db(t, T)}{dT} \theta(t) a(t) dt + \frac{db(0, T)}{dT} r_0 \tag{12}$$

Numerical methods, well developed for integral equations, should be employed.

It is established that all zero-coupon rates, under the CIR model, have noncentral χ^2 distributions and remain positive. Economic rationale, nonnegative rates, and analytic tractability have made the CIR model deservedly popular; it is one of the most attractive and useful interest rate models. It is also consistent with the Japanese market and some periods of the U.S. rate history when rates were very low.

The Squared Gaussian Model

To describe the squared Gaussian model (SqG model, and also known as the quadratic model), we employ a linear differential equation (4) only to define an auxiliary variable $x(t)$; we then define the short rate in a form of its square:

$$\begin{aligned} dx &= -a(t)xdt + \sigma(t)dz \\ r(t) &= [R(t) + x(t)]^2 \end{aligned} \tag{13}$$

For convenience, we removed previously used *arbitrage-free* function $\theta(t)$ from the first

equation and introduced a deterministic calibrating function $R(t)$ to the second equation serving the same purpose. Note that we could have introduced the HW model similarly by defining the short rate as $r(t) = R(t) + x(t)$. Ito's lemma allows us to convert model (13) to a single stochastic differential equation for the short rate:

$$\begin{aligned} dr &= [2R'\sqrt{r} - 2a(r - R\sqrt{r}) + \sigma^2]dt \\ &\quad + 2\sigma\sqrt{r} dz \end{aligned} \tag{14}$$

where R' stands for dR/dt . The SqG model has an apparent similarity to the CIR model in that its volatility term is proportional to the square root of the short rate, too. However, comparing stochastic equations (14) and (9) we see that they have different drift terms.

The SqG model has been studied by Beaglehole and Tenney (1991), Jamshidian (1996), and Pelsser (1997), among others. The most notable fact established for the SqG model is that any zero-coupon rate $r_T(t)$ is quadratic in $x(t)$ that is linear in the short rate $r(t)$ and its square root $\sqrt{r(t)}$:

$$\begin{aligned} (T - t)r_T(t) &= A(t, T) - B(t, T)\sqrt{r(t)} \\ &\quad - C(t, T)r(t) \end{aligned} \tag{15}$$

Functions A , B , and C satisfy a system of ordinary differential equations:

$$A' = BR' + \sigma^2(\frac{1}{2}B^2 + C) + aRB \tag{16a}$$

with $A(T, T) = 0$

$$B' = aB - 2CR' - 2aCR - 2\sigma^2BC \tag{16b}$$

with $B(T, T) = 0$

$$C' = 1 + 2aC - 2\sigma^2C^2 \tag{16c}$$

with $C(T, T) = 0$

where, for brevity, A' and the like denote derivatives with respect to time t and the dependence of all functions on t and T is omitted. Note that all the terminal conditions are set to zero. Indeed, once t is equal to T , both sides of the relationship (15) must become zero for any value of r ; this is possible if and only if functions A , B , and C turn to zero. Much like

in the CIR model, equation (16c) for the linear term's slope, this time denoted via C , is of a Riccati type (see Boyle, Tian, and Guan, 2002) and can be solved in a closed-end form. In fact, it is identical to already solved equation (10) except it operates with a doubled mean reversion and a doubled volatility. Other equations in (16) and calibration to the initial yield curve can be solved numerically.

The short rate has a noncentral χ^2 distribution with 1 degree of freedom. Long rates are mixtures of normal and χ^2 deviates. Like the CIR model, the SqG model ensures positive rates; the square-root specification of volatility is suitable for many options. Due to some analytical tractability and known form for long rates, the volatility function and mean reversion can be quite accurately calibrated to traded options.

The Black-Karasinski Model

Once a very popular model, the Black-Karasinski model (BK model) expresses the short rate as $r(t) = R(t)\exp[x(t)]$, where, as in the previous case, random process $x(t)$ is normally distributed (see Black and Karasinski, 1991). The short rate is, therefore, lognormally distributed. Assuming the same process for $x(t)$ we can write the stochastic differential equation for the short rate as

$$dr = r \left(\frac{R'}{R} + \frac{1}{2}\sigma^2 - a \ln \frac{r}{R} \right) dt + r\sigma dz \quad (17)$$

The rate's absolute volatility is therefore proportional to the rate's level. Although the entire short-rate distribution is known (including the mean and variance), no closed-form pricing solution is available. This is because the cumulative discount rate, the integral of r , has an unknown distribution. Traditionally, the BK model is implemented on a tree. Calibration to the yield curve and volatility curve can be done using purely numeric procedures. For example, one could iterate to find $R(t)$ period-by-period until all the coupon bonds or zero-coupon bonds (used as input) are priced

exactly. Alternatively, one could find approximate formulas and build a faster, but approximate scheme.

Despite its past popularity, the BK model's main assumption, the rate's *lognormality*, is not supported by the recent rate history. The volatility parameter σ entering the BK model is not the same as the *Black volatility* typically quoted for swaptions or LIBOR caps. For example, selecting $\sigma = 0.15$, $a = 0$ does not ensure 15% volatility even for European options on short rates (caplets). Hence, calibration of the model to volatilities found in the option market is not an easy task.

The Flesaker-Hughston Model

The Flesaker-Hughston model (FH) is an interesting model because it is different from all previously described ones in that it allows for computing the coupon rates analytically (see Flesaker and Hughston, 1996). The model starts with defining a random process $M(t)$, which is any martingale starting from 1, and two deterministic positive functions $A(t)$ and $B(t)$, decreasing with time t . Then, at any point of time t , a zero-coupon bond maturing at T has its price in a rational functional form of $M(t)$:

$$P(t, T) = \frac{A(T) + B(T)M(t)}{A(t) + B(t)M(t)} \quad (18)$$

Taking the natural logarithm of this expression, changing the sign, and dividing it by $T - t$ gives us, of course, the zero-coupon rate. In order to derive a coupon rate $c(t, T)$, let us recall that a coupon-bearing bond generates periodic payments at a rate of c and returns the principal amount (\$1) at maturity. Let us denote the time- t value of this bond as $P^c(t, T)$:

$$P^c(t, T) = \sum_{i=1}^n cP(t, t_i) + P(t, T)$$

where t_i are the timings of coupon payments, with $t_n = T$. To express the par coupon rate c , let us equate this $P^c(t, T)$ to 1 and substitute

postulated expression (18) for all discount factors:

$$c(t, T) = \frac{A(t) - A(T) + [B(t) - B(T)]M(t)}{\sum_{i=1}^n [A(t_i) + B(t_i)M(t)]}$$

$$r(t) = -\frac{A'(t) + B'(t)M(t)}{A(t) + B(t)M(t)} \quad (19)$$

Hence, all coupon rates and the short rate are too rational functions of $M(t)$. If we select a positive martingale process $M(t)$; for example, a lognormal one, $dM = \sigma M dz$, then all rates will stay positive. Functions $A(t)$ and $B(t)$ can fit the initial term structure of rates and volatilities. (See Flesaker and Hughston, 1996, or James and Webber, 2000, for additional details.)

Other Single-Factor Models

There exists a fair amount of “named” models not mentioned in this entry thus far. They differ in specifications of drift and volatility functions. They include the Ho-Lee model, the Black-Derman-Toy model, and the Brennan-Schwartz model. We will briefly review some of them.

A predecessor to the HW model, the Ho-Lee model (HL model) was offered as a discrete-time, arbitrage-free, model (see Ho and Lee, 1986). Its continuous version is equivalent to the HW model with zero mean reversion. Hence, all analytical statements made for the HW model are valid for the HL model.

The Black-Derman-Toy model (BDT model) is a lognormal short-rate model with endogenously defined mean reversion term equal to $\sigma'(t)/\sigma(t)$ (see Black, Derman, and Toy, 1990). This specification means that a constant volatility leads to a zero mean reversion; a growing *short-rate volatility* function $\sigma(t)$ causes a negative mean reversion, thereby destabilizing the process. Once very popular in financial industry, BDT was replaced by the BK model; both of these models are now recognized as outdated.

The Brennan-Schwartz model is a proportional volatility, mean-reverting, short-rate model (see Brennan and Schwartz, 1979). Introduced in 1979 as an equilibrium model, it has some similarity in its volatility specification

to lognormal models; however, rates are not lognormally distributed.

Calibration Issues

The Vasicek model and the original Cox-Ingersoll-Ross model laid the foundation of term structure modeling. Despite their unquestionable historical importance, traders almost never employ them today. The reason is fairly simple: Built with constant parameters, these models can't be calibrated to the market accurately enough. The extensions, known as the Hull-White (“extended Vasicek”) model and the extended CIR model, allow for selecting time-dependent functions $a(t)$, $\sigma(t)$, and $\theta(t)$ so that the model produces exact or very close prices for a large set of widely traded fixed income instruments, ranging from option-free bonds (or swaps) to European (“vanilla”) options on them and more. In particular, function $\theta(t)$ [or $R(t)$] is normally selected to fit the entire option-free yield curve as formula (5) demonstrates. In contrast, functions $a(t)$, $\sigma(t)$ are usually found to match prices of European options. For example, using just a pair of constants (a, σ) one can match exactly prices of two options, for example, a 1-year swaption on the 2-year swap and 10-year swap. Clearly, we can match many more expiration points if we make $a(t)$, $\sigma(t)$ time dependent. In some systems, volatility function is allowed to be time dependent, but mean reversion remains a positive constant. This way, one can fit options' expiration curve only on average, but the model remains stable and robust. Note that a negative mean reversion may destabilize the dynamic process.

As we pointed out, single-factor models possess various degrees of *analytical tractability*. When using the HW model, a large portion of calibration work can be done analytically—starting from formula (5). The CIR model and the SqG model are somewhat analytical, but, practically speaking, require numerical solutions to ordinary differential equations. The BK model has no known solution

at all. A lack of analytical tractability doesn't preclude using numerical methods or efficient analytical approximations that are beyond this entry.

Single-factor models can't be calibrated to all market instruments. For example, each of the models we have considered thus far creates certain dependence of a European option's value (hence the implied volatility) on an option's strike known as *volatility skew*. Once a model is selected, luckily or not (see the next section), the skew implied by it cannot be changed by the model's parameters. Another problem is that all rates are perfectly correlated in any single-factor model. Hence, none of them can replicate values of "spread options" or "curve options," that is, special derivatives that are exercised when the yield curve flattens or steepens. The solution may lie in using multifactor models as discussed further in this entry.

WHICH MODEL IS BETTER?

The HW model, the CIR model, the SqG model, and the BK model are special cases of a more general class of "CEV models" introduced in 1980s:

$$dr = (\text{Drift})dt + \sigma r^\gamma dz \quad (20)$$

Parameter γ is called constant elasticity of variance (CEV). For $\gamma = 0$ we may have the HW model; for $\gamma = 0.5$, the CIR model or the SqG model; for $\gamma = 1$, the BK model. There exist no specific economic arguments supporting the r^γ functional form for volatility. Often, the CEV constant lies between 0 and 1, but it is not necessary.

Measuring Volatility Skew

Blyth and Uglum (1999) linked the CEV constant to the volatility skew; that is, its dependence of the *Black volatility* (also called implied volatility) on the option's strike, found in the swaption market. They argue that market participants should track the Black volatility ac-

ording to the following simple formula:

$$\frac{\sigma_K}{\sigma_F} \approx \left(\frac{F}{K} \right)^{\frac{1-\gamma}{2}} \quad (21)$$

where σ_K is the Black volatility for the option struck at K , σ_F is the Black volatility for the "at-the-money" option struck at today's forward rate, F . Importantly, one can recover the best CEV constant to use in the model by simply measuring the observed skew.

The skew measured for the 5-year option on the 10-year swap quoted for the period of 1998 to 2004 suggests $\gamma = 0.14$ being optimal, on average. This means that the most suitable model lies between the HW model and the CIR/SqG model (Figure 1). It is also seen that low-struck options are traded with a close-to-normal volatility, while high-struck options are traded with a square-root volatility profile. This fact may be a combination of the "smile" effect discussed at the end of this entry and the broker commission demand. As shown a little further, the square-root volatility specification becomes very suitable in a low-rate environment.

The most recent tendency has been clearly toward $\gamma = 0$, that is, *normality* (Figure 2), thereby making the HW model the best single-factor model choice currently. Note that neither the rate history of the 20-year period from 1991 to 2010, nor the available swaption volatility skew data support lognormality, although earlier rate history did appear to support $\gamma > 1$.

Using the Volatility Index

To compare rate models, it is useful to design a market *volatility index*—a single number reflecting the overall level of option volatility deemed relevant to the interest rate market. Levin (2004) describes a method of constructing such an index by first designating a family of at-the-money (ATM) swaptions ("surface"); that is, options on swaps struck exactly at current forward rate. Then, assuming zero *mean reversion*, one can optimize for the single short-rate volatility constant σ (*volatility index*) best matching the swaptions' volatility surface, on average. This

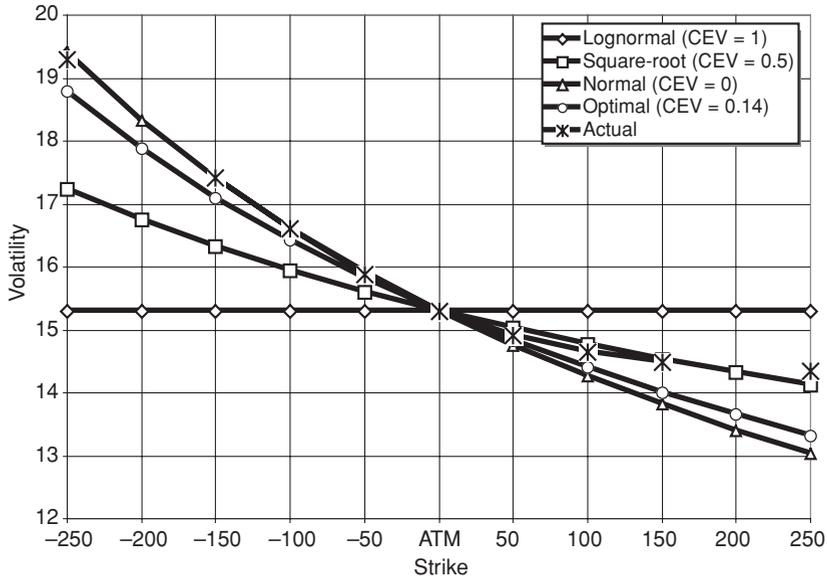


Figure 1 Implied Volatility Skew on 5-Year-into-10-Year Swap (1998–2004 Average)
 *Source of actual volatility: Bank of America; volatility for 200 bps ITM/OTM was not quoted.

measure is model-specific; unlike some other volatility indexes, it is not a simple average of swaption volatilities. The internal analytics of each model, exact or approximate, are used to translate the short rate volatility constant into swaption volatilities used for calibration. Note that this constant-volatility, zero mean rever-

sion setup is employed only to define the index; it is not a recommended setup for pricing complex instruments.

Figure 3 depicts the history of three volatility indexes (sigmas) computed from the beginning of 2000 for the HW model, the BK model, and the squared Gaussian model. Each index is

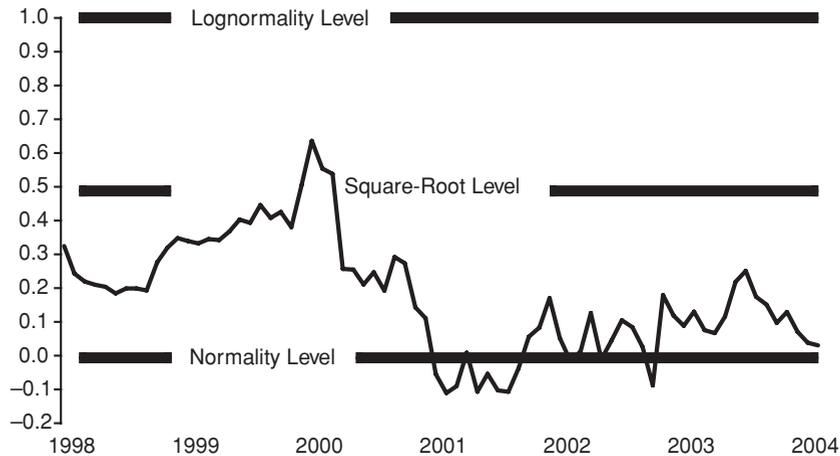


Figure 2 Historical CEV Values

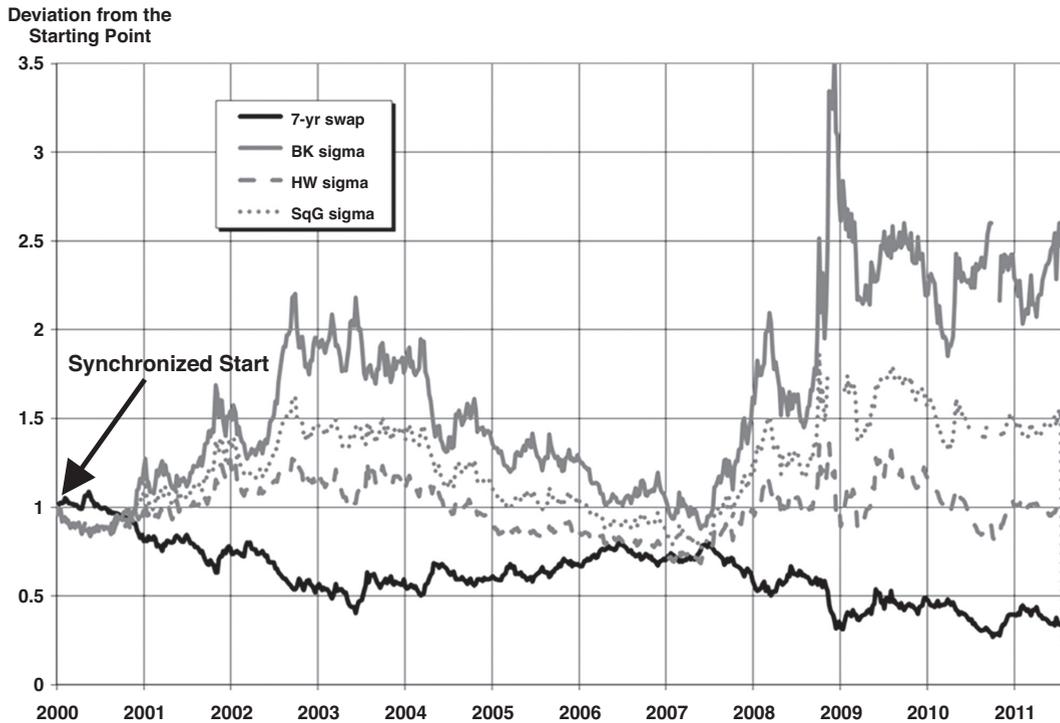


Figure 3 Which Volatility Index Is Most Stable?

calibrated to the same family of equally weighted ATM swaptions on the 2-year swap and the 10-year swap with expirations ranging from 6 months to 10 years. We add for comparison a line for the 7-year rate level, and scale all four lines so that they start from 1.0.

Figure 3 strongly confirms the normalization of the interest rate market; the volatility index constructed for the HW model has gradually become the most stable one. For example, the swap rate plunged a good 60% between January 2000 and June 2003, but the HW volatility index barely changed. The two other models produced volatility indexes that looked mirror-reflective of the rate level (the lognormal model does by far the worst job). A similar observation applies to the 2007–2010 period.

Interestingly enough, the SqG index was stable for most of 2003 and could handle the record-setting rate plunge. This confirms that the square root volatility pattern may outper-

form others when the rates are very low. These findings are consistent with the swaption skew measures we have discussed. This is not a coincidence at all. People who set the market for the ATM swaptions are the same ones who trade out-of- and in-the-money options.

In the sections to follow we will discuss how to extend the short-rate modeling framework to multifactor models and jump-diffusion models, which are often constructed in so-called affine analytical form.

ADDING A SECOND FACTOR TO SHORT-RATE MODELS

Let us consider a fixed income instrument that pays floating coupons indexed to some short rate (such as the 3-month LIBOR). The payer does not want to pay too much in case the curve inverts, so a cap is established equal to the level of some long, say 10-year, rate.

How much is this cap worth? Practically speaking, the curve's inversion is not so rare a phenomenon of the fixed income market. However, if the initial curve is steep, we will greatly undervalue the cap using any of the single-factor models described above. This example highlights the limitation of single-factor modeling: All rates change in unison. Instruments that contain "curve options," that is, asymmetric response to a curve's twist or butterfly moves, cannot be valued using single-factor term structures. Much more complex examples requiring multifactor modeling include American or Bermudan options, certain collateralized mortgage obligations (CMOs) that are much shorter or longer than the collateral itself.

Mathematically, a two-factor normal model can be constructed in a fairly simple way. Suppose that, instead of having one auxiliary Gaussian variable $x(t)$, we have two, $x_1(t)$ and $x_2(t)$, that follow linear stochastic differential equations:

$$\begin{aligned} dx_1 &= -a_1(t)x_1 dt + \sigma_1(t)dz_1 \\ dx_2 &= -a_2(t)x_2 dt + \sigma_2(t)dz_2 \end{aligned} \quad (22)$$

Brownian motions $z_1(t)$ and $z_2(t)$ may have correlated increments, $\text{corr}[dz_1, dz_2] = \rho$. Let us assume that ρ is equal to neither $+1$ nor -1 , and mean reversions $a_1(t)$ and $a_2(t)$ are positive and not identical to one another. These conditions ensure that the system (22) is stable and cannot be reduced to single-factor diffusion.

We now define the short rate simply as $r(t) = R(t) + x_1(t) + x_2(t)$ where deterministic function $R(t)$ is chosen to fit the initial yield curve. The short rate will be normally distributed; it can be shown that such a model possesses analytical tractability similar to the Hull-White single-factor model, see Levin (1998). In particular, the calibrating function $R(t)$ can be computed in a closed-end form given the forward curve, $f(t)$. The long zero-coupon rates are linear in $x_1(t)$ and $x_2(t)$,

$$r_T(t) = A(t, T) + B_{1T}(t)x_1(t) + B_{2T}(t)x_2(t)$$

Functions B 's depend on time t only if the mean reversions a 's do. If a 's are constant, then B 's depend only on maturity T and have a familiar form: $B_{iT} = (1 - e^{-a_i T})/a_i$, $i = 1$ or 2 .

The normal deviates, $x_1(t)$ and $x_2(t)$, bear no financial meaning. However, we can complement the short rate with an independent "slope" variable, $v = x_1 + \beta x_2$ with

$$\beta = -\sigma_1(\sigma_1 + \rho\sigma_2)/\sigma_2(\sigma_2 + \rho\sigma_1) \neq 1$$

The new variable has increments dv mathematically uncorrelated to dr ; it therefore can be interpreted as the driver of long rates independent of the short rate. The underlying processes, $x_1(t)$ and $x_2(t)$, can be transformed differently, thereby creating a pair of state variables with desired financial meanings, see Levin (2001). Levin (1998) developed a three-point calibration method that analytically computes parameters of the two-factor model using volatility of and correlation between the short rate and two arbitrary long rates. The method allows for constructing term structure models with inter-rate correlations selected by the user and maintained steadily over time. The latter property can be achieved by constructing a model with constant mean reversion parameters a_1 and a_2 , and a constant $\sigma_1(t)/\sigma_2(t)$ ratio.

Interestingly enough, all stable two-factor normal models having two real eigenvalues can be presented in the above-written form. Hull and White (1994) introduced a two-factor model that was designed in the form of a single-factor HW model for the short rate (factor 1) with a random long-term equilibrium rate (factor 2). Their approach draws on Brennan and Schwartz (1979). It is now clear that such an appeal to the financial meaning was unnecessary, and the general mathematical approach is as good or even better.

If we transform $x_1(t)$ and $x_2(t)$ nonlinearly, we will get multifactor versions of other previously considered models. For example, we could define the short rate as $r(t) = R(t)\exp[x_1(t) + x_2(t)]$, thereby creating a two-factor lognormal model. As one would expect, these models

inherit main properties of the single-factor parents, but add a greater freedom in changing the curve's shape and calibrating to volatility and correlation structures.

THE CONCEPT OF AFFINE MODELING

Affine modeling is a term introduced by Duffie and Kan (1996). It is a class of term structure models, often multifactor, where all zero-coupon rates are linear functions of factors. Therefore, the zero-coupon bond pricing has an exponential-linear form. Let us revisit the general stochastic model given by

$$dr = (\text{Drift})dt + (\text{Volatility})dz$$

Duffie and Kan showed that the model will be affine if drift and the square of volatility are both linear in rate r , or, more generally, in all market factors. In order to illustrate the main idea, let us denote the drift term as $\mu(x,t)$, the volatility term as $\sigma(x,t)$, and assume for the sake of simplicity that $r = x$, the lone market factor.

Every financial derivative satisfies a partial differential equation, see Duffie (1996). The left-hand side of this equation is equal to the investment's arbitrage-free expected return, which is the product of price (P) by the short rate (r). The right-hand side collects all the terms arising in the course of random behavior of $P(x,t)$: the decay, the drift, the diffusion, and cash received. In particular, a zero-coupon bond receives no cash; its equation is

$$rP(x,t) = \frac{\partial P(x,t)}{\partial t} + \mu(x,t) \frac{\partial P(x,t)}{\partial x} + \frac{1}{2} \sigma^2(x,t) \frac{\partial^2 P(x,t)}{\partial x^2} \quad (23)$$

subject to the terminal condition, $P(x,T) = 1$ (bond pays sure \$1 at maturity regardless of the market conditions). Suppose now that functions $\mu(x,t)$ and $\sigma^2(x,t)$ are linear in x :

$$\begin{aligned} \mu(x,t) &= \alpha_1(t) + \alpha_2(t)x; \\ \sigma^2(x,t) &= \beta_1(t) + \beta_2(t)x \end{aligned}$$

It turns out that the solution to equation (23) will have an exponential-linear form:

$$P(x,t) = \exp[a(t,T) + b(t,T)x]$$

To prove this conjecture, we place the above expressions into equation (23), take all derivatives, and observe that all the terms are either independent of x or linear in x . Collecting them, we get two ordinary differential equations defining unknown functions $a(t,T)$ and $b(t,T)$:

$$\begin{aligned} b'_i(t,T) &= -\alpha_2(t)b(t,T) - \frac{1}{2}\beta_2(t)b^2(t,T) + 1 \\ b(T,T) &= 0 \end{aligned} \quad (24)$$

$$\begin{aligned} a'_i(t,T) &= -\alpha_1(t)b(t,T) - \frac{1}{2}\beta_1(t)b^2(t,T) \\ a(T,T) &= 0 \end{aligned} \quad (25)$$

The terminal conditions for $a(t,T)$ and $b(t,T)$ are dictated by the terminal condition for the price function, $P(x,T) = 1$. Note that equation (24) defines function $b(t,T)$; once it is solved, we can solve (25) for $a(t,T)$.

It is clear that the HW model and the CIR model we considered earlier in the entry were affine. Indeed, in the HW model, β_2 is zero, α_2 is $-a$, β_1 is σ^2 , and (24) becomes a linear differential equation. In the CIR model, β_1 is zero, α_2 is again $-a$, and β_2 is σ^2 ; (24) becomes the Riccati equation (10). In fact, these two models cover all most important specifications of the affine modeling, for the single-factor case. The concept of affine modeling lets us build multifactor models systematically. The two-factor Gaussian model we introduced above was affine, too. Much more complex three-factor affine models were analyzed by Balduzzi et al. (1996) and by Dai and Singleton (2000). Among early works we should mention the model of Longstaff and Schwartz (1992). In their model, both the short rate and its volatility are affine in two factors that follow CIR-like processes.

The Jump-Diffusion Case

All term structure models considered thus far are based on *diffusion*—a continuous random

disturbance known as Brownian motion (Wiener process), $z(t)$. Short rates are somewhat jumpy and may require an addition of the Poisson process for modeling. The jump-diffusion extension to the affine modeling concept has been considered by many researchers (see Duffie and Kan, 1996; Das et al., 1996; and Das, 2000). The key point is that, under certain conditions, addition of jumps does not change the complexity of the problem; long rates remain affine in factors and even equation (24) for $b(t, T)$ remains unaffected.

Under the presence of *jumps*, the main stochastic differential equation for the short rate (or other market factors) gets an additional term as shown below:

$$dr = (\text{Drift})dt + (\text{Volatility})dz + (\text{Jump Volatility})dN$$

where N is the Poisson-Merton jump variable having intensity of λ . When a jump occurs, dN is drawn from the standard normal distribution $N_{[0,1]}$; it stays 0 otherwise. Continuing our affine-model notational style, let us denote the jump volatility term as $\sigma_j(t)$ and the jump intensity as $\lambda(x, t)$. Note that we allow jump's intensity, but not the size, to be factor dependent.

With jumps, the partial differential equation (23) will get one additional term to its right-hand side. If a jump of size δ occurs, the price of a zero-coupon bond, $P(x, t)$ before the jump, will become $P(x + \delta, t)$. The expected change of price can be written as

$$\int_{-\infty}^{\infty} [P(x + \delta, t) - P(x, t)]n_{[0, \sigma_j]}(\delta)d\delta$$

where, as usual, n denotes a normal density function. This expression captures the randomness of the jump's size, not the randomness of the jump's occurrence. Multiplying it by the probability of a jump to occur between t and $t + dt$ (that is, λdt) we get the cumulative expected effect of price change. Finally, dividing by dt we get the annualized return component caused by the jumps. Therefore, the partial-

differential equation (23) will now become a partial integral-differential equation:

$$\begin{aligned} rP(x, t) = & \frac{\partial P(x, t)}{\partial t} + \mu(x, t) \frac{\partial P(x, t)}{\partial x} + \frac{1}{2} \sigma^2(x, t) \frac{\partial^2 P(x, t)}{\partial x^2} \\ & + \lambda(x, t) \int_{-\infty}^{\infty} [P(x + \delta, t) - P(x, t)]n_{[0, \sigma_j]}(\delta)d\delta \end{aligned} \quad (26)$$

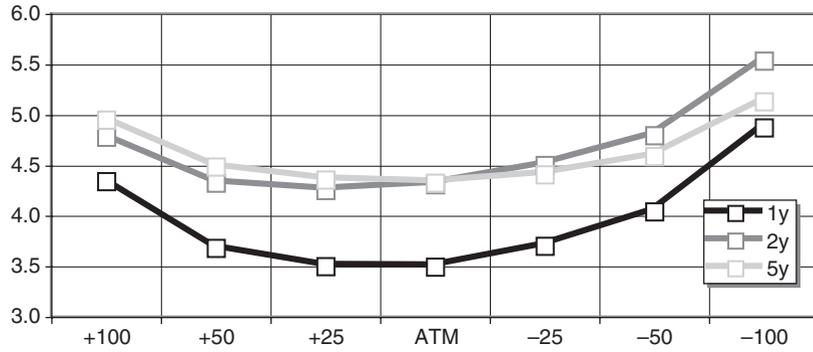
For the diffusion case, we required functions $\mu(x, t)$ and $\sigma^2(x, t)$ to be linear in x . Let us extend this condition to the jump's intensity: $\lambda(x, t) = \gamma_1(t) + \gamma_2(t)x$. It turns out that the exponential-linear form $P(x, t) = \exp[a(t, T) + b(t, T)x]$ still fits the equation. Again, collecting terms, we get two ordinary differential equations defining unknown functions $a(t, T)$ and $b(t, T)$:

$$\begin{aligned} b'_t(t, T) = & -\alpha_2(t)b(t, T) - \frac{1}{2}\beta_2(t)b^2(t, T) \\ & - \gamma_2(t)[e^{\frac{1}{2}b^2(t, T)\sigma_j^2(t)} - 1] + 1 \\ b(T, T) = & 0 \end{aligned} \quad (27)$$

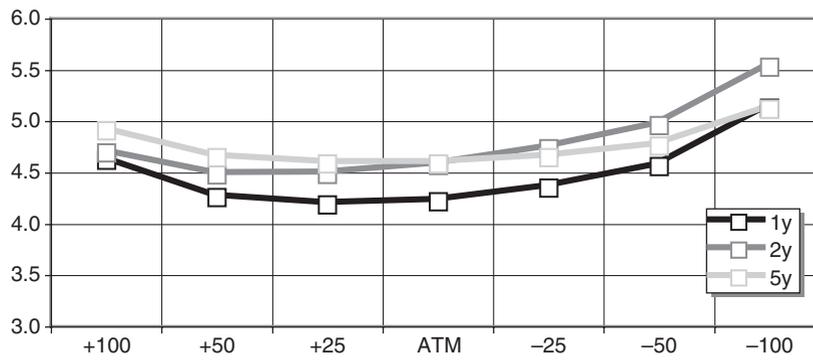
$$\begin{aligned} a'_t(t, T) = & -\alpha_1(t)b(t, T) - \frac{1}{2}\beta_1(t)b^2(t, T) \\ & - \gamma_1(t)[e^{\frac{1}{2}b^2(t, T)\sigma_j^2(t)} - 1] \\ a(T, T) = & 0 \end{aligned} \quad (28)$$

Notably, equation (27) defining function $b(t, T)$ will coincide with previously discussed equation (24) if $\gamma_2 = 0$. If we have a single-factor model, the linear relationship between long rates and the short rate will have a slope of $b(t, t + T)/T$. This slope, found for an affine diffusive model, won't change if we add jumps of factor-independent intensity and size. Hence, in such affine models, jumps and diffusions are equally propagated from the short rate to long rates. Knowing that actually observed long rates are chiefly diffusive and the short rate is notably jumpy, one can conclude that the jump-diffusive setting makes more practical sense within the frame of multifactor modeling.

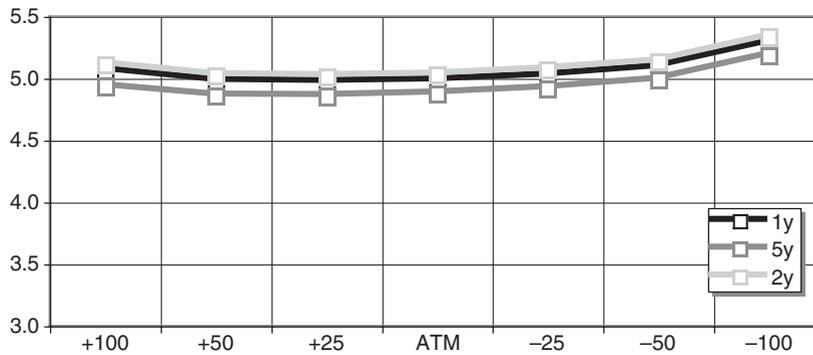
Using jump-diffusion models may be required when valuing options struck away from the current forward rate (that is, the ATM point). Aside from the *volatility skew*, option pricing features *volatility smile*, or simply an



A. 1-month Expiration on Various Swap Tenors



B. 6-month Expiration on Various Swap Tenors



C. 2-year Expiration on Various Swap Tenors

Figure 4 Daily Normalized Volatility Smile for Traded Swaptions (bp/day)
Data are courtesy of Bear Stearns, January 2007.

excessive convexity in σ_K . Revisiting Figure 1, one can notice that the actual dependence of volatility on the strike is more convex than even the optimal CEV model predicts. This is the smile effect, albeit fairly moderate for options on long rates. Smiles for options on shorter rates are very apparent, especially for short ex-

pirations. Figure 4 depicts swaption volatility measured in basis points per day, as a function of strike.

In this normalized scale, all panels of Figure 4 exhibit similar volatility skews, the ones close to normal (CEV = 0). However, the smile effect looks very different in panels A, B,

and C ; it clearly fades with maturity of the underlying rate and the option's expiry. The presence of jumps fattens the distribution tails and inflates out-of-the money or in-the money option values relatively to the ATM values. Therefore, jump modeling can capture the smile effect and explains its dependence on the swap's maturity and the option's expiry: Jumps allowed to occur over a longer time horizon look more like diffusion.

KEY POINTS

- The concept of short-rate modeling serves as a foundation for the fixed-income derivatives market.
- Short-rate models can be single- or multifactor, but their central object is a theoretical risk-free rate. Models employed in the financial markets have to be calibrated to the initial yield curve and simple options; some models let us solve this task analytically.
- There are a number of single-factor models that differ with respect to their distribution of rates, interrate relationships, and ability to fit the swaption market; the Hull-White (normal) model seems to fit the observed volatility skew the best.
- A two-factor normal model can be constructed by borrowing the recipes of so-called "affine" models; such a model can be used to price complex derivatives that are asymmetrically exposed to changes in the yield curve's shape.
- With jumps included, models can be employed to capture volatility "smile," that is, value options struck far out-of- or in-the-money.

ACKNOWLEDGMENTS

The author wishes to thank Andrew Davidson for his help in shaping this material, Will Searle for his comments, and Nancy Davidson for her editorial work.

REFERENCES

- Balduzzi, P., Das, S., Foresi, S., and Sundaram, R. (1996). A simple approach to three factor affine term structure of interest rates. *Journal of Fixed Income* 6: 43–53.
- Beaglehole, D., and Tenney, M. (1991). General solution of some interest rate contingent claim pricing. *Journal of Fixed Income* 1: 69–83.
- Black, F., Derman, E., and Toy, W. (1990). A one factor model of interest rates and its application to the Treasury bond option. *Financial Analysts Journal* (January–February 1990): 33–39.
- Black, F., and Karasinski, P. (1991). Bond and option pricing when short rates are lognormal. *Financial Analysts Journal* (July–August): 52–59.
- Blyth, S., and Uglum, J. (1999). Rates of skew. *Risk* (July): 61–63.
- Boyle, P., Tian, W., and Guan, F. (2002). The Riccati equation in mathematical finance. *Journal of Symbolic Computation* 22, 3: 343–356.
- Brace, A., Gatarek, D., and Musiela, M. (1997). The market model of interest rate dynamics. *Mathematical Finance* 7: 127–155.
- Cox, J. C., Ingersoll, J. E., and Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica* 53: 385–407.
- Cox, J.C., J.E. Ingersoll and S.A. Ross (1985). 'A Theory of the Term Structure of Interest Rates', *Econometrica* 53: 385–407.
- Dai, Q., and Singleton, K. (2000). Specification analysis of affine term structure models. *Journal of Finance* 55, 5: 1943–1978.
- Davidson, A., Sanders, A., Wolff, L.-L., and Ching, A. (2003). *Securitization: Structuring and Investment Analysis*. Hoboken, NJ: John Wiley & Sons.
- Das, S. (2000). Interest rate modeling with jump-diffusion processes. In N. Jegadeesh and B. Tuckman (eds.), *Advanced Fixed-Income Valuation Tools* (pp. 162–189). Hoboken, NJ: John Wiley & Sons.
- Das, S., Balduzzi, P., Foresi, S., and Sundaram, R. (1996). A simple approach to three factor affine models of the term structure. *Journal of Fixed Income* 6, 3: 43–53.
- Duffie, D. (1996). *Dynamic Asset Pricing Theory*, 2nd edition. Princeton, NJ: Princeton University Press.
- Duffie, D., and Kan, R. (1996). A yield-factor model of interest rates. *Mathematical Finance* 6, 4: 379–406.
- Flesaker, B., and Hughston, L. (1996). Positive interest. *Risk* 9, 1: 46–49.
- Heath, D., Jarrow, R., and Morton, A. (1992). Bond pricing and the term structure of interest rates:

- A new methodology for contingent claim valuation. *Econometrica* 60, 1: 77–105.
- Ho, T. S. Y., and Lee, S. B. (1986). Term-structure movements and pricing interest rate contingent claims. *Journal of Finance* 41, 5: 1011–1029.
- Hughston, L., ed. (1996). *Vasicek and Beyond*. London: Risk Publications
- Hull, J., and White, A. (1990). Pricing interest-rate derivative securities. *The Review of Financial Studies* 3, 4: 573–592.
- Hull, J., and White, A. (1994). Numerical procedures for implementing term structure models II: Two factor models. *Journal of Derivatives* 2, 2: 37–48.
- James, J., and Webber, N. (2000). *Interest Rate Modeling: Financial Engineering*. Hoboken, NJ: John Wiley & Sons.
- Jamshidian, F. (1996). Bond, futures and option valuation in the quadratic interest rate models. *Applied Mathematical Finance* 3: 93–115.
- Levin, A. (1998). Deriving closed-form solutions for Gaussian pricing models: A systematic time-domain approach. *International Journal of Theoretical and Applied Finance* 1, 3: 349–376.
- Levin, A. (2004). Interest rate model selection. *Journal of Portfolio Management* 30, 2: 74–86.
- Longstaff, F., and Schwartz, E. (1992). Interest rate volatility and the term structure: A two-factor general equilibrium model. *Journal of Finance* 47, 4: 1259–1282.
- Pelsser, A. (1997). A tractable interest rate model that guarantees positive interest rates. *Review of Derivatives Research* 1: 269–284.

Static Term Structure Modeling in Discrete and Continuous Time

DAVID AUDLEY, PhD

Senior Lecturer, The Johns Hopkins University

RICHARD CHIN

Investment Manager, New York Life Investments

PETER C. L. LIN

PhD Candidate, The Johns Hopkins University

SHRIKANT RAMAMURTHY

Consultant, New York, NY

Abstract: The term structure of interest rates represents the cost of (return from) borrowing (lending / investing) for different terms at any one moment in time. The term structure is most often specified for a specific market such as the U.S. Treasury market, the bond market for double-A rated financial institutions, the interest rate market for LIBOR and swaps, and so on. The term structure is usually specified via a rate or yield for a given term or the discount to a cash payment at some time in the future. These are often summarized mathematically through a wide variety of models. In addition, term structure models are fundamental to expressing value and risk, and establishing relative value across the spectrum of instruments found in the various interest-rate or bond markets. Static models of the term structure are characterizations that are devoted to relationships based on a given market and do not serve future scenarios where there is uncertainty. Standard static models include those known as the spot yield curve, discount function, par yield curve, and the implied forward curve. Instantiations of these models may be found in both a discrete- and continuous-time framework. An important consideration is establishing how these term structure models are constructed and how to transform one model into another.

The objective of this entry is to describe the principles and approaches for a deterministic model of the term structure of interest rates. This is done first in a discrete-time setting, followed by a more analytical development in a continuous-

time setting. We provide an eclectic mixture of ideas from the academic literature in concert with adaptations well known to practitioners.

Computational implementation of anything as complex as interest rate term structure

models naturally engenders the rigorous adherence to, yet clever application of, some arcane ideas from software/system engineering. These are beyond the scope of this introduction, but such topics include numerical recipes; mechanisms to ensure internal consistencies during development and build-up; tests for internal consistency, verification, and validation of completed applications (e.g., put-call parity, cash-and-carry arbitrage, and others); parameterization of models and applications from the markets; and the utility of advanced computer architectures.

A deterministic approach to the term structure of interest rates (or simply, the term structure) may be appropriately thought of as a static modeling approach. This is distinguished from a dynamic model of term structure. The chief distinction is that in a static term structure model, no accommodation is made of the course of interest rates over time. On the other hand, a dynamic model explicitly incorporates how *interest rates* change over time and therefore needs to admit a notion of uncertainty in considering the future course of interest rates. The following discussion will concentrate on static models. First, we address a taxonomy for term structure models in some additional detail.

INTRODUCTION TO TERM STRUCTURE MODELING

The *term structure of interest rates* (or term structure) is simply a price or yield relationship among a set of securities that differ only in the timing of their cash flows or their term until maturity. These securities invariably have a specified set of other attributes in common so that the study of the term relationship is meaningful.

It is common to think of the *term structure* as consisting of the current-coupon U.S. Treasury issues only. This restriction is not necessary since it is possible to define other term structures derived from other securities. For exam-

ple, it is meaningful to define the term structure of sets of coupon or principal Treasury strips. Other examples include off-the-run Treasury issues, agency debentures, LIBOR/interest-rate swaps, or the notes of single-A rated banks and finance companies. The set of securities used to define a term structure is called the reference set. A *market sector* (sometimes referred to as a *market* or a *sector*) consists of all those instruments described by a specific term structure. There is the market sector of coupon or principal Treasury strips, off-the-run *Treasuries*, agency debentures, interest-rate swaps, and single-A rated banks and finance companies, and so forth. Very often, the reference set for a market sector may have restrictions on the structure (noncallable only), liquidity (recent issues only), or price (close to par only) of the securities that make up the set.

The relationship expressed by the term structure is traditionally the par-coupon yield relationship, hence the terminology: *yield curve*. This also is not a necessary restriction. In general, the term structure could be the *discount function*, the *spot-yield curve*, or some other expression of the price or yield relationship between the securities. Given the widespread usage of the (*par*) *yield curve* for the Treasury market, it is not surprising that many market sectors are defined from a reference set derived from the Treasury market. For example, the reference set that defines the agency debenture market is a set of yield spreads to the *on-the-run* Treasuries, so that a 5-year debenture issued by an agency may be priced at par to yield 15 basis points more than the current 5-year Treasury issue. If the Treasury issue is trading at a 6.60% yield to maturity, the par priced agency issue has a 6.75% coupon. By inference, from the spread quote of 15 basis points, the reference yield for the 5-year term is 6.75%. Similar statements can be made for the interest-rate swap and the corporate bond markets.

It needs to be emphasized that the reference set of bonds used to define the term structure

of interest rates and the resulting term structure itself are not one and the same. Indeed, the term structure, as a complete description of the entire yield curve, ultimately can be used to analyze all manner of option-laden, index-amortizing swaps or debentures that are in the same market sector. The “vanilla” reference set consists of individual bonds that are used mainly to define the term structure or to derive its defining relationships—spot-yield curve, spot-rate process, discount function, and the like.

Theories about the term structure of interest rates fall into two categories:

- *Qualitative theories* seek to explain the shape of the yield curve based on economic principles. Three theories attract the widest attention: the expectations, *liquidity preference*, and *preferred habitat* (or hedging pressure) theories.
- *Quantitative theories* seek to mathematically characterize the term structure (often in harmony with one of the qualitative theories).

Usually, a quantitative theory about the term structure of interest rates culminates in a mathematical model, a term structure model that exhibits useful properties. Specifically, a term structure model is the mathematical representation of the relationship among the securities in a market sector. This formalizes the distinction between the reference set used to define a market sector and a term structure model.

TERM STRUCTURE MODELS

The simplest and most familiar term structure model is the (semilogarithmic) graph of the U.S. Treasury yield curve (once found daily in the *Wall Street Journal* and in the business section of many newspapers). This model is useful mainly as a visualization of the yield relationship between the most recently issued shorter-term Treasury instruments and bonds. The graph can be characterized by a mathematical equation and is one example of the set of interpolation models of the term structure.

These “connect-the-dots” models can be useful in providing a quantitative way to price bonds outside the current-coupon Treasury issues, but their utility is rather limited. Bonds that are valued through a linear-interpolation technique may not be “fairly” valued in the sense that an average yield may not be equal to the “par-coupon” yield corresponding to the same date. Later we provide a discussion of how the par-coupon curve is constructed to be fairly valued in comparison to the set of reference (Treasury) issues.

The term structure model as described above simply provides a snapshot of the relationship between the yields for selected Treasury maturities on a given day. It is often required that term structure models exhibit additional “analytic” properties. One such property is the consistency associated with the preclusion of riskless arbitrage when the term structure model is used for pricing. More will be said about this later. For now, it is intended merely to indicate that the “visualization” of the yield relationship to term may be neither completely useful nor adequate.

More generally, term structure models are called on to describe the evolution of a set of interest rates over time. This motivates the following distinction in classifying term structure models:

- *Static models* of the term structure offer a mechanism to establish the “present value of a future dollar” in a deterministic economy. That is, no allowance for uncertainty or interest-rate volatility is explicitly incorporated into the model.
- *Dynamic models* of the term structure, in contrast to static models, explicitly allow for uncertainty in the future course of interest rates.

Ideally, a dynamic model of the term structure should have useful static models embedded within. That is, with no contingency on the receipt of a future cash payment or when there is an assumption of negligible volatility, a

dynamic model should correspond to a consistent static model.

The essence of term structure modeling is the process of converting the market description of a sector's reference set (the data) into a mathematical set of relationships that characterizes all issues in a sector. This is by no means trivial to do correctly. For example, the same model that correctly values a note in the Treasury market should also correctly value an option on that note, the futures contract into which that note may be deliverable, and an option on that futures contract. It should also reveal if the traded basis on that note is rich or cheap relative to the cash, futures, and options markets. It should also be able to describe any stripping or reconstitution opportunities between coupon and principal strips and the cash market. These analyses should not be the result of several models, but of a single term structure model.

A key element of the modeling process is to eliminate distinguishing characteristics associated with each constituent of the reference set. For example, in the on-the-run set of Treasury issues, there are bills as well as notes and bonds. The bills have different conventions for day counting, pricing, and yield expression from those of the coupon-paying issues of the sector. These characteristics need to be removed prior to developing the mathematical relation of the term structure model (as do the distinguishing characteristics for notes and bonds). In this simple example, a model of the Treasury term structure might be the spot curve or the discount function, as opposed to a "connect-the-dots" model to which no yield adjustments have been made.

The mathematical relationship of a term structure model can be used to characterize all issues in a sector. As is the case for the Treasury sector, every instrument can be considered a collection of zero-coupon bonds (the maturities of which correspond to the coupon/principal payment dates, the denominations of which

correspond to the amount of coupon/principal paid). Accordingly, the discount function or equivalently, its corresponding spot-yield curve, furnishes a pricing technique for each zero-coupon bond and, therefore, for each of the instruments. With this insight, the utility of equivalence between the spot-yield curve and discount function, which are derived from the original reference set, is readily apparent.

We begin with the familiar, *discrete-time* modeling approach. That is, units of time quanta are defined (usually in terms of compounding frequency) and financial manipulations are indexed with integer, multiple periods. We continue to build on the discussion by introducing the *continuous-time* analogies to the concepts developed for discrete-time modeling. Continuous-time modeling allows financial manipulations to be freed from discretization artifacts (such as compounding frequency) and provides an algebraic framework that more naturally and rigorously accommodates "rate" as a concept of change. In addition, this approach opens up a huge field of applicable mathematics with the attendant opportunity for abstraction. For example, continuous-time models free the analyst from artificial a priori assumptions about interest-rate lattices; allow concentration on the financial analyses at hand; defer time-step issues to final implementation of an algorithm; and let the analyst choose an approach based on convenience, speed, and accuracy.

DISCRETE-TIME MODELS OF THE TERM STRUCTURE

In the discrete-time framework, we introduce some fundamental concepts in term structure theory. These include the discount function, the spot rate and spot yield, and the forward rate. While these initially may appear to be esoteric in nature, they are in fact closely interrelated quantities that directly represent the term structure, or act to influence the course

of future interest rates in an arbitrage-free environment. In this section these concepts are shown to be incorporated into the different expressions that describe the various qualitative term structure theories, such as the expectation, preferred habitat, and liquidity preference hypotheses. The continuous-time term structure model is evolved next from the same underlying premises as found in discrete time, thereby speeding the exposition.

DISCOUNT FUNCTION

The discount function incorporates market yield-curve information to express the present value of a future dollar as a function of the term to its receipt. As such, the discount function is a valid expression of the term structure of interest rates by virtue of the price/yield relationship. Since the discount function is used to quantify the value of a future dollar, the discount function also provides a direct means to value a coupon-paying bond since the coupon and principal payments are simply scalar multiples of a single dollar. As a result, the discount function can be used as a reference check for other quantitative term structure models.

Quantitative term structure models ultimately deal with the analysis of pure discount bonds. (Discount bonds, or zero-coupon bonds, are the simplest types of bonds to analyze as there is only the repayment of par at maturity. Further, all other bonds can be built from a series of discount bonds and options on discount bonds.) As a consequence of modeling the yield movements of discount bonds, term structure models describe their price movements since the price/yield relationship allows the term structure to be analyzed in terms of either price or yield.

This relationship is addressed further later in this entry, in which the term structure model is expressed in terms of price as a function of rate and time.

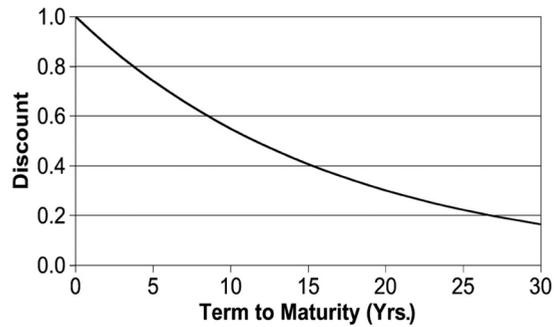


Figure 1 Discount Function

If it is assumed that the discount bond pays one dollar at maturity, then the present value of the bond is some decimal fraction less than one. For a set of discount bonds of increasing maturities, there is the corresponding set of present values starting from approximately 0.999 and decreasing thereafter. This set of present values is called the *discount function* and is shown in Figure 1.

The discount function is the term-to-maturity relationship of the present value of a future unit of cash flow. More formally, for a cash flow, CF , received after a term, T , from today, t , the present value, PV , of that cash flow is discounted, d , from the future value CF as expressed by the relation

$$PV(t, T) = d(t, T) \times CF(t, T) \tag{1}$$

where

$PV(t, T)$ = present value of the cash flow at t

$d(t, T)$ = discount at t for a cash flow received T after t

$CF(t, T)$ cash = flow received at $t + T$

As we are able to generate the discount function, d , for all terms-to-maturity, T , this can be a valid representation of the term structure of interest rates. Indeed, the discount function reflects the Treasury term structure when the discount function exactly reprices the current-coupon Treasury issues.

Deriving the Discount Function for On-the-Run Treasuries

More generally, let $P(t, i) = P_i(t)$ be the set of closing prices on (date) t for the set of current-coupon Treasury bonds (where the index, i , associates a specific issue among several)

$P(t, 3\text{-month})$: price of the 3-month (13-week) bill, at time t

$P(t, 6\text{-month})$: price of the 6-month (26-week) bill, at time t

$P(t, 52\text{-week})$: price of the 1-year (52-week) bill, at time t

$P(t, 2\text{-year})$: price of the 2-year note, at time t

...

$P(t, 30\text{-year})$: price of the 30-year bond, at time t

Each of these instruments has its own time series of cash flows, each with its own individual term-to-maturity (or better, term-to-payment). For the Treasury bills, the cash flows and associated terms-to-maturity are

3-month bill: $CF(t, T(3\text{-month}, 1))$,

6-month bill: $CF(t, T(6\text{-month}, 1))$,

and for the periodic instruments,

2-year note: $CF(t, T(2\text{-year}, 1))$,

$CF(t, T(2\text{-year}, 2))$, ..., $CF(t, T(2\text{-year}, 4))$,

...

30-year bond: $CF(t, T(30\text{-year}, 1))$,

$CF(t, T(30\text{-year}, 2))$, $CF(t, T(30\text{-year}, 60))$

The term to each of the cash flows, $T(i, j) = T_{i,j}$, is specific to the instrument and the context of the notion of "today," t for the purpose of establishing a present value. (In this sense, the dependence on t has been suppressed and it might be more precise to specify T as $T(t; i, j)$, but we believe this to be unnecessary.) The index j is the sequence of the cash flow in the time series for security i . Finally, in general, cash flows only exist in a futures sense. If $T(i, j)$ is less than zero (the cash flow has already been paid), then

those j -cash flows are not included in the series (although this is not an issue for the current-coupon Treasury issues).

The present value of a coupon-paying instrument is simply the sum of the discounted present values of the cash flows that make up the coupon payments and the payment of principal. Accordingly, for the discount function to model the Treasury term structure (i.e., the market sector defined by the on-the-run Treasury reference set), the following equations must be simultaneously satisfied. In this way, the discount function will reprice the current-coupon Treasury issues.

$$P(t, 3\text{-month}) = d(t, T(3\text{-month}, 1)) \times CF(t, T(3\text{-month}, 1))$$

$$= d(t, T_{1,1}) \times CF(t, T_{1,1})$$

$$P(t, 6\text{-month}) = d(t, T(6\text{-month}, 1)) \times CF(t, T(6\text{-month}, 1))$$

$$= d(t, T_{2,1}) \times CF(t, T_{2,1})$$

$$P(t, 2\text{-year}) = P(t, 3) = \sum_{j=1}^4 d(t, T_{3,j}) \times CF(t, T_{3,j})$$

...

$$P(t, 30\text{-year}) = P(t, 8) = \sum_{j=1}^{60} d(t, T_{8,j}) \times CF(t, T_{8,j})$$

The last cash flow of each series consists of the principal payment and, for the notes and bond, one coupon payment. The solution to these simultaneous equations furnishes many distinct points of term in which the discount function is defined; the long bond alone may have as many as 60 term points. Depending on the circumstances surrounding each auction, there may be as many as over 90 distinct points of term defining the discount function.

As with the earlier "connect-the-dots" model for the yield curve, in which the yield points were connected to generate intermediate values for the term structure, similar ideas can be used to accommodate the cash flows that do not fall on one of the terms, $T(i, j)$, enumerated above. In fact, interpolation techniques using spline functions may be applied to create

a continuous discount-function curve, as in Oldrich and Fong (1982).

The discount function forms the basis for the development of a term structure model, as will be developed further in later sections. As the discount function is an expression of the term structure based on price, there is no ambiguity of compounding periodicity, as with yield-based term structure models. The discount function simply expresses the nondimensional, fractional, present value of a unit cash flow to be received after some term. The term may be specified in a unit of time (e.g., years, months, or days) or in periods, in which the period length is a unit of time.

SPOT YIELD CURVE

With the assumption of a compounding convention (usually semiannual), the discount function can be used to derive the equivalent Treasury zero-coupon structure—sometimes referred to as the *spot-yield curve*. In this case, the spot-yield curve is an equivalent term structure representation based on yield that provides a view of the term structure that is more familiar. The equivalence between these two forms of the term structure is used later in this entry.

The *spot yield*, R , is related to the discount function, d , through a price/yield relation. By definition, the present value at t , $PV(t, n)$, of a cash flow received n periods in the future, $CF(t, n)$, has the spot yield, $R(t, n)$, through the relation

$$PV(t, n) = \frac{CF(t, n)}{[1 + R(t, n)]^n} \tag{2}$$

We use the discrete notion of integer periods, with each period of length P , to keep the math simple at this point. The more general case of a noninteger world is treated when a continuous time model is introduced.

Comparing equations (2) and (1) provides the relation between the spot yield and the discount function

$$d(t, n) = \frac{1}{[1 + R(t, n)]^n} \tag{3}$$

where

$d(t, n)$ = discount of a cash flow received n periods after t

$R(t, n)$ = n -period spot yield on t

The spot-yield curve is just the set of spot yields for all terms-to-maturity. In contrast, the spot rate is simply the one-period rate prevailing on t for repayment one period later. In the above notation, the spot rate is denoted $R(t, 1)$.

We can generalize the earlier comment about coupon-paying bonds in terms of the set of spot yields. The present value of a coupon-paying instrument is simply the sum of the discounted (present value) of the cash flows that make up the coupon payments and the payment of principal. The analogy to equation (2) for a coupon-paying bond using spot yields is

$$PV(t, n) = \frac{CF(t, 1)}{[1 + R(t, 1)]} + \frac{CF(t, 2)}{[1 + R(t, 2)]^2} + \dots + \frac{CF(t, n)}{[1 + R(t, n)]^n} \tag{2a}$$

Similarly, the analogy to equation (1) for a coupon-paying bond using the discount function is given by

$$PV(t, n) = d(t, 1) \times CF(t, 1) + d(t, 2) \times CF(t, 2) + \dots + d(t, n) \times CF(t, n) \tag{1a}$$

IMPLIED FORWARD RATE

A consequence of the discount function, spot yield, and spot rate is the immediate relation to the (implied) *forward rates*. The implied forward rate is the spot rate embodied in today's yield curve for some period in the future. The forward rate generally is regarded as an indication of future spot rates in an arbitrage-free economy. In the absence of arbitrage and uncertainty, the future spot rate, by definition, is equal to the forward rate. In the arbitrage-free term structure model discussed later, it can be shown that the future spot rate continuously converges toward the forward rate as the spot rate evolves over time.

Specifically, the one-period forward rate, F , can be determined from the spot yields as follows. Consider the one-period and two-period spot yields; the forward rate, F , may be found from

$$(1 + R(t, 2))^2 = (1 + R(t, 1)) \times (1 + F(t, 1, 1)) \quad (4)$$

where

$R(t, 2)$ = two-period spot yield on t

$R(t, 1)$ = one-period spot rate on t

$F(t, 1, 1)$ = one-period forward rate one period from t

This relation follows from the no-arbitrage assumption intrinsic in the concept of forward rates. The calculation of the forward rate presumes that an investment today for two periods provides the same return as a one-period investment today immediately rolled into another one-period investment one period from now. That is

$$PV(t) = \frac{CF(t, 2)}{[1 + R(t, 2)]^2} \quad (5)$$

$$PV(t) = \frac{CF(t, 2)}{[1 + R(t, 2)] \times [1 + F(t, 1, 1)]} \quad (6)$$

By equating equations (5) and (6), equation (4) results.

Deriving Forward Rates from Spot Yields

Implied from the term structure, through the spot-yield curve, is a set of forward rates. These forward rates may be iteratively defined from the above and written as follows

$$(1 + R(t, n))^n = (1 + R(t, n - 1))^{n-1} \times (1 + F(t, 1, n - 1))$$

where in addition to the earlier notation, $F(t, 1, n - 1)$ = one-period forward rate $n - 1$ periods from t , and noting, through substitution, that

$$(1 + R(t, n))^n = (1 + R(t, 1)) \times (1 + F(t, 1, 1)) \times (1 + F(t, 1, 2)) \times \dots \times (1 + F(t, 1, n - 1)) \quad (7)$$

this furnishes the first $n - 1$ one-period forward rates.

The relation among spot yield, spot rate, and forward rates, equation (7), can be combined with equation (2) to furnish a method for calculating the present value, at t , of a single n -period future cash flow based on a series of one-period forward rates

$$PV(t, n) = \frac{CF(t, n)}{[1 + R(t, 1)] \times \dots \times [1 + F(t, 1, n - 1)]} \quad (8)$$

Since the present value of a coupon-paying security is simply the sum of the discounted present value of the cash flows that make up the coupon payments and the payment of principal (see equations (1a) and (2a)), the analogy to equation (8) for determining the present value of a coupon-paying bond is

$$PV(t, n) = \frac{CF(t, 1)}{[1 + R(t, 1)]} + \frac{CF(t, 2)}{[1 + R(t, 1)] \times [1 + F(t, 1, 1)]} + \dots + \frac{CF(t, n)}{[1 + R(t, 1)] \times \dots \times [1 + F(t, 1, n - 1)]}$$

This equation may be used to define multi-period forward rates.

Deriving Forward Rates from the Discount Function

The discount function provides a direct method for generating forward rates. The one-period forward return $n - 1$ periods from t is obtained through the following

$$1 + F(t, 1, n - 1) = \frac{d(t, n - 1)}{d(t, n)} \quad (9)$$

Equation (9) may be derived from earlier equations, or from the following argument that creates a synthetic forward position. For each unit of cash delivered n periods from today, t , we pay $d(t, n)$. We take a long position in this zero. We also short $d(t, n)/d(t, n - 1)$ units of cash to be delivered $n - 1$ periods from t . For this we receive $d(t, n - 1)$ times $d(t, n)/d(t, n - 1)$, or simply $d(t, n)$, units. There is no net change in our cash position today. After $n - 1$ periods we pay out $d(t, n)/d(t, n - 1)$ and after n periods

receive one unit of cash. Thus the forward price per unit, FP , to be paid $n - 1$ periods from now is

$$FP(t, 1, n - 1) = \frac{d(t, n - 1)}{d(t, n)}$$

where

$FP(t, 1, n - 1)$ = forward price of a one-period unit of cash $n - 1$ periods from now

The forward price then gives the forward one-period rate, $n - 1$ periods from t as

$$FP(t, 1, n - 1) = \frac{1}{1 + F(t, 1, n - 1)}$$

Equating these results in equation (9).

TERM STRUCTURE IN A CERTAIN ECONOMY

As discussed earlier, term structure models describe the evolution of interest rates over time. Often, future interest rates are expressed in terms of the future spot rate. If the future spot rate (or equivalently, the future rate of return on a bond) is known, the future term structure of interest rates may be found from the previously established interrelationships between the spot rate and the discount function or spot yield. In fact, it is this relationship between the spot rate and the discount function that is used to motivate the formulation of the term structure models described later as a function of the spot rate. As a precursor to a generalized term structure theory, we first discuss the ramifications for a term structure in a certain economy. (In this context, "certain" refers to an economy with a lack of randomness, in other words, a lack of uncertainty.)

If the future course of interest rates is known with certainty, then arbitrage arguments demand that future spot rates be identical to forward rates. In the notation presented in equation (7), this is equivalent to noting that

$$R(t + nP, 1) = F(t, 1, n) \tag{10}$$

for $n = 1, 2, 3, \dots$ and where P is the term of the period. If this condition were violated, say, for example,

$$F(t, 1, n) > R(t + nP, 1)$$

then the same arbitrage argument may be made as before: If we buy the synthetic forward (this is a long position in a unit zero to be delivered $n + 1$ periods from today, t); and short $d(t, n + 1)/d(t, n)$ units of cash to be delivered n periods from today, t , no cash changes hands today. However, after n periods, we pay the forward price, FP ,

$$FP(t, 1, n - 1) = \frac{1}{1 + F(t, 1, n - 1)}$$

to receive one unit of cash after $n + 1$ periods. Also, after n periods, at $t + nP$, we sell the one-period unit zero for a price of

$$\frac{1}{1 + R(t + nP, 1)}$$

We know we can do this since there is no uncertainty in the economy. If, as assumed, $F(t, 1, n) > R(t + nP, 1)$, then after n periods the long and short positions yield a positive net cash flow, or a riskless arbitrage, of

$$\frac{1}{1 + R(t + nP, 1)} - \frac{1}{1 + F(t, 1, n)} > 0$$

after n periods with no uncertainty and with no net investment. Arbitrageurs will exploit the imbalance of the n -period forward rate with the spot rate n periods from now by continuing to buy the synthetic forward until demand outstrips supply. In this scenario, the synthetic forward price goes up, and the forward rate, $F(t, 1, n)$, goes down to $R(t + nP, 1)$ —with predictable effect on $d(t, n + 1)$ and/or $d(t, n)$. On the other hand, if $F(t, 1, n) < R(t + nP, 1)$, we may reverse our positions, and the same argument shows that $F(t, 1, n)$ will increase to $R(t + nP, 1)$.

Using the no-arbitrage condition in a certain economy, equation (10), in the present value expression from the implied forward-rate expression, equation (8) (which always holds irrespective of assumptions about the economy),

we have,

$$\begin{aligned} PV(t, n) &= \frac{CF(t, n+1)}{[1 + R(t, 1)] \times [1 + R(t+P, 1)] \times \cdots \times [1 + R(t+nP, 1)]} \\ &= \frac{CF(t, n+1)}{[1 + R(t, n+1)]^{n+1}} \end{aligned} \quad (11)$$

This means that the certain return of holding an $n + 1$ period zero until maturity is the same as the total return on a series of one-period bonds over the same period. Later we will discuss the various forms of equation (11) from various qualitative term structure theories.

Given equation (11), we have, at time P (one period) later,

$$\begin{aligned} PV(t+P, n) &= \frac{CF(t, n+1)}{[1 + R(t+P, 1)] \times \cdots \times [1 + R(t+nP, 1)]} \end{aligned}$$

so we find that the single-period return on a long-term zero is

$$\frac{PV(t+P)}{PV(t)} = 1 + R(t, 1) \quad (12)$$

Since the term-to-maturity was not specified, equation (12) must be true for zeros of any maturity. That is, the return realized on every discount bond over any period is equal to one plus the prevailing spot rate over that period. This will be expanded upon later.

Alternatively, we can use our relation for the discount function in equation (1), noting

$$PV(t+T, n) = d(t+P, n) \times CF(t, n+1)$$

and

$$PV(t, n) = d(t, n+1) \times CF(t, n+1)$$

and restate equation (12) in its discount-function based form:

$$\frac{d(t+P, n)}{d(t, n+1)} = 1 + R(t, 1)$$

While these developments for the certain economy may appear trivial and obvious, they serve as a guide for modeling the term structure under uncertainty as well.

TERM STRUCTURE IN THE REAL WORLD—NOTHING IS CERTAIN

In the real-world economy, the future course of interest rates contains uncertainty. In attempting to deal with uncertainty, however, it would not be inconceivable that a belief in the efficiency of the market would prompt one to use the term structure and the relation between forward rates and spot rates as indicators of expectation about the future. Indeed, market efficiency states that prices reflect all available information bearing on the valuation of the instrument. *Equilibrium* supply and demand for fixed-income instruments reflect a market-cleared consensus of the economic future. As uncertainty represents a departure from this consensus, the expected equilibrium offers a natural starting point for analysis.

Expectations Hypothesis

The expectations theory of the term structure of interest rates offers a good starting point for dealing with an uncertain future. Actually, there is a whole family of expectations theories. Broadly, the expectations theory states that the expected one-period rate of return on an investment is the same, regardless of the maturity of the investment. That is, if the investment horizon is one year, it would make no difference to invest in a one-year instrument, a two-year instrument sold after one year, or two sequential six-month instruments.

The most common form of this statement uses equation (10) as the basis for the theory. This is referred to as the unbiased expectations hypothesis, which states that the expected future spot rate is equal to the forward rate, or

$$E[R(t+nP, 1)] = F(t+kP, 1, n-k)$$

for $k = 0, 1, \dots, n-1$, and where $E[\cdot]$ is the expectation operator.

Using this relation, we find from equation (8) that the present value in an economy

characterized by unbiased expectations is

$$PV(t, n) = \frac{CF(t, n+1)}{[1+R(t, 1)] \times \{1+E [R(t+P, 1)]\} \times \dots \times \{1+E [R(t+nP, 1)]\}} \quad (13)$$

Therefore, the unbiased expectations hypothesis concludes that the guaranteed return from buying an $(n + 1)$ period bond and holding it to maturity is equivalent to the product of the expected returns from holding one-period bonds using a strategy of rolling over a series of one-period bonds until maturity.

Alternatively, the return-to-maturity expectations hypothesis is based on equation (11). Here we find that present value in such an economy is

$$PV(t, n) = \frac{CF(t, n+1)}{E \{ [1+R(t, 1)] \times [1+R(t+P, 1)] \times \dots \times [1+R(t+nP, 1)] \}} \quad (14)$$

The return-to-maturity expectations hypothesis assumes that an investor would expect to earn the same return by rolling over a series of one-period bonds as buying an $(n + 1)$ -period bond and holding it to maturity.

The last version of the expectations hypothesis that we will mention (there are others) is the local-expectations hypothesis (or risk-neutral hypothesis). This hypothesis is based on equation (12), or equivalently, its associated discount function-based equation. Under this hypothesis, the expected rate of return over a single period is equal to the prevailing spot rate of interest. Applying these expressions recursively gives

$$\begin{aligned} PV(t) &= \frac{E [PV(t + P)]}{[1 + R(t, 1)]} \\ &= E \left\{ \frac{PV(t + 2P)}{[1 + R(t + P, 1)] \times [1 + R(t, 1)]} \right\} \\ &= CF(t, n + 1) \times E \left\{ \frac{1}{[1 + R(t, 1)] \times [1 + R(t + P, 1)] \times \dots \times [1 + R(t + nP, 1)]} \right\} \end{aligned} \quad (15)$$

Equations (13), (14), and (15) are clearly different in that the coefficient of the cash flow,

$CF(t, n + 1)$, received $n + 1$ periods in the future is a different expression in each case. Furthermore, by the principle from mathematical analysis known as Jensen's inequality, only one of the expressions can be true if the future course of interest rates is uncertain.

In fact, in discrete time, we find that bond prices given by the unbiased and return-to-maturity hypotheses are equal but less than that given by the expectations hypothesis. Although the three hypotheses are different, in discrete time, any of these hypotheses is an acceptable description of equilibrium.

In the next section, term structure modeling in continuous time is developed. Equations (13), (14), and (15) have continuous-time analogs, which (as in discrete time) are different from one another. This is again due to Jensen's inequality. Unlike in discrete time, however, only the local expectations hypothesis is acceptable as a statement of equilibrium because the expected returns under each of these hypotheses are not consistent with those implied in a general equilibrium, as noted by Cox et al. (1981).

Preferred Habitat Hypothesis

Crucial alternatives to the expectations theory of the term structure of interest rates are theories that add an element of risk when conferring the expected rate of return for bonds of different maturities; that is, the indifference assumption that was stated earlier no longer holds. If the investment horizon is one year, it does make a difference whether to invest in a one-year instrument, a two-year instrument sold after one year, or two sequential six-month instruments. The preferred habitat theory argues that we first must know the investment horizon to determine relative risk among bonds. In the simple example, the horizon is one year. The one-year instrument is safest for this horizon. Under the preferred habitat theory, the investor would require a higher rate of return on both the two-year and six-month instrument.

Liquidity Preference Hypothesis

The liquidity preference theory can be considered a special case of the preferred habitat theory. Here, it is held that investors demand a risk premium as compensation for holding longer-term bonds. In addition, since the variability of price increases with maturity, the risk premium demanded by investors increases. As a special instance of the preferred habitat theory, the liquidity preference theory says that as all investors have a habitat of a single period, the shortest-term bond is judged safest.

With each of these theories, one can assess their efficacy only in the context of the general economy. Specifically, we assume that the economy is one in which investors have an inclination to consume, as well as to invest (in fact, even in a diverse set of risky investments). With a specification of utility of consumption and wealth, as well as a formal expression for risk aversion, the risk-based term structure theories can be viewed in the context of markets. Given that risk-based term structure theories can be viewed in the context of a defined market, the following conclusions can be made.

Term premiums are monotonic in maturity (or term). Interest-rate risk is inherently intertemporal. That is, it is a multiperiod phenomenon, in which an unexpected interest-rate change at any period affects all future returns and risk compounds over time. The traditional notion of preferred habitat seems difficult to reconcile with real markets. As it turns out, the traditional notion omits the importance of risk aversion. As we incorporate a varying need to hedge against interest-rate changes, the theory converges to a more acceptable view of markets. The generalization of these economic analyses has led to what has been called an *eclectic theory* of the term structure that recognizes and accommodates the many factors that play a role in shaping the term structure. Expectations of future events, risk preferences, and the characteristics of a variety of investment alternatives are all important, as are the individual preferences (habitats) of market participants about the timing of

their consumption. It is this eclectic theory that one needs to embrace in the development of the dynamic term structure.

CONTINUOUS-TIME MODELS OF THE TERM STRUCTURE

Now we discuss how the earlier concepts of discount function, spot rate, spot yield, and forward rate have their analogies in the continuous-time domain. It will be seen that while the mathematics are slightly more complex, the roles that each of these quantities play in the term structure of interest rates remain unchanged.

In summary, the priced-based representation of the term structure, or the discount function, facilitates both the mathematical formulation of the problem and its subsequent solution. Once the term structure equation is solved explicitly in terms of price, the price/yield equation (in continuous time) is used to convert the term structure to its equivalent representation in terms of yield.

Given the intertemporal nature of the term structure and the apparent efficiency of the market to incorporate information, it is assumed that the market acts instantaneously, and that a period in time is but an instant. This is the underlying premise for continuous-time models in economics and finance.

Traditional fixed-income analysis assumes that compounding occurs at discrete points or over finite intervals, typically on a semiannual basis. However, as the compounding period grows ever shorter, discrete compounding is replaced by continuous compounding. We expand our original equation (2) for the present value (at t), $PV(t, T)$, of a cash flow received T years from today, $CF(t, T)$, which is invested at the spot yield, $R(t, T)$, to be

$$PV(t, T) = CF(t, T)e^{-TR(t, T)} \quad (16)$$

Equation (16) is the fundamental price/yield relationship for the case of continuous

compounding of a discount bond and is the direct analog of the price/yield relationship shown in equation (2) for discrete compounding.

DISCOUNT FUNCTION

For a pure discount bond that pays one dollar at maturity, $CF(t, T) = 1$. Let P be the price of the pure discount bond. Thus equation (16) becomes

$$P(t, T) = e^{-TR(t,T)} \tag{17}$$

Combining the above with equation (16), which equates the price of a discount bond to the discount function, we obtain

$$P(t, T) = e^{-TR(t,T)} = d(t, T) \tag{18}$$

Equation (18) provides an expression for the relationship between the discount function d and the spot yield R , and is the continuous-time analogy to equation (3).

Spot Rate

In the previous section, the spot rate was defined as the one-period rate of return. Under continuous compounding, the spot rate r is defined as the continuously compounded instantaneous rate of return. Stated another way, the spot rate is the return on a discount bond that matures in the next instant. The spot rate is really an expression of the concept that a discount bond with a specified term-to-maturity and yield is equivalent to a series of instantaneously maturing discount bonds that are continuously reinvested at a rate r until the final term T . This is discussed in the following section.

Spot Yield

If the spot rate is a known function of time, then a loan amount W that is invested at the spot rate r will grow by an increment dW that is given by

$$dW(t) = W(t)r(t)dt \tag{19}$$

where

$dW(t)$ = incremental increase in the value of the loan from time t to time $t + dt$

$W(t)$ = value of loan at time t

$r(t)$ = spot rate at time t

To find the value of the loan W at maturity, integrate equation (19)

$$\int_t^{t+T} \frac{DW(\tau)}{W(\tau)} = \int_t^{t+T} r(\tau)d\tau$$

$$W(t) = W(t + T) \exp\left(-\int_t^{t+T} r(\tau)d\tau\right) \tag{20}$$

If W is a discount bond, $W(t)$ is equal to the present value $P(t, T)$ and the value of $W(t + T)$ is one. Equation (20) is rewritten as

$$P(t, T) = \exp\left(-\int_t^{t+T} r(\tau)d\tau\right) \tag{21}$$

From equation (17), the price P is expressed in terms of its spot yield R . By equating (17) and (21), we obtain the following expression for the *spot yield* in terms of the spot rate

$$R(t, T) = \frac{1}{T} \int_t^{t+T} r(\tau)d\tau \tag{22}$$

Equation (22) is a general expression that always holds.

Another view of the relationship between the spot yield and the spot rate is that instead of continuously reinvesting at the spot rate r for a fixed maturity T to obtain the spot yield R , if the term-to-maturity grows ever shorter, the spot yield R approaches the spot rate r "in the limit." r may be stated as

$$r(t) = R(t, T = 0) = \lim_{T \rightarrow 0} R(t, T) \tag{23}$$

Graphically, the spot rate at $t = 0$ may be visualized as the yield corresponding to the point at which the spot-yield curve intercepts the yield axis.

FORWARD RATE

The forward rate, $F(t_0, t)$ is the marginal rate of return for extending an investment to an additional increment of term at $t > t_0$. The forward rate is defined by

$$R(t, T) = \frac{1}{T} \int_t^{t+T} F(t, \tau) d\tau \quad (24)$$

Comparing the above notations for the forward rate with that in equation (4), note that the parameter “1” from the previous parameter set (denoting one time period) is no longer present. In the continuous-time domain, one time period collapses to just an instant.

Rearranging and applying Leibniz’s rule, the above becomes

$$\begin{aligned} \frac{d}{dT} [TR(t, T)] &= \frac{d}{dT} \int_t^{t+T} F(t, \tau) d\tau \\ &= F(t, t + T) \\ &= F(t, s) \end{aligned} \quad (25)$$

where s is the maturity date. The above equations relate the forward rate to the spot yield R . As with the case of discrete compounding, the forward rate may be expressed similarly in terms of the discount function $d(t, T)$ or the spot rate $r(t)$.

From equations (17), (18), and (25),

$$F(t, t + T) = \frac{-d}{dT} \ln [d(t, T)] \quad (26)$$

where $\ln[\]$ is the natural logarithm.

Separately, from equations (22) and (24),

$$\begin{aligned} r(t) &= \lim_{T \rightarrow 0} R(t, T) \\ r(t) &= \lim_{T \rightarrow 0} R(t, T) \frac{1}{T} \int_t^{t+T} F(t, \tau) d\tau \\ &= \lim_{T \rightarrow 0} \frac{1}{T} F(t, \hat{t}) T \quad (t < \hat{t} < t + T) \\ &= F(t, t) \end{aligned} \quad (27)$$

Under a certain economy, equations (22) and (27) show that the spot rate needs to be equal to the forward rate to preclude arbitrage. In the case in which the spot-yield curve $R(t, T)$ (and consequently the term structure) is defined, it follows that the spot rate needs to be equal to

the instantaneous forward rate over the term of the discount bond for equation (27) to hold true (see equation (7) for the analogy in the case of discrete compounding).

Since R is the yield of a discount bond and the term structure of interest rates is the set of spot yields as a function of maturity, equation (22) defines the term structure when the evolution of the spot rate is a known function of time. However, in general, the spot rate is not known; only the current spot rate is known from the current spot-yield curve. Nevertheless, term structure theory expands the basic relationship that is shown in equation (22), namely that the yield of a discount bond is a function of the spot rate. This is discussed in more detail in the next section when the spot rate assumes the form of a stochastic differential equation.

TERM STRUCTURE IN CONTINUOUS TIME

As stated in the previous section, the term structure of interest rates describes the relationship between the yields of default-free, zero-coupon securities as a function of maturity. Consequently, the term structure may be envisioned as a continuous set of yields for zero-coupon securities over a range of maturities.

Equation (18) describes the price/yield relationship for a single zero-coupon bond of a given maturity. As the term-to-maturity T spans the range of possible maturities within the term structure, the associated spot yields are generated for each maturity point, that is, R is a function of the term T . Furthermore, for any one value of T , the spot yield will vary as a function of the time t . In general, the spot yield R is a function of the term-to-maturity T , the time t and the spot rate r (as shown by equation (22)). R may be expressed as

$$R = R(r, t, T) \quad (28)$$

Equation (28) describes the functional form of the term structure in terms of the spot yield R . In order to describe the term structure completely,

an equation is needed that mathematically specifies the form of the relationship between the spot yield R and the term T over time t .

Such an equation for the term structure may be found by considering that the term structure may be expressed equivalently in terms of the prices of discount bonds (i.e., through the discount function). Thus equation (17) may be rewritten as

$$R(r, t, T) = -\frac{1}{T} \ln[P(r, t, T)] \quad (29)$$

where $\ln[\]$ is the natural logarithm.

If an expression for $P(r, t, T)$ can be found that defines the value of a zero-coupon bond at different points in time and for varying terms T , then the term structure of interest rates has been defined fully. Alternatively, equation (29) provides an equivalent description of the evolution of the term structure over time in terms of the spot yield.

KEY POINTS

- There are three main static models for the term structure of interest rates: the spot yield curve, the discount function, and the curve of implied forward rates; straightforward trans-

formations allow moving from one model to the other.

- These representations exist in both discrete-time and continuous-time versions and may be readily constructed from market data.
- Static models of the term structure suit valuation and comparisons of fixed-income instruments for which there is no dependency (contingency) on future events.
- Even though implied forward rates provide an arbitrage-free forecast for the future course of interest rates, static models do not admit uncertainty about the future.
- There are three main explanations for the future course of interest rates in equilibrium: the expectations hypothesis, the preferred habitat hypothesis, and the liquidity preference hypothesis.

REFERENCES

- Cox, J. C., Ingersoll, J. E., Jr., and Ross, S. A. (1981). Re-examination of traditional hypotheses about the term structure of interest rates. *Journal of Finance* (September): 769–799.
- Vasicek, O. A., and Fong, H. G. (1982). Term structure modeling exponential spline. *Journal of Finance* (May): 339–348.

The Dynamic Term Structure Model

DAVID AUDLEY, PhD

Senior Lecturer, The Johns Hopkins University

RICHARD CHIN

Investment Manager, New York Life Investments

PETER C. L. LIN

PhD Candidate, The Johns Hopkins University

SHRIKANT RAMAMURTHY

Consultant, New York, NY

Abstract: The term structure of interest rates represents the cost of (return from) borrowing (lending/investing) for different terms at any one moment in time. The term structure is most often specified for a specific market such as the U.S. Treasury market, the bond market for double A rate financial institutions, the interest rate market for LIBOR and swaps, and so on. The term structure is usually specified via a rate or yield for a given term or the discount to a cash payment at some time in the future. These are often summarized mathematically through a wide variety of models. In addition, term structure models are fundamental to expressing value, risk, and establishing relative value across the spectrum of instruments found in the various interest-rate or bond markets. Dynamic models of the term structure are characterizations that are specifically established to consider future market scenarios where there is uncertainty. As such they are rooted in probability, stochastic process, and martingale theory. Standard models include those derived from assumptions that include a short-rate or a forward rate process as an explanatory factor for the evolution of markets. Instantiations of these models include a general zero-coupon bond pricing equation and the LIBOR market model. An important consideration includes expressing the market price of risk that allows for the complexity of the term structure of interest rates to exist without arbitrage, as found from the traded markets. This consideration provides a platform to analyze bond and interest rate derivatives in the risk-neutral setting or with a real-world/objective probability measure.

Modern financial markets are predicated on the notions of contingency and uncertainty. Many recent financial innovations are directed at coping with the uncertainty of markets and the contingency of obligations. As part

of this evolutionary process, dynamic models of securities and their behavior in the markets are at the forefront of financial economic research and application. In the fixed-income markets, this condition dominates and

drives the need for dynamic term structure models.

The dynamic term structure model of a market sector, as defined by a reference set of securities, is a mathematical set of relationships that can be used to characterize any security in that market sector in which market uncertainty dominates the expected timing and receipt of cash flows. There are several qualitative essentials that need to be accommodated by a useful modeling approach. The ability to value fixed-income securities at any point in time (present or future) for conventional or forward settlement is a necessary first step. This is especially true in the valuation of compound or derivative instruments. Indeed, before the value of a bond option may be determined, the ability to calculate the (probabilistic) expected value of the bond on the future exercise date (conditioned on current market condition) is needed. Complementing this, reasonable variations from this expectation also need to be determined and weighed relative to the expected outcome. It is essentially this same idea that allows for the analysis of a futures contract, an interest-rate cap, or an option on a swap. In addition, to determine the performance risk that results from market moves, a rationale for incorporating market changes needs to be embedded into the modeling process.

With these premises in mind, the following assertions regarding dynamic models for the term structure of interest rates are postulated:

- The model must have the capability to extrapolate into the future an *equilibrium* evolution of the *term structure of interest rates*, given its form on a specified day, and must preclude riskless arbitrage.
- The model must allow a probabilistic description of how the term structure may deviate from its expected extrapolation while maintaining the model's equilibrium assumption.
- The model must embody a rationale to incorporate perturbations from the equilibrium that correspond to the economic fundamentals that drive the financial markets.

A technical discussion of *term structure* models is really equivalent to a discussion of the (zero-coupon or) spot-yield curve. The theory of the term structure of interest rates focuses on a term structure model that models the movement of the spot (zero-coupon) yield over time. Such term structure models are developed where any coupon-paying bond may be viewed in terms of its constituent zero-coupon bonds and analyzed in the context of this term structure model.

In this entry we focus on arriving at dynamic term structure models that respond to these imperatives. We first describe a dynamic term structure model in the case of objective (or real-world) probability measures. The assumptions, derivation, and parameterizations of the general model are described. We then indicated how this dynamic term structure model represents zero-coupon bonds, coupon-paying bonds, and determines par-coupon and horizon yield curves. It can also be used to model option-laden bonds and derivatives. The key feature of this model is dependence on a short-rate model as the (single) explanatory factor.

Next, a dynamic term structure model in a risk-neutral measure is presented. It is here that connections between the risk-neutral and the real-world setting are made; the importance of the forward rate model as the key explanatory factor is identified; and the implementation of computational imperatives in the context of applying the model to interest rate derivatives are identified.

KEY ELEMENTS IN A DYNAMIC TERM STRUCTURE MODEL

The following key ideas guide the development of dynamic term structure models:

- *Equilibrium*
- *Arbitrage-free*
- *Continuous time/continuous state*
- *Spot rate/forward rates* as underlying variable
- *Completeness of markets*

These five principles not only provide an elegant mathematical formulation of the term structure of interest rates, but also one that is applicable to a number of different market sectors and situations. Later we look at alternatives to the spot rate as the underlying variable and introduce a concept that highlights the market-clearing consequence of equilibrium—namely, the consensus of a fair market as embodied in the idea of a martingale in probability theory and forward rates as the underlying variable.

EQUILIBRIUM

General equilibrium models of the economy describe the basic workings of the macro economy as a function of a given “state variable.” This implies that the production processes and assets that constitute the economy are determined by the value of the state variable. Cox, Ingersoll, and Ross (CIR; 1985) showed that this general equilibrium model of the economy may be used to derive a model for the term structure of interest rates in terms of this state variable. Such an approach is considered to be a general equilibrium model of interest rates in that the interest-rate model is a consequence of a general economic model.

In contrast to general equilibrium models, “partial equilibrium” models assume a particular form of the interest-rate process as a given. This type of approach does not require the particular interest-rate process to be a result of some greater underlying theory. Examples of partial equilibrium models are those of Vasicek (1977), Ho and Lee (1986), and Black, Derman, and Toy (1990), among others. In addition, partial equilibrium models are calibrated exogenously to the current term structure of interest rates. Without this exogenous information, partial equilibrium models cannot quantify the term structure.

On the other hand, general equilibrium models theoretically can specify a term structure independently of any bond-market information. It has been observed, though, that such a term structure (as provided by earlier general equi-

librium models) may not be consistent with the entire market term structure. For this reason and due to the difficulty that some term structure practitioners have had in quantifying the parameters in the CIR model, many implementers of term structure models have pursued the development of partial equilibrium models.

We approached these issues in the development of this term structure model in a variety of ways. While the model described herein is not purely a general equilibrium model, we began with the basic CIR model as a starting point and then further generalized that model’s stochastic interest-rate process. Furthermore, we developed an approach for the specification of CIR-type model parameters such that the derived term structure was consistent with the observed market term structure. Thus, drawing upon a cornerstone in term structure theory, we developed an extension to the CIR model that can be readily applied to the financial marketplace.

ARBITRAGE-FREE

One underlying principle that the term structure model under discussion shares with many of the above-mentioned references is that the term structure is *arbitrage-free*. This concept, an extension of the arbitrage-free principles found in the Black-Scholes options theory for commodity and equity markets, states that the term structure observes a given relationship among its constituent parts and that purely arbitrary yield-curve shapes do not occur. Given today’s yield curve, subsequent yield curves are assumed to evolve in a “rational” manner that precludes riskless arbitrage. This indicates that the prices of bonds defining the yield curve move in such a way that it is not possible to create a portfolio of securities that always will outperform another portfolio without entailing any risk or net investment; in other words, there is no “free lunch.” The arbitrage-free principle plays an important role in the mathematical pricing of fixed-income securities.

CONTINUOUS TIME/CONTINUOUS STATE

Another distinguishing feature of this term structure model is the strict adherence to the *continuous-time/continuous-state* approach to the modeling of stochastic processes. This assumes that interest rates and bond prices move in a continuous fashion over time, rather than in discrete jumps. Thus a spot-yield curve may be found for any point in time during the life of a bond, rather than only at specific points (such as a coupon payment date). This concept is consistent with the notion of a continuous yield curve and allows for the use of continuous stochastic calculus.

Continuous Probability Distributions

Furthermore, the generality of the transitional probability density function, as a complete specification of the statistical properties of the rate process, is maintained throughout the term of the bond. This is in contrast to the common approach of describing individual sample paths or scenarios, as found in Monte Carlo approaches to security analysis. The ability to extend the analyses to compound, derivative instruments is unimpaired through the use of this transitional probability density function. Moreover, the continuous-time/continuous-state approach avoids the computational issues associated with the number of sample paths analyzed. Since the complete specification of the statistical properties is maintained, it is as if an infinite number of sample paths are run.

Numerical Solution Technique

The computer numerical solution technique that accompanies the continuous-time formulation is one that is well known in the engineering and physical sciences as the Crank-Nicholson finite-difference method for the solution of partial differential equations (PDEs).

This solution technique has been used extensively in the study of aerodynamics and fluid flow, and has the flexibility to focus its computational efforts in areas that require greater numerical precision, such as the time period surrounding an option exercise period. This is in contrast to binomial interest-rate lattices, which are constrained to jump, for example, in six-month intervals, such as in some commercially available applications.

COMPLETENESS OF MARKETS

One of the key ideas in developing financial models—especially term structure models—is formulating valuation in the context of a replicating portfolio. That is, for a given structure, a portfolio is formed that replicates or hedges the instrument with the same risk-return properties. Then the replicating portfolio dictates the value of the given structure. Otherwise, a self-financing riskless arbitrage can be engaged. Presumably, price convergence would result given sufficient market awareness. Essentially, a market is *complete* if this can be always done with a certain characterization of uniqueness.

DYNAMIC TERM STRUCTURE MODEL

The formulation and implementation of the term structure model needs to be completely general so as to be applicable across a broad range of fixed-income markets in a straightforward and consistent manner. For example, once the value of the fixed-income instrument is found, the value of its derivative (such as its futures contract) also may be found. Furthermore, it is possible to value the quality and delivery options within the bond futures contract. These effects also can be incorporated when valuing an option on the bond futures contract.

General Assumptions

The analytical model that describes spot-rate movement is a one-factor, mean-reverting, diffusion process model. The model assumes:

1. The evolution of interest rates is a continuous process and may be described by a single variable, that is, by the instantaneous spot rate, which is the return on an investment over an infinitesimally short period of time. This allows for the use of continuous-time mathematics, which requires greater technical sophistication, but which increases the flexibility of the mathematical modeling process.
2. The model assumes that interest rates move in a random fashion, which is known as Brownian motion or a Wiener process. The Wiener process has been used in the physical sciences to describe the motion of molecular particles as they diffuse (or spread) over time and space.
3. The term structure of interest rates is assumed to be represented by a Markov process, which states that the future movement in interest rates depends only on the current term structure and that all past information is embodied in the current term structure.
4. The term structure is arbitrage free in that a portfolio of securities derived from the term structure is constrained to have an instantaneous rate of return that is equal to the risk-free rate. Future movements in interest rates are similarly constrained so that the possibility of riskless profit is precluded. This implies that there are a sufficient number of sophisticated investors who will take advantage of any temporary mispricing in the marketplace, thus quickly diluting any arbitrage opportunities that exist.

Technically, an arbitrage-free term structure indicates that a portfolio of securities derived from the term structure may be constructed such that the portfolio instantaneously returns the risk-free rate. Since the above holds true for

any arbitrary set of maturities in this portfolio of securities, it is said to be true for all maturities. This indicates that all securities that comprise the term structure are related in a common fashion. This commonality is expressed through the concept of the market price of risk, which is the incremental return over the risk-free rate that is required for incurring a given amount of additional risk. In this context, risk is measured by the variance of a bond's rate of return. A result of the arbitrage-free nature of the term structure is that all securities share the same market price of risk. As we demonstrate at the end of the entry, the risk premium is one component of the market price of risk.

1. The price of a default-free, zero-coupon (discount) bond at any point in time continuously depends on the spot rate, time, and maturity of the bond. This models the interaction between the bond's price and the probabilistic movement in the spot rate. This is an extension of the point discussed earlier that stated the yield of a discount bond is a function of the spot rate.
2. The market is efficient in that all investors have the same timely access to relevant market information. Furthermore, investors are rational and there are no transaction costs.

SPOT-RATE MODEL

As a result of assumptions 1 through 3 above, the equation that describes the diffusion process for the movement in the spot rate is given by equation (1)

$$dr = k(\theta - r)dt + \sigma\sqrt{r}dz \quad (1)$$

where

r = spot rate, the instantaneous rate of return

dr = infinitesimal change in the spot rate

k = mean reversion constant

θ = "target" spot rate, which will be expressed as a function of time

dt = infinitesimal change in time

σ = volatility of r
 dz = infinitesimal change in the random variable z (a characterization of the Weiner process)

There are many alternatives to the form (1) (see, for example, Hull, 2009) and while this model has some attractive features, we in no way argue that it is “best.” It is just useful and has been shown to work well in practice. Its features include the following.

Mean Reversion

Equation (1) states that the rate r changes with respect to time and the degree of randomness. The first term on the right-hand side of equation (1) states that the “drift” in the spot rate over time is proportional to the difference between the rate r and θ . As r deviates from θ , the change in r is such that r has a tendency to revert back to θ , a feature that is known as *mean reversion*. The presence of mean reversion imposes a centralizing tendency such that rates are not expected to go to extremely high or low levels. In addition, mean reversion precludes the existence of negative interest rates in our interest-rate model, given that the initial interest rates are positive.

One can easily derive a closed-form expression for θ as a function of time. Note that θ is not assumed to be constant, which is usually the case for the traditional CIR approach.

Effect of Randomness

The second term on the right-hand side of equation (1) states that the contribution to the change in r due to randomness is driven by movements in the random variable z . The variable z is normally distributed with a mean of zero and a variance that is proportional to time. This indicates that the amount of random “noise,” as represented by the variable z , may be any positive or negative value, but that its expected value is zero. In addition, as time passes, the variance increases so that the “amplitude” of

the noise also increases. The variables σ and r , which are coefficients of dz in equation (1), show that the change in r also depends on the level of volatility and interest rates. The variable z has its own defined level of uncertainty so that as volatility and rate change, the overall degree of uncertainty is influenced by the level of these variables.

Endogenous Parameterization (Tuning the Model)

Equation (1) describes the rate in terms of the parameters k , σ , and θ . The volatility parameter σ is specified externally so that it reflects either the historical level of volatility or the volatility that is currently present in the market. Secondly, θ reflects the current term structure such that the future movements in r are influenced by today’s term structure. Finally, the mean reversion constant k determines the speed of adjustment of r back to θ . In order for the interest-rate model to be of any utility, the parameter k is chosen to be consistent with the observed market prices of bonds comprising the current yield curve, while θ is derived directly from the current yield curve. This process of determining k and θ “parameterizes” the model to the observed yield curve.

There are several variations of equation (1) that exist within the academic literature that appear to be similar to equation (1); see, for example, Chan et al. (1992). However, the details surrounding the functional form of each term in equation (1) and the associated parameterization process can result in very different models. The specification of parameters for this term structure model is driven by the requirement to be able to precisely reprice the set of securities that constitute the reference yield curve. A properly calibrated term structure model needs to be able to define a bond whose cash flow characteristics match those of an on-the-run issue exactly and then have the price of that constructed bond match exactly the market price of the Treasury issue. By repeating this process

for each of the on-the-run issues, the mean reversion constant and the *risk premium* that are appropriate over the range of reference issues may be quantified.

As a technical side note, the term structure model needs to satisfy internal consistency checks, and the parameter specification process plays a part in the internal system for checks and balances. For the set of chosen parameters, the price furnished by the term structure model—as the solution to a PDE—needs to be equal to that provided by applying the discount function to the cash flows of the specific on-the-run issue, as explained earlier. Thus the discount function is a direct means of verifying the results of the term structure model. In fact, the PDE may be decomposed into two coupled ordinary differential equations (ODE) in the absence of any embedded options. Thus prices obtained from the PDE, ODE, and discount-function approaches all need to be identical.

Calculation of the Spot Rate

The solution to equation (1) is obtained through computer numerical solution techniques and accounts for the current value of the spot rate (as an initial condition) and its level of volatility. As time moves forward, the solution expresses the probable distribution of the spot rate as the spot rate propagates through time. Thus, at any point in time, it is possible to calculate the probability distribution of the spot rate. It was discussed previously that the price of a bond depends on the spot rate so that the spot-rate probability distribution is also the probability distribution for the bond price. This is useful in calculating the probability that an embedded call or put option will be exercised, which is the probability that the price of a particular bond is greater than or less than, respectively, the specified strike price at exercise.

The calculation of the probabilities is made possible by assuming a specific mathematical form for the random variable z , or a Wiener process. Generally, a probability distribution func-

tion is described by its mean and variance as functions of time. If these quantities are known, then the probability of different spot rates is known. The Wiener process assumption states that the statistical variance for the random variable z varies with the length of time under consideration. As time increases, the variance of z also increases. The known change in the variance of z is subsequently translated (in a known fashion) to the change in the variance of the rate r , which may be used to obtain the desired probability in terms of r . In general, we use the solution of the Kolmogorov (forward or backward) equation to establish an expression for the probability density of the short rate.

BOND-PRICE VALUATION MODEL

As a consequence of assumptions 4 and 5 (the price of a default-free discount bond depends continuously on the spot rate), it can be shown that the price of a discount bond of term T is expressed as

$$\frac{\partial P}{\partial t} = rP - [k(\theta - r) + \lambda\sigma r] \frac{\partial P}{\partial r} - \frac{1}{2}\sigma^2 r \frac{\partial^2 P}{\partial r^2} \quad (2)$$

where

- P = price of zero-coupon bond for time t and rate r
- $\partial P/\partial t$ = partial derivative of price with respect to time
- $\partial P/\partial r$ = partial derivative of price with respect to rate
- $\partial^2 P/\partial r^2$ = second partial derivative of price with respect to rate
- λ = "risk premium"

The "risk premium" is the variable that represents the additional return over the risk-free rate that the market requires for holding a longer-term instrument. This is determined from the current term structure. In addition to the bond price equation, to represent the behavior of the instrument, boundary

conditions on the solution to (2) need to be prescribed. These conform to given circumstances, but in the simplest case, they include cash flows and constraints on P as r converges toward zero or becomes arbitrarily large.

Developing the Bond-Price Equation

A development of the *bond-price valuation model* (for the *zero-coupon bond*) follows in a straightforward manner. Arguments of variables are suppressed except when needed to clarify dependencies.

Equation (1) describes the process for the propagation of the spot rate and is given by

$$dr = k(\theta - r)dt + \sigma\sqrt{r}dz$$

If we assume that P is a function of the two variables r and t expressed as the following $P = P(r, t)$, then Ito's lemma (see Shreve, 2004) provides that

$$dP = \left[k(\theta - r) \frac{\partial P}{\partial r} + \frac{\partial P}{\partial t} + \frac{1}{2} \sigma^2 r \frac{\partial^2 P}{\partial r^2} \right] dt + \sigma\sqrt{r} \frac{\partial P}{\partial r} dz$$

To apply the principal of an arbitrage-free term structure, consider the representation of evolutions of the price to be

$$dP = \mu P dt - \rho P dz$$

where

$$\mu = \frac{1}{P} \left[a \frac{\partial P}{\partial r} + \frac{\partial P}{\partial t} + \frac{1}{2} b^2 \frac{\partial^2 P}{\partial r^2} \right]$$

$$\rho = -\frac{1}{P} b \frac{\partial P}{\partial r}$$

Any security W_i with maturity s_i is subject to the same relationship such that

$$dW_i = \mu_i W_i dt - \rho_i W_i dz$$

Consider a portfolio W consisting of owning an amount of W_1 and shorting an amount of W_2 such that

$$W = W_2 - W_1$$

where

$$W_2 = \left[\frac{\rho_1}{\rho_1 - \rho_2} \right] W$$

and

$$W_1 = \left[\frac{\rho_2}{\rho_1 - \rho_2} \right] W$$

Thus

$$dW = dW_2 - dW_1$$

Substituting for dW_1 and dW_2 yields

$$dW = \left[\frac{\mu_1 \rho_2}{\rho_1 - \rho_2} \right] W dt - \left[\frac{\rho_2 \rho_1}{\rho_1 - \rho_2} \right] W dz - \left[\frac{\mu_1 \rho_2}{\rho_1 - \rho_2} \right] W dz + \left[\frac{\rho_1 \rho_2}{\rho_1 - \rho_2} \right] W dz = \left[\frac{\mu_2 \rho_1 - \mu_1 \rho_2}{\rho_1 - \rho_2} \right] W dt$$

Since the stochastic element dz disappears, the rate of return on the portfolio W is equal to the riskless rate r . Therefore,

$$dW = rW dt$$

where we see it must be that

$$r = \frac{\mu_2 \rho_1 - \mu_1 \rho_2}{\rho_1 - \rho_2}$$

This gives the following relationship

$$r \rho_1 - r \rho_2 = \mu_2 \rho_1 - \mu_1 \rho_2$$

or, equivalently,

$$\frac{\mu_2 - r}{\rho_2} = \frac{\mu_1 - r}{\rho_1}$$

Since the maturities s_1 and s_2 were chosen arbitrarily, the above is true for any maturity s . Therefore, the term

$$\frac{\mu - r}{\rho}$$

is not a function of maturity and may be written as

$$\frac{\mu - r}{\rho} = q(t, r)$$

where $q(t, r)$ is the market price of risk.

Applying separation of variables, we choose $q(t, r)$ to be the following

$$q(t, r) = \lambda(t) \sqrt{r}$$

where $\lambda(t)$ is the risk premium, which can be shown to be

$$\lambda(t) = \frac{1}{2} \frac{\sigma}{k} \left[1 - e^{-kt} \right]$$

(As the term extends, the premium is higher.)

We see, therefore, that

$$\frac{\mu - r}{\rho} = q(t, r) \Rightarrow \mu = r + \lambda(t)\sqrt{r}\rho$$

or that the expected return of a bond is equal to the riskless rate plus another term related to the risk premium.

With $\rho = -\frac{1}{P}b\frac{\partial P}{\partial r}$, the above becomes

$$\mu = r + \lambda\sqrt{r} \left(-\sigma\sqrt{r} \frac{\partial P}{\partial r} \frac{1}{P} \right)$$

Substituting the above into $dP = \mu P dt - \rho P dz$ gives (where $\frac{\partial P}{\partial r} < 0$)

$$dP = \left(r - \lambda\sigma r \frac{\partial P}{\partial r} \frac{1}{P} \right) P dt - \rho P dz$$

Equating the coefficients of dt between the above and

$$dP = \left[k(\theta - r) \frac{\partial P}{\partial r} + \frac{\partial P}{\partial t} + \frac{1}{2} \sigma^2 r \frac{\partial^2 P}{\partial r^2} \right] dt + \sigma\sqrt{r} \frac{\partial P}{\partial r} dz$$

gives

$$\frac{\partial P}{\partial t} = rP - [k(\theta - r) + \lambda\sigma r] \frac{\partial P}{\partial r} - \frac{1}{2} \sigma^2 r \frac{\partial^2 P}{\partial r^2}$$

where, at maturity, we have the boundary condition

$$P(r, t) = 1$$

This completes the derivation of equation (2).

Next, if we assume a separation of variables for $P(r, t)$ of the form

$$P(r, t) = \exp [C(t) - B(t)r]$$

it can be derived that the target spot rate, $\theta(t)$, is of the form

$$\theta(t_0 + T) = -\frac{d}{dT} \ln d(t_0, T) - \frac{1}{k} \frac{d^2}{dT^2} \ln d(t_0, T)$$

or

$$\theta(t_0 + T) = F(t_0, t_0 + T) + \frac{1}{k} \frac{d}{dT} F(t_0, t_0 + T)$$

which will provide a solution to equation (2) that will exactly reprice the reference set where the discount function $d(t_0, T)$ and the forward rates $F(t_0, t_0 + T)$ are derived from the reference set using spline functions. Furthermore, this property is true for all volatilities when the above-specified risk premium is used.

THE TERM STRUCTURE

Equation (2) is a PDE whose solution is obtained through a numerical finite-difference technique. The solution gives the price P of the bond for different times and spot rates, and can be visualized as a three-dimensional surface for which the height of the surface is the price of the bond and the location of the point (i.e., longitude and latitude) is given by the time and spot rate. The solution takes into account that the bond's price is par at maturity, regardless of the level of interest rates. As the solution steps back from the maturity date, the price of the bond may be calculated for varying levels of the spot rate and the familiar price/rate graph may be drawn for this time-step. (Not all bond prices are equally likely to occur since interest-rate movements and the probabilities associated with these movements are described by equation (1).)

As the solution process continues backward from maturity to the present day, the theoretical price corresponding to today's spot rate can be calculated. Once the price behavior of a bond is known, the value of an option on that bond may also be calculated. In general, the expected value of the bond may be determined at any time from the present to maturity under the expectation operation over the solution to (2) and the probability density function for r .

Since the solution to equation (2) furnishes the price as a function of time and rate, equation (14) of the previous section may be solved to provide the zero-coupon yield for a bond with the term-to-maturity T . As the term T is varied, the entire term structure may be obtained.

(The obtained term structure, in general, can take a variety of shapes. If the current spot rate is below the current value of the long-term rate, θ , the obtained term structure will be upward sloping. If the current spot rate is substantially above the long-term rate, the obtained term structure will be inverted to downward sloping. For spot-rate values in between, the term structure will be humped, displaying both upward sloping and downward sloping segments. Thus an attractive feature of the term structure model is the ability to obtain term structure specifications that are consistent with those that have been observed historically.)

APPLICATIONS OF THE TERM STRUCTURE MODEL

We conclude this entry with a description of the application of the term structure model developed in the previous section in the valuation of fixed-income securities. For the simple case of noncallable bonds, many term structure models can be used to determine value. In fact, the spline-fit discount function is a very straightforward method of calculating the value of such a bond. However, when option-embedded bonds or compound instruments are considered, using the PDE approach is opportune to reflect the specific nature of the option features. As this entry demonstrates, the PDE-based term structure model is but the first step that leads to a greater assortment of analytical financial tools.

Zero-Coupon Bonds

Most yield curves, such as the U.S. Treasury curve, are expressed in terms of the yields of coupon-bearing bonds, not zero-coupon bonds. Thus a procedure is required to translate the current-coupon yield curve to an initial zero curve (i.e., the current term structure) expressed in terms of a spot-yield curve. One of several methods may be employed; see Vasicek and

Fong (1982). In summary, a reference set of securities is chosen to represent the yield curve, and each of the cash flows from this set of securities is treated as a zero-coupon bond that is part of the term structure. Since each of the reference securities has a known market price, the price/yield relationship, along with a curve-fitting process, is applied sequentially to each of the cash flows to derive the current term structure. This process establishes the set of initial conditions necessary to predict the evolution of the term structure.

If the actual zero-coupon yields are compared to the theoretical zero-coupon yields, then the richness or cheapness of the zero-coupon market may be gauged. Since the discount function may be constructed from any reasonable set of reference bonds, if the reference bonds consisted of off-the-run Treasury issues that are commonly stripped and/or reconstituted, then the corresponding theoretical zero curve should be indicative of the shape and level of the market strip curve.

Additionally, as the Treasury curve flattens or steepens, the theoretical zero curve changes accordingly to reflect the new shape of the Treasury curve. Consequently, as the Treasury curve steepens or flattens, the degree of anticipated yield-spread widening or tightening in the zero market may be estimated.

Coupon-Paying Bonds

While our discussion thus far applies mainly to the price of a zero-coupon bond, it is more common to encounter coupon-paying bonds. To value coupon-paying bonds, we simply sum the present values of each of the coupon payments to determine the price. As discussed earlier, each coupon is treated as an individual zero-coupon bond.

Determination of the Theoretical Fair Value

Once the term structure is defined, it may be used to value any collection of cash flows and

serves as the standard of fair value. The theoretical price of a security that is calculated in this manner may be compared to its actual market price. Any difference in price that results indicates whether the security is rich or cheap relative to its fair value. If the market price is equal to the fair value, then the security is said to be fairly priced.

Generally, Treasury securities are chosen to represent the basis for fair value and most securities (such as corporate and government-agency debt obligations) are cheap to Treasuries. However, if there are a sufficient number of securities from a particular sector or issuer, these issues may be used as the reference set of securities and a new yield curve may be defined to be the standard of fair value. Thus corporate, agency, or municipal debt issues may be compared to their own family of securities or to their own sector to determine their relative value within the specified sector.

Determination of Par-Coupon and Horizon Yield Curves

A *par-coupon* yield curve is a theoretical yield curve comprised of par-priced bonds along the maturity spectrum. Each of these par-priced bonds is constructed from the same discount function, which in turn is derived from a specified set of reference bonds. Since the discount function is defined continuously at different maturity points and cash-flow dates (via a spline-fitting procedure, for example), the par-coupon bonds corresponding to these same points may be determined.

The procedure for constructing a par-coupon bond involves an iterative process in which an initial coupon is assumed. For a given maturity date and associated coupon-payment dates, the cash flows and cash-flow dates are known for the assumed coupon level. The present value of each of the cash flows is found through the discount function, and the sum of the present values is compared to a price of par. The coupon then is varied until a par-priced bond is found. The process may be repeated for as many ma-

turity points as desired to construct an entire par-coupon yield curve.

A par-coupon yield curve is helpful in pricing bonds with off-the-run maturities. Often the question arises as to what exactly is the comparable Treasury yield when pricing off-the-run bonds. Depending on the fixed-income market sector, the comparable Treasury yield may be that of a specific Treasury note, or it may be an interpolated yield. The par-coupon curve provides a more technically rigorous means of calculating the interpolated yield, as opposed to a simple straight-line interpolation scheme.

Another application of the concept of the par-coupon yield curve is the *horizon yield curve*, the par-coupon yield curve for a future horizon date. Since the discount function may be determined as a function of time, the corresponding horizon yield curves at various points in time also may be found. The horizon yield curve is one way to help visualize how the present yield curve may evolve in the future in an arbitrage-free environment. (Of course, as new information is incorporated into the marketplace as time passes, the actual yield curve may deviate from the horizon yield curve. However, a horizon yield curve may still be calculated that reflects particular views about the future movements in both short-term and long-term rates.)

Yield-Curve Shocks and Shifts

The shape of the yield curve is governed by exogenous (*real-world*) factors. As the Federal Reserve alters its monetary policy, or as the inflation outlook changes, the yield curve responds accordingly. These perturbations to the curve can be characterized as “shocks” to short-term rates and as “shifts” to long-term rates. A shock can occur when there is a sudden and unexpected event that causes short-term rates to jump, even though the overall economic fundamentals have not changed.

The clearest example of a shock is the crash of 1987, during which investors fled to the safety of the Treasury market. During October 19, short-term rates dropped by approximately 90

to 100 basis points as investors sought a temporary safe haven. At the same time, long-term rates fell by about 20 to 30 basis points. Since the crash was a market phenomenon, rather than an altering of economic fundamentals, it is characterized as a shock to the system. (This is described mathematically within the term structure model as a change to the initial condition of the differential equation, where the differential equation remains the same. The solution to the differential equation shows how the entire yield curve responds to a shock in short-term rates.)

A shift in the yield curve results from a change in the economic landscape where federal budgetary concerns or inflation outlooks can affect the view on long-term interest rates. (In contrast to a shock, the term structure model represents a shift as a respecification of the parameters to the differential equation, while the initial condition has remained unchanged. The most general situation can consist of a combination of shocks and shifts.)

The basic premise underlying the shocked and/or shifted horizon yield curve is that the curve evolves in an arbitrage-free manner as prescribed by the term structure model despite alterations to the curve. Thus, even though a shock or a shift has occurred, the entire yield curve responds in such a way as to preclude arbitrage. As a result of different combinations of shocks and shifts of varying magnitudes, a series of horizon yield curves can be found for different yield-curve steepening and flattening scenarios.

TERM STRUCTURE OF FORWARD RATES

The financial markets can be viewed as a “game” with bids and offers between participants. To characterize fairness among the participants, the concept of a *martingale* (from probability theory) is introduced. Briefly, a martingale $M(t)$ is a stochastic process with finite first moment for any t and where

$$E [M(s)|F_t] = M(t) \text{ for } s > t$$

with F_t denoting that the conditioning is on a given filtration or data set. Additionally, a portfolio may be thought of as a quantity vector representing a particular set of positions (Øksendal, 2007). If the market is fair, then the discounted future value of any portfolio should be the same as today’s portfolio value when an appropriate discounting methodology is employed. However, in the objective (or real) world, equipped with the real-world measure, discount functions vary according to individual risk preferences, each associated with its own sector/market consensus. It is tedious to quantify these preferences for every case. So, instead of working under the real-world measure, we seek to explore an artificial probability measure under which every situation is risk-neutral. This probability measure is called the *risk-neutral* measure.

Modern pricing theory for financial derivatives is based on replicating a given derivative’s payoff by putting together a self-financing portfolio consisting of the underlying assets and *risk-free* bonds. By buying a derivative and selling its replicated portfolio (or vice versa), the self-financing portfolio is found to be risk-free. Constructing such a risk-free portfolio is beyond the scope of this discussion, but understanding and utilizing the existence and uniqueness of this replicating strategy is the key for what follows (see Björk, 2009). Next, we first examine the derivation of a *risk-neutral probability measure* from a forward-rate model. Then we look at a general *no-arbitrage* condition for the bond market. Finally, we address some practical issues and solutions in a conceptual fashion.

HEATH, JARROW, AND MORTON MODEL OF THE TERM STRUCTURE

Heath, Jarrow, and Morton (1992) proposed a general condition for no-arbitrage using the instantaneous forward-rate curve dynamics. The

instantaneous forward-rate is defined as

$$F(t, T) := -\frac{\partial \ln B(t, T)}{\partial T}$$

where $B(t, T)$ is the zero-coupon bond price at time t and maturity T . This stochastic process is usually written in a differential form

$$dB(t, T) = \alpha(t, T)dt + \sigma(t, T)dW(t)$$

where α and σ satisfy the usual conditions for an Ito process and $W(t)$ is a standard Brownian motion (under the real-world measure). Here, $F(0, T)$ is the initial *forward-rate term structure*. In many situations, instantaneous forward rates are fundamental building blocks for modeling fixed-income contingent claims. For example, a bond price process can be derived from Ito's lemma such that

$$\begin{aligned} \frac{dB(t, T)}{B(t, T)} = & \left[F(t, t) - \int_t^T \alpha(t, u)du \right. \\ & \left. + \frac{1}{2} \left(\int_t^T \sigma(t, u)du \right)^2 \right] dt \\ & - \int_t^T \sigma(t, u)dudW(t) \end{aligned}$$

Details can be found in Shreve (2004). Also, the money market account can be written as

$$\begin{aligned} \frac{dM(t)}{M(t)} &= F(t, t)dt \text{ (or equivalently,} \\ M(t) &= e^{\int_0^t F(u, u)du} \end{aligned}$$

A discount factor, $D(t) = M^{-1}(t)$, is defined similarly. A variation of this setting is one where we use the notation T to represent time-to-maturity (also called term). This alternative model is closer to the market reality because the curve won't shorten and will validate rolling-over trading strategies. For simplicity we set T to be maturity in the rest of this entry.

Let's first assume the existence of a risk-neutral probability measure, which is equivalent to imposing the local expectations

hypothesis, that is,

$$\widehat{E} \left[\frac{dB(t, T)}{B(t, T)} \middle| F_t \right] = F(t, t)dt$$

where the expectations $\widehat{E}[\cdot]$ is taken under this risk-neutral measure. Therefore the discounted bond-price processes $D(t)B(t, T)$ is a martingale for all T , that is,

$$\widehat{E} [D(s)B(s, T)|F_t] = D(t)B(t, T) \text{ for } t \leq s \leq T$$

This hypothesis also implies that the short rate evolves along today's instantaneous forward rate curve. Refer to Björk (2009) or Shreve (2004) for more details. Based on the martingale property we can then derive the HJM no-arbitrage condition shown in Heath et al. (1992) that

$$\alpha(t, T) = \sigma(t, T) \int_t^T \sigma(t, u)du$$

That is, the drift term of the instantaneous forward-rate curve process is tightly defined by the volatility term. This remarkable result tells us that only volatilities matter when modeling interest rates under a risk-neutral measure. Since the martingale property is imposed on all zero-coupon bonds to ensure fairness, arbitrage trades are precluded. If a pricing model is designed only for a derivatives pricing purpose, further investigation on risk premium is not necessary. This is an important point. For once the HJM no-arbitrage condition is applied to a particular model, the existence of a risk-neutral measure is assumed and the risk premium is zero. Nonetheless, not every modeler appreciates the consequence of ignoring the risk premium—especially when an asset and its derivative are priced congruently. For example, mortgage-backed derivatives usually involve prepayment statistics, which cannot be quantified under a risk-neutral measure, and the risk premium is usually given exogenously. The answer of which model should be used is based on the modeler's discretion involving calibration, implementation, and market assumptions, which we will talk about a bit more below.

MARKET PRICE OF RISK

Let the *market price of risk* be denoted by $\Theta(t)$. By the HJM no-arbitrage condition

$$\alpha(t, T) - \sigma(t, T)\Theta(t) = \sigma(t, T) \int_t^T \sigma(t, u)du$$

which shows that the risk premium can be written as

$$\Theta(t) = \frac{\alpha(t, T)}{\sigma(t, T)} - \int_t^T \sigma(t, u)du$$

If Θ exists, then the market is arbitrage-free. Moreover if Θ is unique, then the market is complete. For a multifactor model, completeness can be shown by nonsingularity of the volatility matrix. A remark can be made here that risk premiums are determined endogenously by the HJM no-arbitrage condition following from the local expectations hypothesis. This market price of risk identified in the HJM model is, however, a constant function of all maturities. The lack of flexibility limits the interest rate curve evolution under the real-world measure. In other words, if the curve dynamic is initially set up under a risk-neutral measure, then $\Theta(t)$ is usually impossible to find so that the “model-derived” real-world interest rates could satisfy the “real” real-world statistics.

BOND PRICING

When the market is assumed to be arbitrage-free and complete, zero-coupon bonds can then be derived under a unique risk-neutral measure that

$$\frac{dB(t, T)}{B(t, T)} = F(t, t)dt - \int_t^T \sigma(t, u)du d\tilde{W}(t)$$

The rate of return for any bond is the same as the short rate; nonetheless, the bond-price process is not Markov for a general forward-rate model. This result is critical when it comes to derivatives pricing since Monte Carlo simulation is often the only approach, and it can be slow and imprecise. Furthermore, no closed-form solution for bond dynamics can be

given, thus there is no closed-form solution for bond derivatives. Besides the computational issues due to the complexity in bond dynamics, the HJM framework cannot be used for log-normally distributed forward rates since, under the continuous compounding environment, the process “explodes” with positive probability. Therefore, practitioners seek eclectic methods to resolve the issues. A powerful tool invented for interest-rate derivatives pricing is the technique of “changing the numeraire,” discussed next.

CHANGE OF NUMERAIRE

The *numeraire* is a traded asset used for measuring value. Given a numeraire, all other prices are measured relative to this asset. In general, risk-neutral measures can have various forms in terms of different numeraires. For instance, if a money market account is used as a numeraire, it is the tradition risk-neutral measure as we see in the Black-Scholes option pricing setting. In a traditional risk-neutral world, the general evaluation form is written as

$$V(t) = \hat{E} [D(T)V(T)|F_t]$$

where $V(T)$ is the payoff of a contingent claim maturing at time T , and $V(0)$ is its price at time 0. Normally interest rates and underlying assets are assumed to be uncorrelated. This assumption makes the evaluation of the expectations above easier, but it is obviously invalid when a derivative V is based on interest rates. Further investigation in separating the derivative value process and the discount factor has been established by Geman et al. (1995).

In a traditional risk-neutral world, every discounted traded-asset price process is a martingale. When we take, for example, a zero-coupon bond with maturity T as our numeraire, the drift term of any other discounted traded-asset price process is adjusted according to this zero-coupon bond volatility. The new measure based on the zero-coupon bond numeraire is the T -forward risk-neutral measure. Consequently

we have

$$V(t) = B(t, T)E^T [V(s)|F_t]$$

where $E^T [\cdot]$ is the expectation under the T -forward risk-neutral measure. When the money market account is used as the numeraire, this adjustment to the drift term is unnecessary since the money market account process has zero volatility. In this new pricing equation the discount factor is taken out of the bracket and replaced with the zero-coupon bond discount. Therefore, the expectation is performed solely on the derivative V .

MARKET MODELS

For practitioners, the continuous compounding framework is unnecessary since most interest rates, such as LIBOR, for example, have only 1-week, 1-month, 3-month, 6-month, and 1-year investing intervals. Therefore, adopting the general no-arbitrage condition under the HJM framework, Brace et al. (1997) created a model for simple forward rates, which are compounded under a discrete-time framework. Based on the *change of numeraire* technique, forward rate processes are martingales under specific forward risk-neutral measures. This phenomenon can be justified via analyzing a bond portfolio used to create the payoff of a forward rate agreement: Let $\hat{F}(t, T, T + \tau)$ denote the process of a simple forward rate for the period $[t, T]$ with tenor τ . Then

$$\hat{F}(t, T, T + \tau) = \frac{B(t, T) - B(t, T + \tau)}{\tau B(t, T + \tau)}$$

Here $B(t, T + \tau)$ serves as the numeraire and transforms the traditional risk-neutral probability into a forward risk-neutral probability. By Ito's lemma, the forward rate dynamic can therefore be written as

$$\frac{d\hat{F}(t, T, T + \tau)}{\hat{F}(t, T, T + \tau)} = \gamma(t, T, T + \tau)d\tilde{W}^{T+\tau}(t)$$

where

$$\begin{aligned} \gamma(t, T, T + \tau) = & \frac{1 + \tau \hat{F}(t, T, T + \tau)}{\tau \hat{F}(t, T, T + \tau)} \\ & \times \left[\int_T^{T+\tau} \sigma(t, u) du \right] \end{aligned}$$

The main advantage of the *LIBOR market model* is set on the practical side. First, if γ are assumed to be nonstochastic, then forward rates are log-normal, which coincides with Black's pricing formula. Moreover, the consequence that interest rates are nonnegative and zero-coupon bond prices are nonzero under Monte Carlo simulations makes the model widely accepted. Therefore, for the past two decades, the LIBOR market model has been highly developed for various applications including the LIBOR swap market. Derivations and implementations of these market models can be found in Brigo and Mercurio (2006) and Rebonato (2002, 2004).

INTEREST RATE DERIVATIVES

An interest-rate cap consists of a portfolio of caplets that provide insurance against rising borrowing costs. Let $C(T)$ denote a caplet with maturity T on a simple τ -LIBOR forward rate \hat{F} over time interval $[t, T]$. The payoff of this LIBOR caplet is

$$C(T, T) = L (\hat{F}(T, T, T + \tau) - K)^+$$

where L is the principal amount and K is the strike rate. Under the market model setting with deterministic forward-rate volatilities, the caplet price can be written in Black's formula by

$$\begin{aligned} C(0, T) = & B(0, T)L [\hat{F}(0, T, T + \tau)N(d_1) - KN(d_2)] \\ d_1 = & \frac{\ln\left(\frac{\hat{F}(0, T, T + \tau)}{K}\right) + \frac{1}{2} \int_0^T \gamma^2(u, T, T + \tau) du}{\sqrt{\int_0^T \gamma^2(u, T, T + \tau) du}} \\ d_2 = & d_1 - \sqrt{\int_0^T \gamma^2(u, T, T + \tau) du} \end{aligned}$$

in which the volatility structure is flat with respect to the caplet strike prices. Despite this limitation, the model becomes the building block for replicating exotic interest-rate derivatives since the implied volatilities can be derived from several plain-vanilla traded derivatives. The information determined from this smaller scale market is then extended to characterize the whole-market dynamic. The operation usually involves interpolating, and many techniques are introduced in Rebonato (2002).

For pricing exotic interest-rate derivatives, interpolation on implied volatilities is often necessary, though undesirable because the HJM no-arbitrage condition cannot hold in most cases. LIBOR serial options, for example, are not as actively traded, so the prices are calculated based on the LIBOR cap/floor market. A serial option has two different maturities for the underlying forward rate agreement different from the option itself. Despite the availability of a closed-form solution, the needed volatility input for Black's formula turns out to be a partial integration from time t to the option maturity, and this information is not available from the cap/floor market. Therefore, further heuristic treatment is usually undertaken to connect the dots, in which case the curve would behave in explicit patterns and allow arbitrage.

DESIGNING YOUR NEXT MODEL

No single model is perfect in general for all assets in any market environment. The trade-offs between convenience and accuracy are evaluated by individual trading desks, quantitative analysts, and ultimately validated by the market. Nonetheless, when presenting a new model, three aspects are usually evaluated.

From a financial aspect, a model must be able to price the underlying asset(s) and its derivatives simultaneously. The market for an asset

and its derivatives are congruent, and there is no logic in pricing them separately, thereby risking "model" arbitrage. For example, we construct an interest-rate model for LIBOR-swaps curve in the real world and organically embed it in the model to price LIBOR derivatives such as LIBOR caps, floors, or even serial options in a risk-neutral world. Another example is for an underlying bullet bond and its callable counterpart. A callable bond is a bullet bond with an issuer-long, embedded American-style call option; however, the bullet bond price is determined under the real-world measure and the embedded option can be priced in the risk-neutral world. Therefore, a good model should be able to value a callable bond by valuing the bullet bond and the embedded American-style call option simultaneously.

From a mathematical standpoint, a model must be able to exhibit equivalency under different measures by explicitly characterizing the market price of risk. This mathematical component builds the bridge connecting the real world and a risk-neutral world. A complete financial market infers the existence of a unique market price of risk; but we should emphasize that whether a market is complete or not does not depend on the existence of a set of complete traded assets, but on the existence of an entity that can make the market if an arbitrage opportunity is revealed. Therefore, modern financial markets create not only hedging tools but an intangible equilibrium, which validates the underlying mathematical assumptions.

Finally, as we employ computation, this aspect demands that models/derivatives that require Monte Carlo analysis must be simulated by the same algorithm efficiently under different measures. This issue is more important in interest-rate modeling since there may be a trade-off between satisfying the mathematical requirements of a model and employing a computational implementation. Finding a model that satisfies both criteria is not trivial even

though the markets are assumed to be complete. We specifically use the word “efficiently” to implicitly indicate that a model can be simulated by a recombination tree for American-style options.

Dynamic term structure models represent a highly developed condition where finance, mathematics, and computation come together. As opposed to the case with static term structure models where the term structure appears explicitly, for dynamic models the term structure of interest rates is usually implicitly embedded in models that engage in representing risk/value relative to current conditions for lending and borrowing over the spectrum of terms available in the market. Preclusion of arbitrage is fundamental for these models. We have shown two approaches to dynamic term structure models, one depending on a representation through the spot rate, the other depending on a representation through implied forward rates. In each case the relationship between the objective and risk-neutral world (measure) has been exploited to ensure coherence between underlying asset prices and any resulting derivative. Here, the value of the asset and the derivative each depend on a representation of the same determining condition of interest rates.

KEY POINTS

- Dynamic term structure models of interest rates readily admit uncertainty in valuation/risk analyses requiring a characterization of future market scenarios.
- In building dynamic term structure models it is important that equilibrium, in an arbitrage-free sense, is represented and that variations from the equilibrium may be represented in an appropriate, probabilistic sense through a choice of stochastic processes and probability measures.
- Two approaches in explaining the future course of interest rates embody the short-rate model or an evolution of forward rates.
- Common methods for analyzing fixed-income/interest-rate instruments include formulation through a risk-neutral measure or by maintaining a real-world (objective) probability measure. Each has its own merit.
- The market price of risk is the key link between the risk-neutral and objective probability measures.

REFERENCES

- Björk, T. (2009). *Arbitrage Theory in Continuous Time*, 3rd ed. Oxford: Oxford University Press.
- Black, F., Derman, E., and Toy, W. (1990). A one factor model of interest rates and its application to Treasury bond options. *Financial Analysts Journal* (January/February): 33–39.
- Brace, A., Gatarek, D., and Musiela, M. (1997). The market model of interest rate dynamics. *Mathematical Finance* 4: 127–155.
- Brigo, D., and Mercurio, F. (2006). *Interest Rate Models—Theory and Practice*, 2nd ed. New York: Springer.
- Chan, K. C., Karolyi, G. A., Longstaff, F. A., and Sanders, A. B. (1992). An empirical comparison of alternative models of the short-term interest rate. *Journal of Finance* (July): 1209–1227.
- Cox, J. C., Ingersoll, J. E., Jr., and Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica* (March): 385–407.
- Geman, H., El Karoui, N., and Rochet, J-C. (1995). Changes of numéraire, changes of probability measure and option pricing. *Journal of Applied Probability* 32: 443–458.
- Heath, D., Jarrow, R. A., and Morton, A. (1992). Bond pricing and the term structure of interest rates: A new methodology. *Econometrica* 60, 1: 77–105.
- Ho, T. S. Y., and Lee, S. B. (1986). Term structure movements and pricing interest rate contingent claims. *Journal of Finance* (December): 1011–1029.
- Hull, J. C. (2009). Interest rate derivatives: Models of the short rate. Chapter 30 in *Options, Futures,*

- and Other Derivatives*. Upper Saddle River, NJ: Pearson.
- Oksendal, B. (2007). *Stochastic Differential Equations*, 6th ed. New York: Springer.
- Rebonato, R. (2002). *Modern Pricing of Interest-Rate Derivatives*. Princeton, NJ: Princeton University Press.
- Rebonato, R. (2004). *Volatility and Correlation*, 2nd ed. Chichester, UK: Wiley.
- Shreve, S. E. (2004). *Stochastic Calculus for Finance II: Continuous-Time Models*, 2nd ed. New York: Springer.
- Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics*, 177–188.
- Vasicek, O. A., and Fong, H. G. (1982). Term structure modeling exponential spline. *Journal of Finance* (May): 339–348.

Essential Classes of Interest Rate Models and Their Use

PETER FITTON

Manager, Scientific Development CreditXpert Inc.

JAMES F. McNATT, CFA

Executive Vice President, ValueWealth Services

Abstract: Models of the *term structure of interest rates* have become increasingly important in financial modeling. However, the understanding of these models by practitioners has not always kept pace with the breadth of the application of these models. In particular, misinterpretation of the proper uses of a particular model can lead to significant errors. The confusion regarding these models has arisen because of the overuse and misuse of the term “arbitrage-free.”

In this entry, we attempt to clear up some of the most commonly misconstrued aspects of interest rate models: the choice between an arbitrage-free or equilibrium model, and the choice between risk neutral or realistic parameterizations of a model. These two dimensions define four classes of model forms, each of which has its own proper use.

Much of the confusion has arisen from overuse and misuse of the term “arbitrage-free.” Virtually all finance practitioners believe that market participants quickly take advantage of any opportunities for risk-free arbitrage among financial assets, so that these opportunities do not exist for long; thus, the term “arbitrage-free” sounds as if it would be a good characteristic for any model to have. Simply based on these positive connotations, it almost seems hard to believe that anyone would not want their model to be arbitrage-free. Briefly,

in the world of finance this expression has the associations of motherhood and apple pie.

Unfortunately, this has led some users (and even builders) of interest rate models to link uncritically the expression “arbitrage-free” with the adjective “good.” One objective of this entry is to show that arbitrage-free models are not appropriate for all purposes. Further, we show that just because a model uses the arbitrage-free approach does not mean that it is necessarily good, even for the purposes for which arbitrage-free models are appropriately used.

Another common confusion ensues from implicitly equating the terms “arbitrage-free” and “risk neutral.” This arises partly from the fact that, in the academic and practitioner literature, there have been very few papers that have applied the arbitrage-free technique to a model that was not in risk-neutral form. We explain

the reason for this below. The natural result is that the terms have sometimes been used interchangeably. In addition, since quantitative risk management is a relatively new concept to the finance community, most well-known papers have focused only on the application of interest rate models to simple valuation and hedging problems. These have not required either the realistic or equilibrium approaches to modeling. This lack of published work has led to a mistaken belief that an arbitrage-free, risk-neutral model is the only valid kind of term structure model. In this entry, we intend to dispel that notion.

CATEGORIZATION OF APPROACHES TO TERM STRUCTURE MODELING

Arbitrage-Free Modeling

Arbitrage-free models take certain market prices as given and adjust model parameters in order to fit the prices exactly. Despite being called “term structure” models, they do not in reality attempt to emulate the dynamics of the term structure. Instead, they assume some computationally convenient, but essentially arbitrary, random process underlying the yield curve, and then add time-dependent constants to the drift (mean) and volatility (standard deviation) of the process until all market prices are matched. To achieve this exact fit, they require at least one parameter for every market price used as an input to the model.

For valuation, it is possible to produce reasonable current prices for many assets without having a realistic term structure model, by using arbitrage-free models for interpolation among existing prices. To this end, the trading models used by most dealers in the over-the-counter derivatives market employ enormous numbers of time-dependent parameters. These achieve an exact fit to prices of assets in particular classes, without regard to any differences

between the behaviors of the models and the actual behavior of the term structure over time. Placed in terms of a physical analogy, the distinction here is between creating a robot based on a photograph of an animal, and creating a robot based on multiple observations of the animal through time. While the robot produced using only the photograph may *look* like the animal, only the robot built based on behavioral observations will *act* like the animal. An arbitrage-free model is like the former robot, constructed with reference to only a single point in time; that is, a snapshot of the fixed-income marketplace.

Equilibrium Modeling

In contrast to arbitrage-free models, equilibrium term structure models are truly models of the term structure process. Rather than interpolating among prices at one particular point in time, they attempt to capture the behaviors of the term structure over time. An equilibrium model employs a statistical approach, assuming that market prices are observed with some statistical error, so that the term structure must be estimated, rather than taken as given. Equilibrium models do not exactly match market prices at the time of estimation, because they use a small set of state variables (fundamental components of the interest rate process) to describe the term structure. Extant equilibrium models do not contain time-dependent parameters; instead they contain a small number of statistically estimated constant parameters, drawn from the historical time series of the yield curve.

Risk-Neutral Probabilities:

The Derivative Pricing Probability Measure

When we create a model for pricing interest rate derivatives, the “underlying” is not the price of a traded security, as it would be in a model for equity options. Instead, we specify a random

process for the instantaneous, risk-free spot interest rate, the rate payable on an investment in default-free government bonds for a very short period of time. For convenience, we call this interest rate “the short rate.” Financial modelers have chosen to create models around the short rate because it is the only truly riskless interest rate in financial markets. An investment in default-free bonds for any noninstantaneous period of time carries *market risk*, the chance that the short rate will rise during the term of the investment, leading to a decline in the investment’s value.

As with any risky investment, an investor in bonds subject to market risk expects to earn a risk-free return (that is, the return from continuously investing at the short rate, whatever that may be) plus a risk premium, which could increase or decrease as the term of the investment increases. Thus, the spot rate for a particular term is composed of the return expected under the random process for the short rate up to the end of that term, plus a term premium, an additional return to compensate the investor for the interest rate risk of the investment. The term premium offered in the market depends on the aggregate risk preference of market participants, taking into account their natural preferences for securities that conform to their investment (term) needs.

Let r_t be the short rate at time t . Let $D(t, T)$ be the price, at time t , of a discount bond paying one dollar at time T . Let $s(t, T)$ be the spot rate at time t for the term $(T-t)$. Finally, let $\phi(T-t)$ be the term premium (expressed as an annual excess rate of return) required by investors for a term of $(T-t)$. All rates are continuously compounded. We can then write,

$$D(t, T) = \frac{1}{e^{s(t,T) \times (T-t)}} = \frac{1}{e^{\phi(T-t) \times (T-t)}} E \left[\frac{1}{e^{\int_t^T r_s ds}} \right] \tag{1}$$

The second term in the two-term expression above is a discount factor that reflects the expected return from investing continuously at

the short rate for the term $(T-t)$. The first term is the additional discount factor that accounts for the return premium that investors require to compensate them for the market risk of investing for a term of $(T-t)$. The use of an integral in the expression for the expected short rate discount factor is necessary because the short rate is continuously changing over the bond’s term.

From this description and formula, it may seem necessary to know the term premium for every possible term, in addition to knowing the random process for the short rate, in order to value a default-free discount bond. This is not the case, however. As in the pricing of a forward contract or option on a stock, we can use the mathematical sleight-of-hand known as *risk-neutral valuation* to find the relative value of a security that is derivative of the short rate.

The principle of risk-neutral valuation as it applies to bonds and other interest rate derivatives is that, regardless of how risk averse investors are, we can identify a set of spot rates that value discount bonds correctly relative to the rest of the market. We do not have to identify separately the term premium embedded in each spot rate in order to use it to discount future cash flows. This fact can be used to make the valuation of all interest rate derivatives easier by risk adjusting the term structure model; that is, by changing the probability distribution of the short rate so that the spot rate of every term is, under the new model, equal to the expected return from investing at the short rate over the same term. This is accomplished by redefining the model so that, instead of being a random process for the short rate, it is a random process for the short rate plus a function of the term premium. If we specify the process for r_t^* in such a way that

$$r_s^* = r_s + \phi(s-t) + \phi'(s-t) \times (s-t) \tag{2}$$

at every future point in time s (accomplished by adjusting the rate of increase of r_t upward) then

we can write,

$$\begin{aligned}
 D(t, T) &= \frac{1}{e^{\int_t^T r_s ds}} \\
 &= E \left[\frac{1}{e^{\int_t^T (r_s + \phi(T-t)) ds}} \right] = E \left[\frac{1}{e^{\int_t^T r_s^* ds}} \right] \quad (3)
 \end{aligned}$$

By transforming the short rate process in this manner, we have created a process for a random variable which, when used to discount a certain future cash flow, gives an expected present value equal to the present value obtained by discounting that cash flow at the appropriate spot rate. It is important to note that this random variable is no longer the short rate, but something artificial that we might refer to as the *risk-adjusted short rate*.¹

The resulting *risk-neutral model* might be construed as a model for the true behavior of the short rate in an imaginary world of risk-neutral market participants, where there is no extra expected return to compensate investors for the extra price risk in bonds of longer maturity. This impression, while accurate, is not very informative. The important aspect of the risk-neutral model is that the term premiums, whatever their values, that exist in the marketplace are embedded in the interest rate process itself, so that the expected discounted value of a cash flow at the risk-adjusted short rate is equal to the discounted value of the cash flow at the spot rate.²

The value of the *risk-neutral probability measure* is that, under this parameterization, an interest-sensitive instrument's price can be estimated by averaging the present values of its cash flows, discounted at the short-term interest rates along each path of the short rate under which those cash flows occur. In contrast, valuing assets under the model before it was risk adjusted would require a more complicated discounting procedure that applied additional discount factors to the short rate paths to compensate for market risk; however, the price obtained under both approaches would be the same. For this rea-

son, we use randomly generated scenarios from risk-neutral interest rate models for pricing.

To sum up, there is nothing magical about risk neutrality. There are any number of changes of variables we could make to a short rate process that would retain the structure of the model, but have a different (but equivalent) probability distribution for the new variable. We could change the measure to represent imaginary worlds in which market participants were risk seeking (negative term premiums), or more risk averse than in the real world; regardless, as long as we structured the discounting procedure properly we would always determine the same model price for an interest rate derivative. The specific change of variables that produces a risk-neutral model simply makes the algebra easier than the others, because one can ignore risk preferences.

Realistic Probabilities:

The Estimated Market Probability Measure

We have described why risk-neutral interest rate scenarios are preferred for pricing bonds and interest rate derivatives. However, it is important to note that risk-neutral scenarios are not appropriate for all purposes. For example, for scenario-based evaluation of portfolio strategies, realistic simulation is needed. And a computerized system for stress testing asset/liability strategies under adverse movements in interest rates is to actuaries what a wind tunnel is to aerospace engineers. The relevance of the information provided by the testing depends completely on the realism of the simulated environment. Stated differently, the test environment must be like the real environment; if not, the test results are not useful.

The realistic term structure process desired for this kind of stress testing must be distinguished from the risk-neutral term structure process used for pricing. The risk-neutral process generates scenarios in which all term premiums are zero. This process lacks realism; in the real world, term premiums are clearly not

zero, as evidenced by the fact that the implied spot curve from Treasuries has been predominantly upward sloping. This predominantly upward slope reflects an expected return premium for bonds of longer maturity, although at times other configurations of buyer preferences can be inferred; for example, an inverted curve suggests that buyers demand an increasing premium for decreasing the term of their positions.

Thus, the user of an interest rate model must be careful. When generating scenarios for reserve adequacy testing, where the purpose is to examine the effect on a company's balance sheet of changes in the real (risk-averse) world, the user must not use the scenarios from a risk-neutral interest rate model.

WHEN DO I USE EACH OF THE MODELING APPROACHES?

The two dimensions, risk-neutral versus realistic and arbitrage-free versus equilibrium, define four classes of modeling approaches. Each has its appropriate use.

Risk-Neutral and Arbitrage-Free Model

The risk-neutral and arbitrage-free model is the most familiar form of an interest rate model for most analysts. The model has been risk adjusted to use for pricing interest rate derivatives, and its parameters have been interpolated from a set of current market prices rather than being statistically estimated from historical data. It is appropriately used for current pricing when the set of market prices is complete and reliable.

It is worth noting that, just because two models are each both risk neutral and arbitrage-free, we cannot conclude that they will give the same price for a particular interest rate derivative. Two arbitrage-free models will produce the same prices only for the instruments in a subset

common to both sets of input data. The form of the model, and particularly the number of random factors underlying the term structure process, can make a large difference to valuations of the other instruments.

When the market data are sparse, the behavior of the model becomes important. For example, the value of a Bermudan or American swaption depends on the correlations among rates of different maturities. The swaption market is not liquid, nor are its prices widely disseminated, so there is no way to estimate a "term structure of correlations" that would allow a simple arbitrage-free model to interpolate reasonable swaption prices. In this case, a multi-factor model that captures the nature of correlations among rates of different maturities, including the way that those correlations are influenced by the shape of the term structure, will perform better for pricing swaptions than will a one-factor model. Models with good statistical fit to historical correlation series are needed for Bermudan or American options on floating-rate notes, caps, and floors for the same reason. Model behavior is also important for long-dated caps and floors, where there is a lack of reliable data for estimating the "term structure of volatilities" beyond the 5-year tenor.

Risk-Neutral and Equilibrium

There are a number of sources of "error" in quotations of the market prices of bonds, so that the discount rates that exactly match a set of price quotations may contain bond-specific effects, corrupting the pricing of other instruments. These sources, defined as any effects on a bond's market price apart from the discount rates applying to all market instruments, include differences in liquidity, differential tax effects, bid-ask spreads (the bid-ask spread defines a range of possible market prices, implying a range of possible discount rates), quotation stickiness, timeliness of data, the human element of the data collection and reporting process, and market imperfections.

Since arbitrage-free models accept all input prices as given, without reference to their reasonability or comparability to other prices in the input data, they impound in the pricing model any bond-specific effects. In contrast, equilibrium models capture the global behavior of the term structure over time, so security-specific effects are treated in the appropriate way, as noise. For this reason, risk-neutral equilibrium models can have an advantage over arbitrage-free models in that equilibrium models are not overly sensitive to outliers. Also, for current pricing (as distinguished from horizon pricing, described below), equilibrium models can be estimated from historical data when current market prices are sparse. Thus, a risk-neutral and equilibrium model can be used for pricing when the current market prices are unreliable or unavailable.

For most standard instruments, circumstances rarely prevail such that the current market prices needed for estimating an arbitrage-free model are not available. However, such circumstances always prevail for horizon pricing, where the analyst calculates a price for an instrument in some assumed future state of the market. Since arbitrage-free models require a full set of market prices as input, arbitrage-free models are useless for horizon pricing, the future prices being unknown. Thus, the horizon prices obtained under the different values of the state variables in an equilibrium model provide an analytical capability that arbitrage-free models lack.

USING MODELS OF BORROWER BEHAVIOR WITH A RISK-NEUTRAL INTEREST RATE MODEL

Often, an interest rate model is not enough to determine the value of a fixed-income security or interest rate derivative. To value mortgage-backed securities or collateralized mortgage obligations (CMOs), one also needs a prepayment model. To value bonds or interest rate derivatives with significant credit risk, one

needs a model of default and recovery. To value interest-sensitive annuities and insurance liabilities, one needs models of lapse and other policyholder behaviors. In all of these behavioral models, the levels of certain interest rates are important explanatory variates, meaning that, for example, the prepayment speeds in a CMO valuation system are driven primarily by the interest rate scenarios.

Common practice has been to estimate parameters for prepayment, default, and lapse models using regression on historical data about interest rates and other variables. Then, in the valuation process, the analyst uses the interest rates from a set of risk-neutral scenarios to derive estimates for the rates of prepayment, default, or lapse along those scenarios. This borrower behavior information is combined with the interest rates to produce cash flows and, ultimately, prices. Unfortunately, this practice leads to highly misleading results.

The primary problem here is that the regressions have been estimated using historical data, reflecting the *real* probability distributions of borrower behavior, and then used with scenarios from a risk-neutral model, with an *artificial* probability distribution. The risk-neutral model is not a process for the short rate; rather, it is a process for the risk-adjusted short rate. Since the real world is risk averse, the risk-adjusted short rate usually has an expected value much higher than the market's forecast of the short rate; the extra premium for interest rate risk permits one to value optionable default-free bonds by reference to the forward rate curve.

The same procedure can be applied to corporate bonds. Corporate bonds are exposed to default risk in addition to interest rate risk. One may construct a behavioral model of failure to pay based on historical data about default rates and recovery, perhaps using bond ratings as explanatory variates in addition to interest rates. One can then attempt to compute the present value of a corporate bond by finding the expected value of the discounted cash flows from the two models in combination: a risk-neutral model of the Treasury curve, and a

realistic model of default behavior as a function of interest rates and other variables. Because the cash flows of the bond, adjusted for default, will be less than the cash flows for a default-free bond, the model will price the corporate bond at a positive spread over the Treasury curve.

This spread will almost certainly be substantially too low in comparison to the corporate's market price. The reason for this is that, just as investors demand a return premium for interest rate risk, they demand an additional return for default risk. The application of an econometrically estimated model of default to pricing has ignored the default risk premium encapsulated in the prices of corporate bonds. Market practice has evolved a simple solution to this; one adjusts the default model to fit (statistically, in the equilibrium case; exactly, in the arbitrage-free case) the current prices of active corporates in the appropriate rating class. By using the market prices of active corporates to embed the default risk premium in the model, the analyst is really applying the principle of risk-neutral valuation to the default rate. The combined model of risk-adjusted interest rates and risk-adjusted default rates now discounts using the corporate bond spot rate curve instead of the Treasury spot curve.

The same technique of risk neutralizing a model by embedding information about risk premiums derived from current market prices can be applied to prepayment models as well. The results of a prepayment model can be risk adjusted by examining the prices of active mortgage-backed securities. Unfortunately, one can only guess at the appropriate expected return premium for insurance policy lapse risk or mortality risk. Nevertheless, these quantities should be used to "risk neutralize" these models of behavior to the extent practical. The integrity of risk-neutral valuation depends on risk adjusting all variables modeled; otherwise, model prices will be consistently overstated.

A final note can be made in this regard about option-adjusted spread (OAS). OAS can be understood in this context as a crude method to

risk adjust the pricing system to reflect all risk factors not explicitly modeled.

Realistic and Arbitrage-Free

A *realistic, arbitrage-free model* starts by exactly matching the term structure of interest rates implied by a set of market prices on an initial date, then evolves that curve into the future according to the realistic probability measure. This form of a model is useful for producing scenarios for evaluation of hedges or portfolio strategies, where it is important that the initial curve in each scenario exactly matches current market prices. The difficulty with such an approach lies in the estimation; realistic, arbitrage-free models are affected by confounding, where it is impossible to discriminate between model misspecification error and the term premiums. Since the model parameters have been set to match market prices exactly, without regard to historical behavior, too few degrees of freedom remain to estimate both the term premiums and an error term. Unless the model perfectly describes the true term structure process (that is, the time-dependent parameters make the residual pricing error zero at all past and future dates, not just on the date of estimation), the term premiums cannot be determined. The result is that realistic, arbitrage-free models are not of practical use.

Realistic and Equilibrium

Since the arbitrage-free form of a realistic model is not available, the equilibrium form must be used for stress testing, Value-at-Risk (VAR) calculations, reserve and asset adequacy testing, and other uses of realistic scenarios.

Some analysts express concern that, because the predicted initial curve under the equilibrium model does not perfectly match observed market prices, then the results of scenario testing will be invalid. However, the use of an equilibrium form does not require that the predictions be used instead of the current market prices as the first point in a scenario. The

Table 1 When to Use Each of the Model Types

Model Classification	Risk Neutral	Realistic
Arbitrage-free	<ul style="list-style-type: none"> • Current pricing, where input data (market prices) are reliable 	<ul style="list-style-type: none"> • Unusable, since term premium cannot be reliably estimated
Equilibrium	<ul style="list-style-type: none"> • Current pricing, where inputs (market prices) are unreliable or unavailable • Horizon pricing 	<ul style="list-style-type: none"> • Stress testing • Reserve and asset adequacy testing

Table 2 Four Forms of the Black-Karasinski Model

Model Classification	Risk Neutral	Realistic
Arbitrage-free	$du = \kappa(t) (\theta(t) - u) dt + \sigma(t) dz$ <ul style="list-style-type: none"> • u_0 and $\theta(t)$ matched to bond prices • $\kappa(t)$ and $\sigma(t)$ matched to cap or option prices 	$du = \kappa(t) (\theta(t) - \lambda(u,t) - u) dt + \sigma(t) dz$ <ul style="list-style-type: none"> • u_0 and $\theta(t)$ matched to bond prices • $\kappa(t)$ and $\sigma(t)$ matched to cap or option prices • $\lambda(u,t)$ cannot be reliably estimated
Equilibrium	$du = \kappa(\theta - u) dt + \sigma dz$ <ul style="list-style-type: none"> • u_0 statistically fit to bond prices • κ, θ, σ historically estimated 	$du = \kappa(\theta - \lambda(u) - u) dt + \sigma dz$ <ul style="list-style-type: none"> • u_0 statistically fit to bond prices • $\kappa, \theta, \sigma, \lambda(u)$ historically estimated

scenarios can contain the observed curve at the initial date and the conditional predictions at future dates. This does not introduce inconsistency, because the *equilibrium model* is a statistical model of term structure behavior; by taking this approach we explicitly recognize that its predictions will deviate from observed values by some error. In contrast, the use of an *arbitrage-free, realistic model* implicitly assumes that the model used for the term structure process is absolutely correct.

Summary of the Four Essential Classes

Table 1 summarizes the uses of the four Essential Classes of interest rate models. Table 2 shows the mathematical form of a commonly used interest rate model, disseminated by Black and Karasinski (1991), under each of the modeling approaches and probability measures. In each equation, u is the natural logarithm of the short rate.

In the above models, σ is the instantaneous volatility of the short rate process, κ is the rate of mean reversion, θ is the mean level to which the natural logarithm of the short rate is reverting, and λ represents the term premium demanded

by the market for holding bonds of longer maturity. The value of the state variable u at the time of estimation is represented by u_0 .

The realistic model forms can be distinguished from the risk-neutral forms by the presence of the term premium function λ . The difference between the arbitrage-free forms and the equilibrium forms can be discerned in that the parameters of the arbitrage-free forms are functions of time.

KEY POINTS

- Models of the term structure of interest rates are important in financial modeling.
- The most commonly misconstrued aspects of interest rate models are important to understand to make the correct choice between an arbitrage-free or equilibrium model, and the correct choice between risk-neutral or realistic parameterizations of a model.
- A common confusion is the result of implicitly equating the terms “arbitrage-free” and “risk neutral.”
- Arbitrage-free models take certain market prices as given and adjust model parameters in order to fit the prices exactly.

- Equilibrium term structure models are truly models of the term structure process because rather than interpolating among prices at one particular point in time, they attempt to capture the behaviors of the term structure over time.
- The principle of risk-neutral valuation as it applies to bonds and other interest rate derivatives is that, regardless of how risk averse investors are, a set of spot rates that value discount bonds correctly relative to the rest of the market can be identified.
- The two dimensions, risk-neutral versus realistic and arbitrage-free versus equilibrium, define four classes of modeling approaches.
- The risk-neutral and arbitrage-free model is appropriately used for current pricing when the set of market prices is complete and reliable.
- Because equilibrium models capture the global behavior of the term structure over time, so security-specific effects are treated as noise, a risk-neutral and equilibrium model can be used for pricing when the current market prices are unreliable or unavailable.
- For several reasons, realistic, arbitrage-free models are not of practical use.

NOTES

1. This is not the way that risk neutrality is usually presented. Typically, writers have focused on the stochastic calculus, using Girsanov's theorem to justify a change of

probability measure to an equivalent (i.e., an event has zero probability under one measure if and only if it has zero probability under the other measure) martingale measure. This complexity and terminology can obscure the simple intuition that we are making a change of variables in order to restate the problem in a more easily solvable form. For this approach to explaining risk neutral valuation, see Courtadon (1982) or Harrison and Pliska (1981).

2. Note that this is not the same as the expectations hypothesis of the term structure, which holds that the term structure's shape is determined solely by the market's expectations about future rates. The expectations hypothesis is a theory of the real term structure process, whereas the risk-neutral approach is an analytical convenience that takes no position about the truth or falsity of any term structure theory.

REFERENCES

- Black, F., and Karasinski, P. (1991). Bond and option pricing when short rates are lognormal. *Financial Analysts Journal* 4: 52–59.
- Courtadon, G. (1982). The pricing of options on default-free bonds. *Journal of Financial and Quantitative Analysis* 17: 75–100.
- Harrison, J., and Pliska, J. (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and Their Applications* 11: 215–260.

A Review of No Arbitrage Interest Rate Models

GERALD W. BUETOW, Jr., PhD, CFA
President and Founder, BFRC Services, LLC

FRANK J. FABOZZI, PhD, CFA, CPA
Professor of Finance, EDHEC Business School

JAMES SOCHACKI, PhD
Professor of Applied Mathematics, James Madison University

Abstract: Interest rates are commonly modeled using stochastic differential equations. One-factor models use a stochastic differential equation to represent the short rate and two-factor models use a stochastic differential equation for both the short rate and the long rate. The stochastic differential equations used to model interest rates must capture some of the market properties of interest rates such as mean reversion and/or a volatility that depends on the level of interest rates. There are two distinct approaches used to implement the stochastic differential equations into a term structure model: equilibrium and no arbitrage.

In modeling the behavior of interest rates, *stochastic differential equations* (SDEs) are commonly used. The SDEs used to model interest rates must capture some of the market properties of interest rates such as mean reversion and/or a volatility that depends on the level of interest rates. For a *one-factor model*, the SDE is used to model the behavior of the short-term rate, referred to simply as the “short rate.” The addition of another factor (i.e., a *two-factor model*) involves extending the SDE to represent the behavior of the short rate and a long-term rate (i.e., long rate).

There are two distinct approaches used to implement the SDEs into a *term structure model*:

equilibrium and no arbitrage. Each can be used to value bonds and interest rate contingent claims. Both approaches start with the same SDEs but apply the SDE under a different framework to price securities.

Equilibrium models such as those developed by Vasicek (1977), Cox, Ingersoll, and Ross (1985), Longstaff (1989, 1992), Longstaff and Schwartz (1992), and Brennan and Schwartz (1979, 1982) all start with an SDE model and develop pricing mechanisms for bonds under an equilibrium framework. The actual implementation may vary depending on the model. Vasicek and Cox, Ingersoll, and Ross (CIR) develop analytic pricing expressions while Backus, Foresi, and

Telmer (2001) present econometric and recursive approaches to implement the equilibrium models. Brennan and Schwartz use a finite difference scheme that approximates a partial differential equation.

No arbitrage models such as Black and Karasinski, (1991), Black, Derman, and Toy (1990), Ho and Lee (1986), Heath, Jarrow, and Morton (1992), and Hull and White (1990, 1993) begin with the same or similar SDE models as the equilibrium approach but use market prices to generate an interest rate lattice. The lattice represents the short rate in such a way as to ensure there is a no arbitrage relationship between the market and the model. The numerical approach used to generate the lattice will depend on the SDE model(s) being used to represent interest rates.

No arbitrage models are the preferred framework to value interest rate derivatives. This is because they minimally ensure that the market prices for bonds are exact. Equilibrium models will not price bonds exactly, and this can have tremendous effects on the corresponding contingent claims. No arbitrage lattices also allow for a systematic valuation approach to almost all interest rate securities.

Three general SDE functional forms are considered in this entry. The first is the *Hull-White (HW) model*. The HW model is a more general version of the *Ho and Lee (HL)*¹ approach except that it allows for mean reversion. Implementing the HW in a binomial framework removes a degree of freedom, and in this case the HW model collapses to the HL model if a constant time step is retained. The second model we consider is the *Black-Karasinski (BK) model*. The BK model is a more general form of the *Kalotay, Williams, and Fabozzi (KWF) model*.² The BK model (like the HW model) in the binomial setting does not have enough degrees of freedom to be properly modeled and so the time step must be allowed to vary. The third is the *Black, Derman, and Toy model*.

We implement the HW and BK trinomial models using the Hull and White approach.

Within the trinomial setting the time step remains constant and *mean reversion* can be explicitly incorporated. We discuss the SDEs, the properties of the SDEs, the numerical solutions to the SDEs, and the binomial and trinomial interest rate lattices for these models.

The focus of our presentation is on the end user and developer of interest rate models. We will highlight some significant differences across models. Most of these are due to the different distributions that underlie the models. This is done to emphasize the need to calibrate all models to the market prior to their use. By calibrating the models to the market we reduce the effects of the distributional differences and ensure a higher level of consistency in the metrics produced by the models.

The outline of this entry is as follows. In the next section we present the SDEs and some of their mathematical properties. We also use the mathematics to highlight properties of the short rate. We then develop the methodology used to implement our approach in both the binomial and trinomial frameworks. A comparison of some numerical results across the different models including some interest rate risk and valuation metrics is then presented.

THE GENERAL MODELS FOR THE SHORT RATE

The models considered in this entry take the form of the following one-factor SDE:

$$df(r(t)) = [\theta(t) + \rho(t)g(r(t))]dt + \sigma(r(t), t)dz \quad (1)$$

where f and g are suitably chosen functions, θ is determined by the market, and ρ can be chosen by the user of the model or dictated by the market. We will show that θ is the drift of the short rate and ρ is the tendency to an equilibrium short rate. The term σ is the local volatility of the short rate. The term $dz = \varepsilon\sqrt{dt}$ arises from a normally distributed Wiener process, since $\varepsilon \sim N(0,1)$, where $N(0,1)$ is the normal

distribution with mean 0 and standard deviation of 1. This means that the term $\sigma(r(t), t)dz$ has an average or expected value of 0.

Equation (1) has two components. The first component is the expected or average change in rates over a small period of time, dt . This is the component where certain characteristics of interest rates, such as mean reversion, are incorporated. The second component is the unknown or the risk term since it contains the random term. This term dictates the distribution characteristics of interest rates. Depending on the model, interest rates are either normally or lognormally distributed.

The Ho-Lee Model

In the HL model or process $f(r) = r$, $g(r) = 0$, and $\rho = 0$ in equation (1). The HL process is, therefore, given by

$$dr = \theta dt + \sigma dz \tag{2}$$

Since z is a normally distributed Wiener process, we say the HL process is a normal process for the short rate. The solution to equation (2), assuming $r(0) = r_0$ is given by

$$r(t) = r_0 + \int_0^t \theta ds + \int_0^t \sigma dz \tag{3a}$$

where the integral involving σ is a stochastic integral. If θ is constant this can be expressed as

$$r(t) = r_0 + \theta t + \int_0^t \sigma dz \tag{3b}$$

Equation (3b) shows that the HL process models an interest rate that can change proportionally with time t through the constant of proportionality, θ , and a random disturbance determined by σ . That is, the larger θ is in magnitude, the larger the average change in the short rate over time. This is why θ is called the “drift in the short rate.” Also, the smaller θ is, the larger the influence of the random disturbance. The short rate can be negative in

the HL process. This is a shortcoming of the model. Hull (2000) shows that θ is related to the slope of the term structure.

To obtain a numerical approximation for equation (2) we approximate equation (2) by using equations (3a) and (3b). Letting $t_k = k\tau$ and $r_k \approx r(k\tau)$ gives

$$r_{k+1} - r_k = \theta_k \tau + \sigma_k \Delta z_k$$

or

$$r_{k+1} = r_k + \theta_k \tau + \sigma_k \Delta z_k \tag{4}$$

where Δz_k is a numerical (discrete) approximation to dz . Since $dz = \varepsilon \sqrt{dt}$, we can further approximate equation (4) by

$$r_{k+1} = r_k + \theta_k \tau + \sigma_k \varepsilon_k \sqrt{\tau} \tag{5}$$

where ε_k is a random number given by a normal distribution $N(0,1)$. Equation (5) is the form of the expression that is used for r_{k+1} to build the HL binomial tree.

We first consider the solution to equation (5) without the stochastic term when θ is constant. Equation (5) under these requirements is

$$r_{k+1} = r_k + \tau \theta \tag{6a}$$

and the solution is given by

$$r_k = c + k\delta \tag{6b}$$

where c and δ are constants. In particular, $c = r_0$ and $\delta = \theta\tau$. It is seen from this last equation that the mean short rate in the HL process increases or decreases at a constant rate θ over time depending on the sign of θ . As a matter of fact, equation (6b) shows that the short rate grows without bound if $\theta > 0$ and decreases without bound (i.e. becomes very negative) if $\theta < 0$.

The Hull-White Model

In the HW model or process $f(r) = r$, $g(r) = r$, and $\rho = -\phi$. Therefore, the stochastic process for the HW model for the short rate is

$$dr = (\theta - \phi r)dt + \sigma dz \tag{7}$$

The short rate process in the HW model is seen to be normal as in the HL process. We consider the case where the parameters θ and ϕ are constant over time. Note that if $\phi = 0$ the HL process reduces to the HW process. (The HW process will, therefore, be similar to the HL process if ϕ is close to 0.) We will see that the introduction of ϕ in the HW model is an attempt to incorporate mean reversion and to correct for the uncontrolled growth (or decline) in the HL model discussed later.

Eliminating the stochastic term in equation (7) gives the ordinary differential equation

$$dr = (\theta - \phi r)dt \quad (8)$$

whose solution is given by

$$r(t) = \frac{\theta}{\phi} + ce^{-\phi t} \quad (9)$$

where

$$c = r_0 - \frac{\theta}{\phi} \quad (10)$$

If $\phi > 0$ we see from equation (9) that

$$\lim_{t \rightarrow \infty} r(t) = \frac{\theta}{\phi} = \mu$$

Therefore, for positive mean reversion ($\phi > 0$) the HW process will converge to the short rate, μ . Due to this, the term μ , is called the “target” or “long run mean rate.” For negative mean reversion ($\phi < 0$), the short rate grows exponentially over time.

Factoring ϕ in equation (7) leads to

$$dr = \phi(\mu - r)dt + \sigma dz$$

and eliminating the stochastic term leads to

$$dr = \phi(\mu - r)dt$$

We see that if $r > \mu$ then dr is negative and r will decrease and if $r < \mu$ then dr is positive and r will increase. That is, r will approach the target rate μ . The larger ϕ is, the faster this approach to the target rate μ . This is why ϕ is called the “mean reversion” or “mean reversion rate.” It regulates how fast the target rate is reached. However, it does not eliminate the negative rates that can occur in the HL process.

Since the target rate μ is equal to θ/ϕ , we can solve for the drift, θ , or the mean reversion, ϕ . That is,

$$\theta = \mu\phi \quad (11)$$

or

$$\phi = \frac{\theta}{\mu} \quad (12)$$

It is seen from equations (11) and (12) that there is a strong relationship between the drift and mean reversion that can be used to reach any desired target rate. How large the mean reversion should be is an important financial question. Equations (11) and (12) can be used to set target rates. Equations (9) and (10) allow one to determine how long it takes to reach the target rate.

Approximating equation (7) gives us

$$r_{k+1} = r_k + (\theta_k - \phi_k r_k)\tau + \sigma_k \varepsilon_k \sqrt{\tau} \quad (13)$$

If θ and ϕ are constant and we eliminate the stochastic term, then the solution to equation (13) has the form

$$r_k = \alpha\beta^k + \gamma$$

To determine α , β , and γ we substitute this form for r_k into equation (13) under these conditions and obtain that $\beta = (1 - \phi\tau)$, $\gamma = \theta/\phi = \mu$, and $\alpha = r_0 - \mu$. Therefore,

$$r_k = \alpha(1 - \phi\tau)^k + \frac{\theta}{\phi} \quad (14)$$

Note that if $0 < \phi\tau < 2$ then $-1 < 1 - \phi\tau < 1$ and

$$\lim_{k \rightarrow \infty} r_k = \frac{\theta}{\phi} = \mu$$

which is the same result we obtained from equation (9) for the HW SDE. The condition $0 < \phi\tau < 2$ is easily maintained in modeling the short rate.

The Kalotay-Williams-Fabozzi Model

For the KWF process $f(r) = \ln(r)$, $g(r) = 0$, and $\rho = 0$ in equation (1). This leads to the

differential process

$$d \ln(r) = \theta dt + \sigma dz \tag{15a}$$

This model is directly analogous to the HL model. If $u = \ln r$, then we obtain the HL process (equation (2)) for u

$$du = \theta dt + \sigma dz \tag{15b}$$

Because u follows a normal process, $\ln(r)$ follows a normal process and so r follows a lognormal process. Since u follows the same process as the HL and HW models, u can become negative, but $u = \ln(r)$ and $r = e^u$ ensuring r is always positive. Therefore, the KWF model eliminates the problems of negative short rates that occurred in the HL and HW models.

Eliminating the stochastic term in equation (15) we obtain

$$d \ln(r) = \theta(t)dt$$

and

$$du = \theta(t)dt$$

From equation (3a) we have

$$\ln r(t) = u = u(0) + \int_0^t \theta(s) ds$$

since $u(0) = \ln r(0) = \ln r_0$,

$$\ln r(t) = \ln r(0) + \int_0^t \theta(s) ds$$

Taking the exponential of both sides gives us

$$r(t) = r_0 e^{\int_0^t \theta(s) ds} \tag{16}$$

showing that $r(t) > 0$ since $r(0) > 0$. Therefore, if $\theta(t) > 0$ the short rate in the KWF process can grow without bound and if $\theta(t) < 0$ the short rate in the KWF process can decay to 0.

From equation (5) for the HL process the discrete approximation to equation (15b) is

$$u_{k+1} = u_k + \theta_k \tau + \sigma_k \varepsilon_k \sqrt{\tau} \tag{17a}$$

and the exponential of this equation gives the discrete approximation to equation (15a):

$$r_{k+1} = r_k e^{\theta_k \tau + \sigma_k \varepsilon_k \sqrt{\tau}} \tag{17b}$$

From equation (17b) and equation (16) we see that the numerical approximation to equation (15a) has similar properties to the solution to the HL SDE. That is, if $\theta(t) > 0$ the short rate can grow without bound and if $\theta(t) < 0$ the short rate can decay to 0.

The Black-Karasinski Model

In the BK model we set $f(r) = \ln r$, $p = -\phi$, and $g(r) = \ln r$ in equation (1) to obtain the SDE

$$d \ln r = (\theta - \phi \ln r)dt + \sigma dz \tag{18a}$$

We now work with equation (18a) using equation (7) for the HW process in a manner similar to how we used results from the HL process to develop the KWF process. If we let $u = \ln r$ in equation (18a) we obtain

$$du = (\theta - \phi u)dt + \sigma dz \tag{18b}$$

which is the HW process for u . Again, note that u has all the same properties as r in the HW model. Since $r = e^u$ in the BK process, $r > 0$. This is the advantage the BK model has over the HW model. Therefore, we see that the BK process is an extension of the KWF process as the HW process is an extension of the HL process. The main difference is the BK is a lognormal extension of the lognormal KWF process. As a matter of fact, if $\phi = 0$ the BK process reduces to the KWF process. Black and Karasinski introduced ϕ to control the growth of the short rate in the KWF process.

From equation (9) we have

$$u(t) = \frac{\theta}{\phi} + ce^{-\phi t}$$

and after taking exponentials

$$r(t) = e^{u(t)} = e^{\frac{\theta}{\phi} + ce^{-\phi t}} \tag{19}$$

For $\phi < 0$ we see that r grows without bound and that for $\phi > 0$

$$\lim_{t \rightarrow \infty} r(t) = e^{\frac{\theta}{\phi}} = \mu$$

The target rate for the BK process is the exponential of the target rate for the HW process.

As in the HW process, from equation (19) (or equations (9) and (10)) we see that

$$c = \ln r_0 - \frac{\theta}{\phi} \quad (20)$$

in the BK process. The closer the initial rate is to the target rate, the faster the BK process converges to the target rate. From equations (19) and (20) we see that if the initial short rate is the target rate, then $r(t) = \mu$ for all t in the BK process, which is analogous to the HW process.

Given the target rate μ , we can solve for the drift or the mean reversion similarly to equations (11) and (12) in the HW model. We have

$$\theta = \phi \ln \mu \quad (21)$$

and

$$\phi = \frac{\theta}{\ln \mu} \quad (22)$$

We discretize $u = \ln r$ in equation (18b) just as we did for the HW SDEs and then let $r = e^u$. This is analogous to how we used the HL discrete process to get the KWF discrete process. The equations corresponding to equation (13) are

$$u_{k+1} = u_k + (\theta_k - \phi_k u_k)\tau + \sigma_k \varepsilon_k \sqrt{\tau} \quad (23a)$$

or after taking the exponential of both sides of equation (23a)

$$r_{k+1} = r_k e^{(\theta_k - \phi_k \ln r_k)\tau + \sigma_k \varepsilon_k \sqrt{\tau}} \quad (23b)$$

For constant θ and ϕ (similarly to equation (14)), the solution to equation (23b) after eliminating the stochastic term is

$$r_k = e^{\alpha(1-\phi r)^k + \frac{\theta}{\phi}} \quad (24)$$

Note from equation (24) that

$$\lim_{k \rightarrow \infty} r_k = e^{\frac{\theta}{\phi}} = \mu$$

for $0 < \phi\tau < 2$. This is similar to the result we obtained from equation (14) for the HW SDEs.

The Black-Derman-Toy Model

The Black-Derman-Toy (BDT) model is a lognormal model with mean reversion, but the mean reversion is endogenous to the model.

The mean reversion in the BDT model is determined by market conditions.

The equation describing the interest rate dynamics in the BDT model has $f(r) = \ln r$ and $g(r) = \ln r$ in equation (1) as in the BK model. Therefore, the short rate in the BDT model follows the lognormal process

$$d \ln r + [\theta(t) + \rho(t) \ln r]dt + \sigma(t)dz$$

However, in the BDT model $\rho(t) = \frac{d}{dt} \ln \sigma(t) = \frac{\sigma'(t)}{\sigma(t)}$ giving us

$$d \ln r = \left(\theta(t) + \frac{\sigma'(t)}{\sigma(t)} \ln r \right) dt + \sigma(t)dz \quad (25a)$$

Making the substitution $u = \ln r$ leads to

$$du = \left(\theta(t) + \frac{\sigma'(t)}{\sigma(t)} u \right) dt + \sigma(t)dz \quad (25b)$$

Notice the similarity in equations (25) and the equations (18) of the BK model. We expect

$$\frac{\sigma'(t)}{\sigma(t)}$$

to behave similarly to $-\phi(t)$ in the BK model. This expression should give mean reversion in the short rate when it is negative. That is, we expect that if $\sigma'(t) < 0$ (implying $\sigma(t)$ is decreasing) then the BDT model will give mean reversion. On the other hand, when $\sigma'(t) > 0$ (implying $\sigma(t)$ is increasing) the short rates in the BDT model will grow with no mean reversion. If $\sigma(t)$ is constant in the BDT model, then $\sigma'(t) = 0$ so $\rho = 0$ and equation (25a) becomes the KWF model (equation (15)). Therefore, we will only study the case of varying local volatility for the BDT model.

Eliminating the stochastic term in equation (25) leads to

$$\begin{aligned} d \ln r &= du = \left(\theta(t) + \frac{\sigma'(t)}{\sigma(t)} u \right) dt \\ &= \left(\theta(t) + \frac{\sigma'(t)}{\sigma(t)} \ln r \right) dt \end{aligned} \quad (26)$$

Solving this equation for u , as we did in the KF and BK models, gives us

$$u(t) = \left[\frac{u(0)}{\sigma(0)} + \int_0^t \frac{\theta(s)}{\sigma(s)} ds \right] \sigma(t)$$

or

$$r(t) = e^{\left(\frac{\log(r_0)}{\sigma_0} + \int_0^t \frac{\theta(s)}{\sigma(s)} ds\right)\sigma(t)} = e^{\frac{\sigma(t)\log(r_0)}{\sigma_0}} e^{\sigma(t) \int_0^t \frac{\theta(s)}{\sigma(s)} ds}$$

or

$$r(t) = r_0 e^{\frac{\sigma(t) - \sigma_0 \log(r_0)}{\sigma_0}} e^{\sigma(t) \int_0^t \frac{\theta(s)}{\sigma(s)} ds} \tag{27}$$

Note that the BDT mean short rate depends on the local volatility. If the local volatility has a decreasing structure, then the first exponential term in equation (27) has a negative exponent and will cause a decrease in the short rate and vice versa if the local volatility has an increasing structure. It is important to note that mean reversion in the BDT model comes from the local volatility structure (i.e., it is endogenous).

We now consider numerical solutions to the BDT process. To discretize equation (25a) for the BDT model we start off again by approximating du in equation (25b) by u to get

$$u_{k+1} = u_k + (\theta_k + \rho_k u_k)\tau + \sigma_k \varepsilon_k \sqrt{\tau} \tag{28}$$

The exponential of equation (28) gives us

$$r_{k+1} = r_k e^{[(\theta_k + \rho_k \ln r_k)\tau + \sigma_k \varepsilon_k \sqrt{\tau}]} \tag{29}$$

where

$$\rho_k = \frac{\sigma'_k}{\sigma_k}$$

We approximate this term by

$$\frac{\sigma_{k+1} - \sigma_k}{\tau} \frac{1}{\sigma_k}$$

That is, we approximate σ'_k by a discrete approximation to the derivative. We now have

$$u_{k+1} = u_k + \left(\theta_k + \frac{\sigma_{k+1} - \sigma_k}{\tau} \frac{1}{\sigma_k} u_k \right) \tau + \sigma_k \varepsilon_k \sqrt{\tau}$$

or

$$u_{k+1} = \frac{\sigma_{k+1}}{\sigma_k} u_k + \theta_k \tau + \sigma_k \varepsilon_k \sqrt{\tau} \tag{30}$$

If the random term is 0 equation (30) becomes

$$u_{k+1} = \frac{\sigma_{k+1}}{\sigma_k} u_k + \theta_k \tau \tag{31}$$

In particular, if

$$\frac{\sigma_{k+1}}{\sigma_k} = \alpha$$

where α is a constant then

$$u_k = \alpha^k u_0 + \sum_{j=0}^{k-1} \alpha^j \theta_{k-j-1} \tau$$

The exponential of this gives

$$r_k = r_0 e^{(\alpha^k - 1) \ln r_0} e^{\sum_{j=0}^{k-1} \alpha^j \theta_{k-j-1} \tau}$$

This equation is interesting because $\ln r_0 < 0$. If $\alpha > 1$ then the first exponential term decreases. When $\theta < 0$ the second exponential term also decreases and the BDT short rate should approach a target rate. Conversely, when $\theta > 0$ the second exponential term increases. In this case we can approach a target rate or the second term can dominate. If $\alpha < 1$ then a similar situation arises. Therefore, in order to get meaningful numerical results for the BDT short rates we strongly recommend that α be close to 1 and that the term structure of spot rates not have too large a slope.

The analysis of the equations without the stochastic term presented in this section is important. Recall that the characteristics of the random term are such that average influence of this term will be much smaller than the mean term in the SDEs. Consequently, the properties presented within this section will also hold under more general circumstances. The discrete approximations we developed for the models will be used to build the binomial and trinomial models in the next section. Note that we are highlighting the difference across the models and do not calibrate the models to market information.

For numerical reasons, the BK and HW models are best implemented in the trinomial framework. The HL, KWF, and BDT models are more easily implemented in the binomial framework.³ We will discuss the specifics of this in the next section. For the trinomial framework we use the approach of Hull and White (1994).

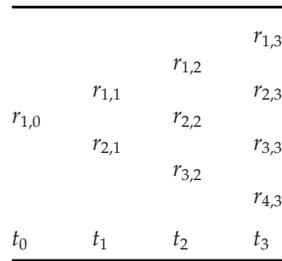


Figure 1 Binomial Lattice

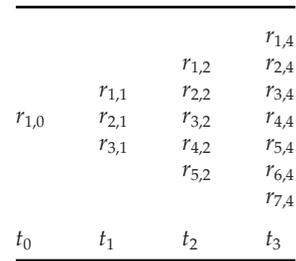


Figure 2 Trinomial Lattice

BINOMIAL AND TRINOMIAL SOLUTIONS TO THE STOCHASTIC DIFFERENTIAL EQUATIONS

In this section we present the *binomial* and *trinomial lattice models* that are obtained for the discretized versions of SDEs given in the previous section. The binomial method models the short rate in a geometrically analogous manner as equities.⁴ The up move has a probability q and so the down move has a probability of $1 - q$. We use $q = 0.5$ within the framework of risk neutrality. This binomial process of two possible moves for the short rate in the next time period is then continued at each time to produce a binomial lattice of interest rates.

The trinomial model is similar in spirit to the binomial except there are three possible states emanating from each node. From each point in time we call the upward-most move the “up move,” the downward-most move the “down move,” and the center move the “middle move.” The probabilities for an up move, middle move, and down move are given by q_1 , q_2 , and q_3 with $q_1 + q_2 + q_3 = 1$.

Interest rate lattices should possess the property of recombination for them to be computationally tractable. That is, from any given node in the binomial model we will require an up move followed by a down move to get to the same point as a down move followed by an up move. This ensures that the number of nodes in the binomial lattice increase by only one at each time step. In the trinomial case recombination

is a little more complicated. From any node in the trinomial lattice an up move followed by a down move will get to the same node as two successive middle moves and as a down move followed by an up move. This ensures that the number of nodes in the trinomial lattice increase by only two at each time step.

Figure 1 represents a binomial short rate lattice and Figure 2 represents a trinomial short rate lattice. The notation $r_{j,k}$ is used to denote the short rate value at level j at time t_k . In the binomial lattice, an up move from $r_{j,k}$ is given by $r_{j,k+1}$ and a down move is given by $r_{j+1,k+1}$. At time t_k there are $k + 1$ possible values for the short rate in the binomial lattice. That is, j ranges from 1 to $k + 1$. In the trinomial model, an up move, middle move, and down move from the short rate $r_{j,k}$ are given by $r_{j,k+1}$, $r_{j+1,k+1}$, and $r_{j+2,k+1}$, respectively. In the trinomial model there are $2k + 1$ possible values for the short rate at time t_k . That is, j ranges from 1 to $2k + 1$. The short rates forming the top of the lattice will be called the up state for the short rates and the short rates forming the bottom of the lattice will be called the down state for the short rates. For the binomial and trinomial model, the up state is the set of short rates $r_{1,k}$ for $0 \leq k \leq n$ and the down state for the binomial case is the set of short rates $r_{k,k}$ for $0 \leq k \leq n$; within the trinomial tree the down state is the set of short rates $r_{2k+1,k}$ for $0 \leq k \leq n$.

Hull-White Binomial Lattice

Since the HW model is a more general version of the HL model we present the binomial version

only for the HW. In the HW binomial lattice the expressions for $r_{j,k}$ that correspond to equation (13) are

$$r_{j,k+1} = r_{j,k} + \theta_k \tau_k - \phi_k r_{j,k} \tau_k + \sigma_k \sqrt{\tau_k} \quad (32)$$

for an up move and

$$r_{j+1,k+1} = r_{j,k} + \theta_k \tau_k - \phi_k r_{j,k} \tau_k - \sigma_k \sqrt{\tau_k} \quad (33)$$

for a down move. (We are using τ_k for Δt_k .)

These equations suggest that in order to have recombination the following must be true:

$$\tau_{k+1} = \tau_k \frac{4 \left(\frac{\sigma_k}{\sigma_{k+1}} \right)^2}{\left[1 + \sqrt{1 + 4 \left(\frac{\sigma_k}{\sigma_{k+1}} \right)^2 \tau_k \phi_{k+1}} \right]^2} \quad (34)$$

Equation (34) illustrates that if you want a constant time step when the local volatility is constant, the mean reversion must be 0. The recombination requirement has put the stringent condition on the HW binomial lattice that the mean reversion is determined by the local volatility. To avoid this problem within the binomial framework we must allow the time step to vary with k in equations (32) through (34). As a matter of fact, for a constant time step,

$$\phi_{k+1} = \frac{\sigma_k - \sigma_{k+1}}{\sigma_k \tau} \quad (35)$$

which can also be solved for σ_{k+1} to give

$$\sigma_{k+1} = \sigma_k (1 - \phi_{k+1} \tau) \quad (36)$$

Equation (36) shows that the mean reversion can be used to match any given local volatility for a constant time step. If the local volatility is decreasing the mean reversion will be positive, and if the local volatility is increasing the mean reversion will be negative. We point out that if a variable time step is used, one does not have to have mean reversion match local volatility.

Black-Karasinski Binomial Lattice

Since the BK model is a more general form of the KWF model, we only present the binomial version for the BK model. The expressions corre-

sponding to equations (32) and (33) of the HW model and from equation (23b) are

$$r_{j,k+1} = r_{j,k} e^{(\theta_k - \phi_k \ln(r_{j,k})) \tau_k + \sigma_k \sqrt{\tau_k}} \quad (37)$$

for an up move and

$$r_{j+1,k+1} = r_{j,k} e^{(\theta_k - \phi_k \ln(r_{j,k})) \tau_k - \sigma_k \sqrt{\tau_k}} \quad (38)$$

for a down move.

Using equations (37) and (38) we can develop equations for the BK binomial lattice that are identical to equations (34) and (36) for the HW binomial lattice. This should be expected since the BK SDE is just a lognormal version of the HW SDE. A crucial point here is that we can use the HW and BK models to match local volatility and to compare results. It is important to point out that the HW and BK binomial lattices have a constant time step. If a variable time step is used, then interpolation is required to give the short rates at the fixed time steps. We do not offer this framework. Instead we present the HW and the BK models in the trinomial framework.

Within the binomial framework, the HW and BK models only approximate the distributional properties of their respective SDEs. The accuracy of the approximation is a function of the mean reversion. As the mean reversion increases, the accuracy decreases. Note that since the HL and KWF models have a zero mean reversion the distributional characteristics of their SDEs are perfectly matched within the binomial framework. This is the reason for using the trinomial method for the HW and BK models.

The Trinomial Lattices

A better way to keep a constant time step and to match the appropriate distributional properties is to use a trinomial lattice instead of a binomial lattice. If we use a trinomial lattice for the HW SDEs, then from equation (13) we use

$$r_{j,k+1} = r_{j,k} + \theta_k \tau - \phi_k r_{j,k} \tau + \alpha_k \sigma_k \sqrt{\tau} \quad (39a)$$

for an up move,

$$r_{j+2,k+1} = r_{j,k} + \theta_k \tau - \phi_k r_{j,k} \tau - \alpha_k \sigma_k \sqrt{\tau} \quad (39b)$$

for a down move, and

$$r_{j+1,k+1} = r_{j,k} + \theta_k \tau - \phi_k r_{j,k} \tau \quad (39c)$$

for a middle move. Similarly, if we use a trinomial lattice for the BK SDEs then from equation (23b) we use

$$r_{j,k+1} = r_{j,k} e^{(\theta_k - \phi_k \ln(r_{j,k}))\tau + \alpha_k \sigma_k \sqrt{\tau}} \quad (40a)$$

for an up move,

$$r_{j+2,k+1} = r_{j,k} e^{(\theta_k - \phi_k \ln(r_{j,k}))\tau - \alpha_k \sigma_k \sqrt{\tau}} \quad (40b)$$

for a down move, and

$$r_{j+1,k+1} = r_{j,k} e^{(\theta_k - \phi_k \ln(r_{j,k}))\tau} \quad (40c)$$

for a middle move.

Note that a constant time step is now used. The expression α_k is used to guarantee recombination. The probabilities of an up, middle, and down move are chosen to give the correct variance.

The No Arbitrage Equations

The procedure to generate the no arbitrage equations for the binomial and trinomial lattices is outlined in the appendix. The no arbitrage polynomial for the short rates in the binomial tree is given by

$$f_i = c_{1,i} \prod_{j=1}^i (1 + r_{j,i} \tau) + \sum_{m=1}^i c_{m+1,i} \prod_{\substack{n=1 \\ n \neq m}}^i (1 + r_{n,i} \tau) \quad (41)$$

where, for $i \geq 3$

$$a_{1,i} = \prod_{n=0}^{i-1} \prod_{m=1}^i (1 + r_{m,n} \tau)$$

$$\begin{aligned} a_{2,i} &= b_{1,i-1}, a_{j,i} = b_{j-2,i-1} + b_{j-1,i-1}, & \text{for } j = 3, \dots, i, \\ a_{i+1,j} &= b_{i-1,i-1}, & \text{and } c_{1,i} = P_{i+1} a_{1,i}, \\ c_{j+1,i} &= q^{i-j} (1-q)^{j-1} a_{j+1,i} & \text{for } j = 1, \dots, i. \end{aligned}$$

We solve equation (41) for θ_i by setting $f_i = 0$. We then use θ_i to compute $r_{j,i}$ for $j = 1, \dots, i$ at the i th period. The bisection method will converge quickly because there is

only one root between -1 and 1 for the HW binomial lattice and one root between 0 and 1 for the BK binomial lattice.⁵

After generating the new rates we let

$$b_{j,i} = \alpha_{j+1,i} \prod_{\substack{m=1 \\ m \neq j}}^i (1 + r_{m,i} \tau)$$

For the variable time step, τ_i we replace the terms $(1 + r_{j,i} \tau)$ by $(1 + r_{j,i} \tau)^{\tau_i/\tau}$ and the terms $(1 + r_{n,i} \tau)$ by

$$(1 + r_{n,i} \tau)^{\tau_i/\tau}$$

in equation (41).

Similarly, the no arbitrage polynomial for the trinomial trees is given by

$$f_i = c_{1,i} \prod_{j=1}^{2i-1} (1 + r_{j,i} \tau) + \sum_{m=1}^{2i-1} c_{m+1,i} \prod_{\substack{n=1 \\ n \neq m}}^{2i-1} (1 + r_{n,i} \tau) \quad (42)$$

where we first let

$$a_{1,i} = \prod_{j=1}^{2i-3} (1 + r_{j,i} \tau)$$

$$a_{2,i} = q_1 b_{1,i-1} a_{2,i-1}, a_{3,i} = q_2 b_{1,i-1} a_{2,i-1} + q_1 b_{2,i-1} a_{3,i-1}$$

$$a_{j,i} = q_3 b_{j-3,i-1} a_{j-2,i-1} + q_2 b_{j-2,i-1} a_{j-1,i-1} + q_1 b_{j-1,i-1} a_{j,i-1},$$

for $j = 4, \dots, 2i - 2$,

$$\begin{aligned} a_{2i-1,i} &= q_3 b_{2i-4,i-1} a_{2i-3,i-1} + q_2 b_{2i-3,i-1} a_{2i-2,i-1} a_{2i,i} \\ &= q_3 b_{2i-3,i-1} a_{2i-2,i-1} \end{aligned}$$

and then let

$$c_{1,i} = P_{i+1} a_{1,i}, c_{j,i} = a_{j,i} \text{ for } j = 2, \dots, 2i + 1$$

We solve equation (42) for θ_i by setting $f_i = 0$ using the bisection method. From this the short rates for either the HW or BK trinomial lattices are determined at step i . We then let

$$b_n = \prod_{\substack{j=1 \\ j \neq n}}^{2i-1} (1 + r_{j,i} \tau)$$

for $n = 1, \dots, 2i - 1$ and then repeat the process. In these derivations $P_i = 1/(1 + R_i \tau)^i$ is the discount factor given by the spot rates (zero curve).

The Hull and White Lattice

We now briefly outline the Hull and White methodology for generating HW and BK trinomial lattices.⁶ The Hull and White methodology uses

$$r_{j,k} = x + (j_k)\Delta\rho \tag{43}$$

for the HW trinomial lattice short rates and

$$r_{j,k} = e^{[x+(j_k)\Delta\rho]} \tag{44}$$

for the BK trinomial lattice short rates.

They choose $\Delta\rho = \sigma\sqrt{3\tau}$ to minimize numerical error and introduce the mean reversion through the probabilities $q_1, q_2,$ and q_3 . Specifically, they use

$$q_1 = \frac{1}{6} + \frac{(j_k)^2\phi^2\tau^2 + (j_k)\phi\tau}{2}$$

$$q_2 = \frac{2}{3} - (j_k)^2\phi^2\tau^2$$

and

$$q_3 = \frac{1}{6} + \frac{(j_k)^2\phi^2\tau^2 - (j_k)\phi\tau}{2}$$

for the up, middle, and down moves at $r_{j,k}$, respectively, since this matches the expected change and variance of the short rate over the next time period. However, as they point out, these probabilities must remain positive. In order to do this they “prune” the upper and lower branches of their lattice at the level j that keeps these probabilities positive. Since q_2 is the only one that can become negative they require the following

$$j < \frac{\sqrt{6}}{3\phi\tau} \approx \frac{0.816}{\phi\tau}$$

At this maximum value of j , Hull and White apply a different branching procedure with different probabilities in order to “prune” the lattice. However, as they point out, using this value of j can lead to computational problems so they actually use the first j satisfying

$$j_k > \frac{3 - \sqrt{6}}{3\phi\tau} \approx \frac{0.184}{\phi\tau}$$

This leads to a reduction in the spread of the rates.

COMPARATIVE STUDY OF THE NUMERICAL SOLUTIONS

In this section a comparison between the methodologies is presented. In particular, we look at the effects of mean reversion and local volatility on the drift and the spread in the short rates. We present numerical results for the term structures, volatility, and mean reversion in Table 1. The table also includes the bond information for use later.

Original Term Structure with No Mean Reversion

We first consider the original term structure with no mean reversion for the HL and HW models. In Figure 3 we present the binomial tree for the HL model and the trinomial for the HW model using the HW trinomial methodology. We use a 10% volatility throughout the trees. We see that the spread in the short rates increases over time in the models as expected.

Table 1 Input Information

Original TS	Volatility	Mean Reversion
6.20%	10.00%	5%
6.16%	10.00%	
6.15%	9.00%	
6.09%	9.00%	
6.02%	8.00%	
6.02%	8.00%	
6.01%	7.00%	
6.01%	7.00%	
6.00%	7.00%	
6.01%	7.00%	
Bond Information for ED, EC, and OAS		
Call Price (Regular Callable)		\$102.50
Put Price (Regular Puttable)		\$95.00
Annual Coupon (\$ per \$100)		\$6.00
Time Option Starts (years from now)		1

a. The Ho-Lee Interest Rate Lattice

									136.31%	
								118.42%	116.31%	
						85.20%	101.50%	98.42%	96.31%	
				54.99%	69.85%	65.20%	81.50%	78.42%	76.31%	
			41.49%	34.99%	49.85%	45.20%	61.50%	58.42%	56.31%	
		28.93%	21.49%	14.99%	29.85%	25.20%	41.50%	38.42%	36.31%	
	6.20%	17.05%	8.93%	1.49%	9.85%	5.20%	21.50%	18.42%	16.31%	
		-2.95%	-11.07%	-5.01%	-10.15%	5.20%	1.50%	1.50%	1.50%	
			-18.51%	-25.01%	-14.80%	-18.50%	-1.58%	-1.58%	-1.58%	
					-30.15%	-34.80%	-18.50%	-21.58%	-23.69%	
							-38.50%	-41.58%	-43.69%	
Time in Years	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0

b. The Hull-White Trinomial Interest Rate Lattice Using the HW Method with No Mean Reversion

										203.31%
									172.58%	185.99%
								142.30%	155.26%	168.67%
							131.70%	124.98%	137.94%	151.35%
						107.78%	114.38%	107.66%	120.62%	134.03%
					84.92%	90.46%	97.06%	90.34%	103.30%	116.71%
				63.71%	67.60%	73.14%	79.74%	73.02%	85.98%	99.39%
			43.65%	46.38%	50.28%	55.82%	62.42%	55.70%	68.66%	82.07%
		24.39%	26.33%	29.06%	32.96%	38.50%	45.10%	38.38%	51.34%	64.75%
	6.20%	7.07%	9.01%	11.74%	15.64%	21.18%	27.78%	21.06%	34.02%	47.43%
		-10.25%	-8.31%	-5.58%	-1.68%	3.86%	10.46%	3.74%	16.70%	30.11%
			-25.63%	-22.90%	-19.00%	-13.46%	-6.86%	-13.58%	-0.62%	12.79%
				-40.22%	-36.32%	-30.78%	-24.18%	-30.90%	-17.94%	-4.53%
					-53.64%	-48.10%	-41.50%	-48.22%	-35.26%	-21.85%
						-65.42%	-58.83%	-65.54%	-52.58%	-39.18%
							-76.15%	-82.86%	-69.90%	-56.50%
								-100.18%	-87.22%	-73.82%
									-104.54%	-91.14%
										-108.46%
Time in Years	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0

Figure 3 The HL Binomial and HW Trinomial Trees for the Original Term Structure with No Mean Reversion

We also see that the HL model can give negative short rates.

In Figure 4 we present the binomial tree for the KWF model, the trinomial for the BK model using the HW trinomial methodology, and the BDT binomial model. The KWF and BK models use the 10% volatility throughout the tree and

no mean reversion. Note the volatile nature of the BDT model. This is due to the time varying volatility structure and the way mean reversion is incorporated into the BDT model through this decreasing volatility structure. Note that all the short rates are positive and that the spread in the rates is significantly less than in Figure 3.

a. The Kalotay, Williams, and Fabozzi Interest Rate Lattice

									12.92%	14.72%
								11.87%	12.05%	
						10.65%		10.58%		
				8.43%	9.76%	8.72%		9.72%	9.87%	
			7.89%	7.99%	8.72%			7.96%	8.66%	8.08%
		7.44%	6.90%	7.14%				7.96%	7.09%	
	6.73%	6.46%	6.54%	6.52%				6.52%	6.61%	
	6.20%	6.09%	5.65%	5.84%				5.34%	5.81%	
		5.51%	5.29%	5.36%				5.34%	5.41%	
			4.98%	4.78%				4.37%	4.75%	
			4.33%	4.39%				4.37%	4.43%	
				3.79%	3.92%			3.58%	3.89%	
					3.59%			3.58%	3.63%	
						3.21%		2.93%	3.19%	2.97%
									2.61%	2.43%
Time in Years	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0

b. The Black-Karasinski Trinomial Interest Rate Lattice Using the HW Method with No Mean Reversion

									23.21%	28.45%
									23.92%	
								19.82%	19.52%	20.12%
						14.08%		16.52%	16.67%	16.92%
					11.31%	11.84%		13.89%	14.02%	13.80%
				9.82%	9.51%	9.96%		9.83%	9.92%	9.76%
			8.60%	8.26%	8.00%	8.37%		8.26%	8.34%	8.21%
		7.25%	7.23%	6.95%	6.73%	7.04%		6.95%	7.01%	6.90%
	6.20%	6.09%	6.08%	5.75%	5.66%	5.92%		5.84%	5.90%	5.81%
		5.12%	5.11%	4.91%	4.76%	4.98%		4.91%	4.96%	4.88%
			4.30%	4.13%	4.00%	4.19%		4.13%	4.17%	4.11%
				3.47%	3.37%	3.52%		3.48%	3.51%	3.45%
					2.83%	2.96%		2.92%	2.95%	2.90%
						2.49%		2.46%	2.48%	2.44%
							2.07%	2.09%	2.05%	2.12%
								1.75%	1.73%	1.78%
									1.45%	1.50%
										1.26%
Time in Years	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0

Figure 4 The BDT and KWF Binomial and the BK Trinomial Trees for the Original Term Structure with No Mean Reversion

Table 2 presents the trinomial lattices for the HW and BK models using the information in Table 1 and a mean reversion of 5%. The volatility is 10%. Notice the pruning that takes place within the lattice when we have mean reversion. This produces lattices that are significantly

different from those shown in Figures 3 and 4. This is a peculiarity of the Hull and White methodology. The pruning is a result of incorporating mean reversion into the model and ensuring that the distributional characteristics of the SDEs are retained.

c. The Black, Derman, and Toy Interest Rate Model

								10.29%	11.39%
							6.47%	9.89%	
					7.36%	9.52%	6.34%	8.59%	
			7.24%	8.12%	6.79%	8.10%	6.21%	7.76%	7.46%
		7.44%	6.30%	6.78%	6.26%	6.90%	6.08%	6.74%	6.47%
6.20%	6.73%	6.09%	5.49%	5.66%	5.77%	5.88%	5.95%	5.85%	5.62%
	5.51%	4.98%	4.78%	4.73%	5.32%	5.00%	5.83%	5.09%	4.88%
				3.95%	4.91%	4.26%	5.71%	4.42%	4.24%
						3.63%	5.59%	3.84%	3.68%
								3.33%	3.20%
Time in Years	1	2	3	4	5	6	7	8	9

Figure 4 (Continued)

Table 2 Trinomial Model

a. The Hull-White Trinomial Interest Rate Lattice Using the HW Method with Mean Reversion of 5%

				83.50%	87.60%	91.92%	96.84%	101.89%	107.24%
			63.14%	66.18%	70.28%	74.60%	79.52%	84.57%	89.91%
		43.51%	45.82%	48.86%	52.96%	57.28%	62.20%	67.25%	72.59%
6.20%	24.39%	26.18%	28.50%	31.54%	35.64%	39.96%	44.88%	49.93%	55.27%
	7.07%	8.86%	11.17%	14.22%	18.32%	22.64%	27.56%	32.61%	37.95%
	-10.25%	-8.46%	-6.15%	-3.10%	1.00%	5.32%	10.24%	15.29%	20.63%
		-25.78%	-23.47%	-20.42%	-16.32%	-12.00%	-7.09%	-2.03%	3.31%
			-40.79%	-37.75%	-33.64%	-29.32%	-24.41%	-19.35%	-14.01%
				-55.07%	-50.96%	-46.64%	-41.73%	-36.67%	-31.33%
Time in Years	1	2	3	4	5	6	7	8	9

b. The Black-Karasinski Trinomial Interest Rate Lattice Using the HW Method with Mean Reversion of 5%

				11.34%	11.87%	11.73%	11.84%	11.67%	12.03%
			9.83%	9.53%	9.99%	9.86%	9.96%	9.81%	10.12%
		8.60%	8.27%	8.02%	8.40%	8.29%	8.38%	8.25%	8.51%
6.20%	7.25%	7.26%	6.95%	6.74%	7.06%	6.98%	7.04%	6.94%	7.16%
	6.09%	6.08%	5.85%	5.67%	5.94%	5.87%	5.92%	5.84%	6.02%
	5.12%	5.11%	4.92%	4.77%	4.99%	4.93%	4.98%	4.91%	5.06%
		4.30%	4.14%	4.01%	4.20%	4.15%	4.19%	4.13%	4.26%
			3.48%	3.37%	3.53%	3.49%	3.52%	3.47%	3.58%
				2.84%	2.97%	2.93%	2.96%	2.92%	3.01%
Time in Years	1	2	3	4	5	6	7	8	9

Table 3 Effective Duration and Effective Convexity Results

Shift==>	-500 bp		-250 bp		Current		250 bp		500 bp	
	Eff. Duration	Eff. Convexity								
<i>Ho Lee</i>										
Callable Bond	3.72119	-31.15230	3.62427	10.51371	3.43354	9.58153	4.19081	-6.18888	4.18588	12.92063
Putable Bond	6.48070	55.51213	5.96968	26.45835	4.82856	41.73014	4.33750	17.68955	3.52379	15.98202
<i>BDT</i>										
Callable Bond	0.98815	0.97643	0.96433	0.92992	5.72746	-100.52077	6.97619	31.91884	6.59872	29.24115
Putable Bond	8.15290	41.20380	7.75444	37.88876	6.94320	136.25219	0.91997	0.84634	0.89929	0.80871
<i>KWF</i>										
Callable Bond	0.98815	0.97643	0.96433	0.92992	5.48099	-8.70115	6.90354	18.94888	6.59875	29.22747
Putable Bond	8.15311	41.26110	7.75438	37.97492	6.02987	132.82680	0.91997	0.84634	0.89929	0.80871
<i>HW-HW</i>										
Callable Bond	3.35706	5.81085	3.24446	8.80890	3.33140	9.55382	3.46677	-9.19552	4.65946	14.99510
Putable Bond	5.82483	23.71025	5.33913	20.81987	4.79375	17.78372	4.14647	14.50538	3.30034	10.76225
<i>BK-HW</i>										
Callable Bond	0.98815	0.97643	0.96433	0.92992	5.21624	-77.28716	6.93694	31.17366	6.56855	28.88729
Putable Bond	8.09134	40.58931	7.70100	37.39723	6.79269	72.05773	0.91997	0.84634	0.89929	0.80871

Comparison of the Models Using Common Risk and Value Metrics

Here we contrast the effective duration, effective convexity, and the option-adjusted spread (OAS) for 10-year callable and putable bonds each with a one-year delay on the embedded option. The information in Table 1 is used for the analysis. We computed the effective duration for the original term structures shown in Table 1 using a yield change of 25 basis points. The original term structure is then shifted up and down in a parallel manner by ± 250 basis points and by ± 500 basis points, respectively. In other words, we computed the effective duration at five different term structure levels using a yield change of 25 basis points.

Table 3 presents the effective duration and convexity results for the two securities for each model. The results are interesting. It is clear that the normal models do not agree with the lognormal models. Specifically, the normal models do not match the characteristics of the price yield

relationship at extreme interest rate levels.⁷ Furthermore, each model gives slightly different results. This is an important finding and must be appreciated by any user of these models.

Table 4 presents the OAS results. We used a market price that is 3% below the model price for the OAS computation. They are consistent with the results in Table 3. Note that the normal models produce OAS values larger than any of the lognormal models. This is due to the distributional differences and the property of allowing very low and negative interest rates. Clearly, normal models are not desirable when evaluating securities with embedded options.⁸

APPENDIX

In this appendix we outline how to obtain equations (41) and (42). For equation (41) we use Figure 1. For equation (42) we use Figure 2.

Table 4 Option-Adjusted Spread Results

	Ho-Lee	BDT	KWF	HW-HW	BK-HW
Callable Bond	0.8454%	0.4785%	0.5449%	0.8350%	0.5063%
Putable Bond	0.5884%	0.4732%	0.5249%	0.5688%	0.4774%

We first solve for $r_{1,1}$ and $r_{2,1}$ in Figure 1. Equating the price from the spot rate term structure with the price from the binomial lattice gives us

$$P_2 = \frac{1}{(1 + R_2\tau)^2} = \frac{qp_{1,1} + (1 - q)p_{2,1}}{1 + r_{1,0}\tau} \quad (\text{A1})$$

Substituting in the discount factors $p_{j,1} = 1/(1 + r_{j,1}\tau)$ for $j = 1, 2$ and clearing fractions we obtain

$$P_2(1 + r_{1,0}\tau)(1 + r_{1,1}\tau)(1 + r_{2,1}\tau) - q(1 + r_{2,1}\tau) - (1 - q)(1 + r_{1,1}\tau) = 0 \quad (\text{A2})$$

We let $r_{1,0} = R_1$. This equation can now be solved for θ_1 .

For the next period in the binomial lattice we have from Figure 1 that

$$P_3 = \frac{1}{(1 + R_3\tau)^3} = \frac{qp_{1,1} + (1 - q)p_{2,1}}{1 + r_{1,0}\tau} = \frac{q\left(\frac{qp_{1,2} + (1 - q)p_{2,2}}{1 + r_{1,1}\tau}\right) + (1 - q)\left(\frac{qp_{2,2} + (1 - q)p_{3,2}}{1 + r_{2,1}\tau}\right)}{1 + r_{1,0}\tau}$$

which reduces to

$$\begin{aligned} &P_3(1 + r_{1,0}\tau)(1 + r_{1,1}\tau)(1 + r_{2,1}\tau)(1 + r_{1,2}\tau) \\ &\quad \times (1 + r_{2,2}\tau)(1 + r_{3,2}\tau) \\ &- q^2(1 + r_{2,1}\tau)(1 + r_{2,2}\tau)(1 + r_{3,1}\tau) \\ &\quad - q(1 - q)[(1 + r_{1,1}\tau) + (1 + r_{2,1}\tau)] \\ &\quad \times (1 + r_{1,2}\tau)(1 + r_{3,2}\tau) - (1 - q)^2(1 + r_{1,1}\tau) \\ &\quad \times (1 + r_{1,2}\tau)(1 + r_{2,2}\tau) = 0 \end{aligned} \quad (\text{A3})$$

We now solve equation (A3) for θ_2 using the bisection method.

From equation (A2) and equation (A3) we can generate the remainder of the no arbitrage equations that give the short rates in the binomial lattice. Note that equation (A2) can be written as

$$c_{1,1}(1 + r_{1,1}\tau)(1 + r_{2,1}\tau) + c_{2,1}(1 + r_{2,1}\tau) + c_{3,1}(1 + r_{1,1}\tau) = 0 \quad (\text{A4})$$

and that equation (A3) can be written as

$$\begin{aligned} &c_{1,2}(1 + r_{1,2}\tau)(1 + r_{2,2}\tau)(1 + r_{3,2}\tau) \\ &\quad + c_{2,2}(1 + r_{2,2}\tau)(1 + r_{3,2}\tau) + c_{3,2}(1 + r_{1,2}\tau) \\ &\quad \times (1 + r_{3,2}\tau) + c_{4,2}(1 + r_{1,2}\tau)(1 + r_{2,2}\tau) = 0 \end{aligned} \quad (\text{A5})$$

We now introduce some variables that will help to generate the coefficients $c_{i,k}$ for the polynomials that determine the interest rates at time period k . We start by doing it for the polynomials in equations (A4) and (A5). This is done in two steps. The first step is to notice how the coefficients are related to the interest rates at the previous time periods. Note that if we let $a_{1,1} = 1 + r_{1,0}\tau$, $a_{2,1} = -1$, and $a_{3,1} = -1$ then $c_{1,1} = P_2a_{1,1}$, $c_{2,1} = qa_{2,1}$, and $c_{3,1} = (1 - q)a_{3,1}$ in equation (A4). In order to generate equation (A5) we first let $b_{1,1} = a_{2,1}(1 + r_{2,1}\tau)$, $b_{2,1} = a_{3,1}(1 + r_{1,1}\tau)$. We can then generate $a_{1,2} = (1 + r_{1,0}\tau)(1 + r_{1,1}\tau)(1 + r_{2,1}\tau)$, $a_{2,2} = b_{1,1}$, $a_{3,2} = b_{1,1} + b_{2,1}$, and $a_{4,2} = b_{2,1}$. It is now seen that $c_{1,2} = P_3a_{1,2}$, $c_{2,2} = q^2a_{2,2}$, $c_{3,2} = q(1 - q)a_{3,3}$, and $c_{4,2} = (1 - q)^2a_{4,2}$. We now let $b_{1,2} = a_{3,1}(1 + r_{2,2}\tau)$

$(1 + r_{3,2}\tau)$, $b_{2,2} = a_{3,2}(1 + r_{1,2}\tau)(1 + r_{3,2}\tau)$, and $b_{3,2} = a_{4,2}(1 + r_{1,2}\tau)(1 + r_{2,2}\tau)$ and continue the process to obtain equation (41).

For the trinomial lattice no arbitrage polynomial we first solve for $r_{1,1}$, $r_{2,1}$, and $r_{3,1}$ in Figure 2. Equating the price from the spot rate term structure with the price from the trinomial lattice gives us

$$P_2 = \frac{1}{(1 + R_2\tau)^2} = \frac{q_1p_{1,1} + q_2p_{2,1} + q_3p_{3,1}}{1 + r_{1,0}\tau}$$

which is similar to equation (A1). Proceeding as in the binomial lattice we find that

$$\begin{aligned} &P_2(1 + r_{1,0}\tau)(1 + r_{1,1}\tau)(1 + r_{2,1}\tau)(1 + r_{3,1}\tau) \\ &\quad - q_1(1 + r_{2,1}\tau)(1 + r_{3,1}\tau) - q_2(1 + r_{1,1}\tau) \\ &\quad \times (1 + r_{3,1}\tau) - q_3(1 + r_{1,1}\tau)(1 + r_{2,1}\tau) = 0 \end{aligned} \quad (\text{A6})$$

As in the binomial case, $r_{1,0} = R_1$ and equation (A6) is solved for θ_1 using the bisection method.

For the next period in the trinomial lattice (Figure 2) gives us

$$P_3 = \frac{1}{(1 + R_3\tau)^3} = \frac{q_1 p_{1,1} + q_2 p_{2,1} + q_3 p_{3,1}}{1 + r_{1,0}\tau}$$

$$= \frac{q_1 \left(\frac{q_1 p_{1,2} + q_2 p_{2,2} + q_3 p_{3,2}}{1 + r_{1,1}\tau} \right) + q_2 \left(\frac{q_1 p_{2,2} + q_2 p_{3,2} + q_3 p_{3,3}}{1 + r_{2,1}\tau} \right) + q_3 \left(\frac{q_1 p_{3,3} + q_2 p_{3,4} + q_3 p_{3,5}}{1 + r_{3,1}\tau} \right)}{1 + r_{1,0}\tau}$$

which simplifies to the following equation similar to equation (A3)

$$P_3 (1 + r_{1,0}\tau) \prod_{j=1}^3 (1 + r_{j,1}\tau) \prod_{j=1}^5 (1 + r_{j,2}\tau)$$

$$- q_1^2 (1 + r_{2,1}\tau) (1 + r_{3,1}\tau) \prod_{j=2}^5 (1 + r_{j,2}\tau)$$

$$- [q_1 q_2 (1 + r_{2,1}\tau) (1 + r_{3,1}\tau) + q_1 q_2 (1 + r_{1,1}\tau)$$

$$\times (1 + r_{3,1}\tau)] \prod_{\substack{j=1 \\ j \neq 2}}^5 (1 + r_{j,2}\tau)$$

$$- [q_1 q_3 (1 + r_{2,1}\tau) (1 + r_{3,1}\tau) + q_2^2 (1 + r_{1,1}\tau)$$

$$\times (1 + r_{3,1}\tau) + q_3 q_1 (1 + r_{1,1}\tau) (1 + r_{2,1}\tau)]$$

$$\times \prod_{\substack{j=1 \\ j \neq 3}}^5 (1 + r_{j,2}\tau) \tag{A7}$$

$$- [q_2 q_3 (1 + r_{1,1}\tau) (1 + r_{3,1}\tau) + q_3 q_2 (1 + r_{1,1}\tau)$$

$$\times (1 + r_{2,1}\tau)] \prod_{\substack{j=1 \\ j \neq 4}}^5 (1 + r_{j,2}\tau)$$

$$- q_3^2 (1 + r_{1,1}\tau) (1 + r_{2,1}\tau) \prod_{j=1}^4 (1 + r_{j,2}\tau) = 0$$

Equation (A7) is also solved for θ_2 using the bisection method. We now proceed as in the binomial lattice case to generate the no arbitrage equation for θ_i given in equation (42).

KEY POINTS

- Interest rates are commonly modeled using stochastic differential equations.
- One-factor models use a stochastic differential equation to represent the short rate and

two-factor models use a stochastic differential equation for both the short rate and the long rate.

-
- The stochastic differential equations used to model interest rates must capture some of the market properties of interest rates such as mean reversion and/or a volatility that depends on the level of interest rates.
 - The approaches used to implement the SDEs into a term structure model include equilibrium and no arbitrage.
 - There are five different term structure models that evolve from three general stochastic differential equations.
 - Without market calibration the models produce very different results.
 - Both the end user and the developer must be aware of these properties in order to properly implement and interpret any results from the models.
 - Even with calibration the models can produce different results. Calibration reduces the differences across the models but does not eliminate them.

NOTES

1. Ho and Lee (1986).
2. Kalotay, Williams, and Fabozzi (1993).
3. See Buetow and Sochacki (2001).
4. See, for example, Cox, Ross, and Rubinstein (1979).
5. See Burden and Faires (1998).
6. For complete details see Hull and White (1994).

7. See Fabozzi, Buetow, and Johnson (2012) for more details on the behavior of putable and callable bonds.
8. Details of these phenomena are provided in Buetow, Hanke, and Fabozzi (2001).

REFERENCES

- Backus, D., Foresi, S., and Telmer, C. (2001). Affine term structure models and the forward premium anomaly. *Journal of Finance* 56: 279–304.
- Black, F., Derman, E., and Toy, W. (1990). A one factor model of interest rates and its application to the Treasury bond options. *Financial Analysts Journal* 46: 33–39.
- Black, F., and Karasinski, P. (1991). Bond and option pricing when short rates are lognormal. *Financial Analysts Journal* 47: 52–59.
- Brennan, M., and Schwartz, E. (1979). A continuous time approach to the pricing of bonds. *Journal of Banking and Finance* 3: 133–155.
- Brennan, M., and Schwartz, E. (1982). An equilibrium model of bond pricing and a test of market efficiency. *Journal of Financial and Quantitative Analysis* 17: 301–329.
- Buetow, G. W., Hanke, B., and Fabozzi, F. J. (2001). The impact of different interest rate models on effective duration, effective convexity and option-adjusted spreads. *Journal of Fixed Income* 11: 41–53.
- Buetow, G. W., and Sochacki, J. (2001). *Binomial Interest Rate Models*. Charlottesville, VA: AIMR Research Foundation.
- Burden, R. L., and Faires, D. (1998). *Numerical Methods*, 2nd ed. Pacific Grove, CA: Brooks/Cole.
- Cox, J., Ingersoll, J., and Ross, S. (1985). A theory of the term structure of interest rates. *Econometrica* 53: 385–408.
- Cox, J., Ross, S., and Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of Financial Economics* 7: 229–264.
- Fabozzi, F. J., Buetow, G. W., and Johnson, R. (2012). Measuring interest rate risk. In F. J. Fabozzi (Ed.), *The Handbook of Fixed Income Securities*, 8th ed. New York: McGraw-Hill.
- Heath, D., Jarrow, R., and Morton, A. (1992). Bond pricing and the term structure of interest rates: A new methodology. *Econometrica* 60: 77–105.
- Ho, T., and Lee, S. (1986). Term structure movements and pricing interest rate contingent claims. *Journal of Finance* 41: 1011–1029.
- Hull, J. (2000). *Options, Futures, and Other Derivatives*, 4th ed. Upper Saddle River, NJ: Prentice Hall.
- Hull, J., and White, A. (1990). Pricing interest rate derivative securities. *Review of Financial Studies* 3: 573–592.
- Hull, J., and White, A. (1993). One factor interest rate models and the valuation of interest rate derivative securities. *Journal of Financial and Quantitative Analysis* 28: 235–254.
- Hull, J., and White, A. (1994). Numerical procedures for implementing term structure models I: Single-factor models. *Journal of Derivatives* 2: 7–16.
- Kalotay, A., Williams, G., and Fabozzi, F. J. (1993). A model for the valuation of bonds and embedded options. *Financial Analysts Journal* 49: 35–46.
- Longstaff, F. (1989). A non-linear general equilibrium model of the term structure of interest rates. *Journal of Financial Economics* 23: 195–224.
- Longstaff, F. (1992). Multiple equilibria and term structure models. *Journal of Financial Economics* 32: 333–344.
- Longstaff, F., and Schwartz, E. (1992). Interest rate volatility and the term structure: A two-factor general equilibrium model. *Journal of Finance* 47: 1259–1282.
- Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics* 5: 177–188.

Trading Cost Models

Modeling Market Impact Costs

PETTER N. KOLM, PhD

Director of the Mathematics in Finance Masters Program and Clinical Associate Professor,
Courant Institute of Mathematical Sciences, New York University

FRANK J. FABOZZI, PhD, CFA, CPA

Professor of Finance, EDHEC Business School

Abstract: Portfolio managers and traders need to be able to effectively model the impact of trading costs on their portfolios and trades.

Trading is an integral component of the equity investment process. A poorly executed trade can eat directly into portfolio returns. This is because equity markets are not frictionless, and transactions have a cost associated with them. Costs are incurred when buying or selling stocks in the form of, for example, brokerage commissions, bid-ask spreads, taxes, and market impact costs.

In recent years, portfolio managers have started to more carefully consider transaction costs. The literature on market microstructure, analysis and measurement of transaction costs, and market impact costs on institutional trades is rapidly expanding.¹ One way of describing transaction costs is to categorize them in terms of *explicit costs* such as brokerage and taxes, and *implicit costs*, which include market impact costs, price movement risk, and opportunity cost. *Market impact cost* is, broadly speaking, the price an investor has to pay for obtaining liquidity in the market, whereas price movement risk is the risk that the price of an asset increases or decreases from the time the investor

decides to transact in the asset until the transaction actually takes place. *Opportunity cost* is the cost suffered when a trade is not executed. Another way of seeing transaction costs is in terms of fixed costs versus variable costs. Whereas commissions and trading fees are fixed, bid-ask spreads, taxes, and all implicit transaction costs are variable.

Portfolio managers and traders need to be able to effectively model the impact of trading costs on their portfolios and trades. In this entry, we introduce several approaches for the modeling of transaction costs, in particular market impact costs.

MARKET IMPACT COSTS

The *market impact cost* of a transaction is the deviation of the transaction price from the market (mid) price² that would have prevailed had the trade not occurred. The price movement is the cost, the market impact cost, for liquidity. Market impact of a trade can be negative if, for example, a trader buys at a price below the

no-trade price (i.e., the price that would have prevailed had the trade not taken place). In general, liquidity providers experience negative costs while liquidity demanders will face positive costs.

We distinguish between two different kinds of market impact costs, temporary and permanent. Total market impact cost is computed as the sum of the two. The temporary market impact cost is of transitory nature and can be seen as the additional liquidity concession necessary for the liquidity provider (e.g., the market maker) to take the order, inventory effects (price effects due to broker/dealer inventory imbalances), or imperfect substitution (for example, price incentives to induce market participants to absorb the additional shares).

The permanent market impact cost, however, reflects the persistent price change that results as the market adjusts to the information content of the trade. Intuitively, a sell transaction reveals to the market that the security may be overvalued, whereas a buy transaction signals that the security may be undervalued. Security prices change when market participants adjust their views and perceptions as they observe news and the information contained in new trades during the trading day.

Traders can decrease the temporary market impact by extending the trading horizon of an order. For example, a trader executing a less urgent order can buy or sell his or her position in smaller portions over a period and make sure that each portion only constitutes a small percentage of the average volume. However, this comes at the price of increased opportunity costs, delay costs, and price movement risk.

Market impact costs are often asymmetric; that is, they are different for buy and sell orders. Several empirical studies suggest that market impact costs are generally higher for buy orders. Nevertheless, while buying costs might be higher than selling costs, this empirical fact is most likely due to observations during rising/falling markets, rather than any true market microstructure effects. For example, a study by

Hu shows that the difference in market impact costs between buys and sells is an artifact of the trade benchmark.³ (We discuss trade benchmarks later in this entry.) When a pre-trade measure is used, buys (sells) have higher implicit trading costs during rising (falling) markets. Conversely, if a post-trade measure is used, sells (buys) have higher implicit trading costs during rising (falling) markets. In fact, both pre-trade and post-trade measures are highly influenced by market movement, whereas during- or average-trade measures are neutral to market movement.

Despite the enormous global size of equity markets, the impact of trading is important even for relatively small funds. In fact, a sizable fraction of the stocks that compose an index might have to be excluded or their trading severely limited. For example, RAS Asset Management, which is the asset manager arm of the large Italian insurance company RAS, has determined that single trades exceeding 10% of the daily trading volume of a stock cause an excessive market impact and have to be excluded, while trades between 5% and 10% need execution strategies distributed over several days.⁴ According to RAS Asset Management estimates, in practice funds managed actively with quantitative techniques and with market capitalization in excess of €100 million can operate only on the fraction of the market above the €5 million, splitting trades over several days for stocks with average daily trading volume in the range from €5 million to €10 million. They can freely operate only on two-thirds of the stocks in the MSCI Europe.

LIQUIDITY AND TRANSACTION COSTS

Liquidity is created by agents transacting in the financial markets when they buy and sell securities. Market makers and brokers–dealers do not create liquidity; they are intermediaries who facilitate trade execution and maintain an orderly market.

Liquidity and transaction costs are interrelated. A highly liquid market is one where large transactions can be immediately executed without incurring high transaction costs. In an indefinitely liquid market, traders would be able to perform very large transactions directly at the quoted bid-ask prices. In reality, particularly for larger orders, the market requires traders to pay more than the ask when buying and to receive less than the bid when selling. As we discussed previously, this percentage degradation of the bid-ask prices experienced when executing trades is the market impact cost.

The market impact cost varies with transaction size: The larger the trade size, the larger the impact cost. Impact costs are not constant in time, but vary throughout the day as traders change the limit orders that they have in the limit order book. A *limit order* is a conditional order; it is executed only if the limit price or a better price can be obtained. For example, a buy limit order of a security XYZ at \$60 indicates that the assets may be purchased only at \$60 or lower. Therefore, a limit order is very different from a *market order*, which is an unconditional order to execute at the current best price available in the market (guarantees execution, not price). With a limit order, a trader can improve the execution price relative to the market order price, but the execution is neither certain nor immediate (guarantees price, not execution).

Notably, there are many different limit order types available such as pegging orders, discretionary limit orders, immediate or cancel order (IOC) orders, and fleeting orders. For example, fleeting orders are those limit orders that are canceled within two seconds of submission. Hasbrouck and Saar find that fleeting limit orders are much closer substitutes for market orders than for traditional limit orders.⁵ This suggests that the role of limit orders has changed from the traditional view of being liquidity suppliers to being substitutes for market orders.

At any given instant, the list of orders sitting in the limit order book embodies the liquidity that exists at a particular point in time. By observing the entire limit order book, impact costs can be calculated for different transaction sizes. The limit order book reveals the prevailing supply and demand in the market.⁶ Therefore, in a pure limit order market, we can obtain a measure of liquidity by aggregating limit buy orders (representing the demand) and limit sell orders (representing the supply).⁷

We start by sorting the bid and ask prices, $p_1^{\text{bid}}, \dots, p_k^{\text{bid}}$ and $p_1^{\text{ask}}, \dots, p_l^{\text{ask}}$, (from the most to the least competitive) and the corresponding order quantities $q_1^{\text{bid}}, \dots, q_k^{\text{bid}}$ and $q_1^{\text{ask}}, \dots, q_l^{\text{ask}}$. We then combine the sorted bid and ask prices into a supply and demand schedule according to Figure 1. For example, the block $(p_2^{\text{bid}}, q_2^{\text{bid}})$ represents the second best sell limit order with price p_2^{bid} and quantity q_2^{bid} .

We note that unless there is a gap between the bid (demand) and the ask (supply) sides, there will be a match between a seller and buyer, and a trade would occur. The larger the gap, the lower the liquidity and the market participants' desire to trade. For a trade of size Q , we can define its liquidity as the reciprocal of the area between the supply and demand curves up to Q (i.e., the "dotted" area in Figure 1).

However, few order books are publicly available and not all markets are pure limit order markets. In 2004, the New York Stock Exchange (NYSE) started selling information on its limit order book through its new system called the *NYSE OpenBook*[®]. The system provides an aggregated real-time view of the exchange's limit-order book for all NYSE-traded securities.⁸

In the absence of a fully transparent limit order book, expected market impact cost is the most practical and realistic measure of market liquidity. It is closer to the true cost of transacting faced by market participants as compared to other measures such as those based upon the bid-ask spread.

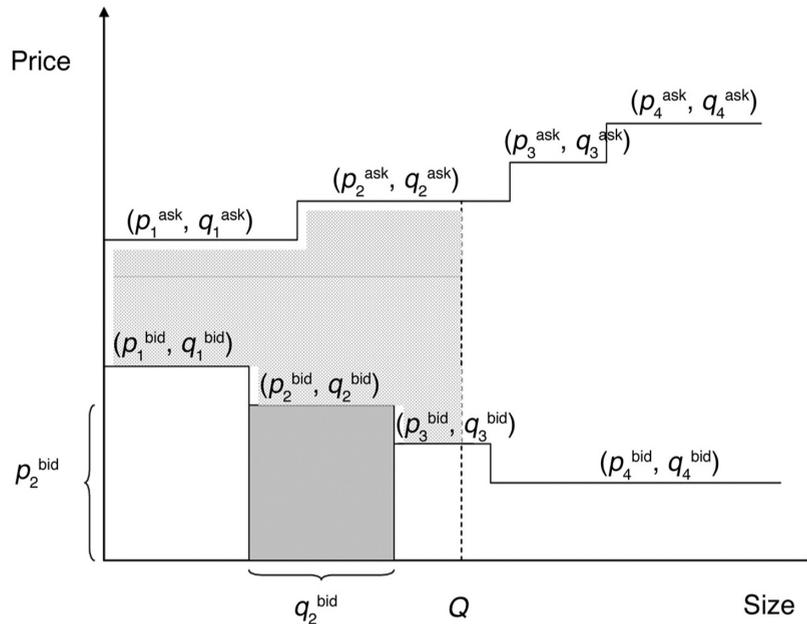


Figure 1 The Supply and Demand Schedule of a Security
 Source: Figure 1A in Domowitz and Wang (2002, p. 38).

MARKET IMPACT MEASUREMENTS AND EMPIRICAL FINDINGS

The problem with measuring implicit transaction costs is that the true measure, which is the difference between the price of the stock in the absence of a money manager's trade and the execution price, is not observable. Furthermore, the execution price is dependent on supply and demand conditions at the margin. Thus, the execution price may be influenced by competitive traders who demand immediate execution or by other investors with similar motives for trading. This means that the execution price realized by an investor is the consequence of the structure of the market mechanism, the demand for liquidity by the marginal investor, and the competitive forces of investors with similar motivations for trading.

There are many ways to measure transaction costs. However, in general this cost is the difference between the execution price and some

appropriate benchmark, a so-called *fair market benchmark*. The fair market benchmark of a security is the price that would have prevailed had the trade not taken place, the no-trade price. Since the no-trade price is not observable, it has to be estimated. Practitioners have identified three different basic approaches to measure the market impact:⁹

1. Pre-trade measures use prices occurring before or at the decision to trade as the benchmark, such as the opening price on the same day or the closing price on the previous day.
2. Post-trade measures use prices occurring after the decision to trade as the benchmark, such as the closing price of the trading day or the opening price on the next day.
3. Same-day or average measures use average prices of a large number of trades during the day of the decision to trade, such as the *volume-weighted average price* (VWAP) calculated over all transactions in the security on the trade day.¹⁰

The volume-weighted average price is calculated as follows. Suppose that it was a trader's objective to purchase 10,000 shares of stock XYZ. After completion of the trade, the trade sheet showed that 4,000 shares were purchased at \$80, another 4,000 at \$81, and finally 2,000 at \$82. In this case, the resulting VWAP is $(4,000 \times 80 + 4,000 \times 81 + 2,000 \times 82)/10,000 = \80.80 .

We denote by χ the indicator function that takes on the value 1 or -1 if an order is a buy or sell order, respectively. Formally, we now express the three types of measures of market impact (MI) as follows

$$MI_{pre} = \left(\frac{p^{ex}}{p^{pre}} - 1 \right) \chi$$

$$MI_{post} = \left(\frac{p^{ex}}{p^{post}} - 1 \right) \chi$$

$$MI_{VWAP} = \left(\frac{\sum_{i=1}^k V_i \cdot p_i^{ex}}{\sum_{i=1}^k V_i} / p^{pre} - 1 \right) \chi$$

where p^{ex} , p^{pre} , and p^{post} denote the execution price, pre-trade price, and post-trade price of the stock, and k denotes the number of transactions in a particular security on the trade date. Using this definition, for a stock with market impact MI the resulting *market impact cost* for a trade of size V , MIC , is given by

$$MIC = MI \cdot V$$

It is also common to adjust market impact for general market movements. For example, the pre-trade market impact with market adjustment would take the form

$$MI_{pre} = \left(\frac{p^{ex}}{p^{pre}} - \frac{p_M^{ex}}{p_M^{pre}} \right) \chi$$

where p_M^{ex} represent the value of the index at the time of the execution, and p_M^{pre} the price of the index at the time before the trade. Market-adjusted market impact for the post-trade and same-day trade benchmarks are calculated in an analogous fashion.

The above three approaches to measure market impact are based upon measuring the fair market benchmark of stock at a point in time. Clearly, different definitions of market impact lead to different results. Which one should be used is a matter of preference and is dependent on the application at hand. For example, Elkins and McSherry, a financial consulting firm that provides customized trading costs and execution analysis, calculates a same-day benchmark price for each stock by taking the mean of the day's open, close, high, and low prices. The market impact is then computed as the percentage difference between the transaction price and this benchmark. However, in most cases VWAP and the Elkins McSherry approach lead to similar measurements.¹¹

As we analyze a portfolio's return over time an important question to ask is whether we can attribute good/bad performance to investment profits/losses or to trading profits/losses. In other words, in order to better understand a portfolio's performance it can be useful to decompose investment decisions from order execution. This is the basic idea behind the *implementation shortfall approach* suggested by Perold (1998).

In the implementation shortfall approach, we assume that there is a separation between investment and trading decisions. The portfolio manager makes decisions with respect to the investment strategy (i.e., what should be bought, sold, and held). Subsequently, these decisions are implemented by the traders.

By comparing the actual portfolio profit/loss (P/L) with the performance of a hypothetical paper portfolio in which all trades are made at hypothetical market prices, we can get an estimate of the implementation shortfall. For example, with a paper portfolio return of 6% and an actual portfolio return of 5%, the implementation shortfall is 1%.

There is considerable practical and academic interest in the measurement and analysis of international trading costs. Domowitz, Glen, and Madhavan (1999) examine international equity

trading costs across a broad sample of 42 countries using quarterly data from 1995 to 1998. They find that the mean total one-way trading cost is 69.81 basis points. However, there is an enormous variation in trading costs across countries. For example, in their study the highest was Korea with 196.85 basis points whereas the lowest was France with 29.85 basis points. Explicit costs are roughly two-thirds of total costs. However, one exception to this is the United States where the implicit costs are about 60% of the total costs.

Transaction costs in emerging markets are significantly higher than those in more developed markets. Domowitz, Glen, and Madhavan argue that this fact limits the gains of international diversification in these countries, explaining in part the documented home bias of domestic investors.

In general, they find that transaction costs declined from the middle of 1997 to the end of 1998, with the exception of Eastern Europe. It is interesting to notice that this reduction in transaction costs happened despite the turmoil in the financial markets during this period. A few explanations that Domowitz et al. suggest are that (1) the increased institutional presence has resulted in a more competitive environment for brokers/dealers and other trading services; (2) technological innovation has led to a growth in the use of low-cost electronic crossing networks (ECNs) by institutional traders; and (3) soft dollar payments are now more common.

FORECASTING AND MODELING MARKET IMPACT

In this section, we describe a general methodology for constructing forecasting models for market impact. These types of models are very useful in predicting the resulting trading costs of specific trading strategies and in devising optimal trading approaches.

Explicit transaction costs are relatively straightforward to estimate and forecast. There-

fore, our focus in this section is to develop a methodology for the *implicit* transaction costs, and more specifically, market impact costs. The methodology is a linear factor-based approach where market impact is the dependent variable. We distinguish between trade-based and asset-based independent variables or *forecasting factors*.

Trade-Based Factors

Some examples of trade-based factors include:

- Trade size
- Relative trade size
- Price of market liquidity
- Type of trade (information or informationless trade)
- Efficiency and trading style of the investor
- Specific characteristics of the market or the exchange
- Time of trade submission and trade timing
- Order type

Probably the most important market impact forecasting variables are based on absolute or relative trade size. Absolute trade size is often measured in terms of the number of shares traded, or the dollar value of the trade. Relative trade size, on the other hand, can be calculated as number of shares traded divided by average daily volume, or number of shares traded divided by the total number of shares outstanding. Note that the former can be seen as an explanatory variable for the temporary market impact and the latter for the permanent market impact. In particular, we expect the temporary market impact to increase as the trade size to the average daily volume increases because a larger trade demands more liquidity.

Each type of investment style requires a different need for immediacy.¹² Technical trades often have to be traded at a faster pace in order to capitalize on some short-term signal and therefore exhibit higher market impact costs. In contrast, more traditional long-term value strategies can be traded more slowly. These types of strategies

can in many cases even be liquidity providing, which might result in negative market impact costs.

Several studies show that there is a wide variation in equity transaction costs across different countries.¹³ Markets and exchanges in each country are different, and so are the resulting market microstructures. Forecasting variables can be used to capture specific market characteristics such as liquidity, efficiency, and institutional features.

The particular timing of a trade can affect the market impact costs. For example, it appears that market impact costs are generally higher at the beginning of the month as compared to the end of it.¹⁴ One of the reasons for this phenomenon is that many institutional investors tend to rebalance their portfolios at the beginning of the month. Because it is likely that many of these trades will be executed in the same stocks, this rebalancing pattern will induce an increase in market impact costs. The particular time of the day a trade takes place does also have an effect. Many informed institutional traders tend to trade at the market open as they want to capitalize on new information that appeared after the market close the day before.

As we discussed earlier in this entry, market impact costs are asymmetric. In other words, buy and sell orders have significantly different market impact costs. Separate models for buy and sell orders can therefore be estimated. However, it is now more common to construct a model that includes dummy variables for different types of orders such as buy/sell orders, market orders, limit orders, and the like.

Asset-Based Factors

Some examples of asset-based factors are:

- Price momentum
- Price volatility
- Market capitalization
- Growth versus value
- Specific industry or sector characteristics

For a stock that is exhibiting positive price momentum, a buy order is liquidity demanding and it is, therefore, likely that it will have higher market impact cost than a sell order.

Generally, trades in high volatility stocks result in higher permanent price effects. It has been suggested by Chan and Lakonishok (1997) and Smith et al. (2001) that this is because trades have a tendency to contain more information when volatility is high. Another possibility is that higher volatility increases the probability of hitting and being able to execute at the liquidity providers' price. Consequently, liquidity suppliers display fewer shares at the best prices to mitigate adverse selection costs.

Large-cap stocks are more actively traded and therefore more liquid in comparison to small-cap stocks. As a result, market impact cost is normally lower for large caps.¹⁵ However, if we measure market impact costs with respect to relative trade size (normalized by average daily volume, for instance), they are generally higher. Similarly, growth and value stocks have different market impact cost. One reason for that is related to the trading style. Growth stocks commonly exhibit momentum and high volatility. This attracts technical traders that are interested in capitalizing on short-term price swings. Value stocks are traded at a slower pace and holding periods tend to be slightly longer.

Different market sectors show different trading behaviors. For instance, Bikker and Spierdijk (2007) show that equity trades in the energy sector exhibit higher market impact costs than other comparable equities in nonenergy sectors.

A Factor-Based Market Impact Model

One of the most common approaches in practice and in the literature in modeling market impact is through a linear factor model of the form:

$$MI_t = \alpha + \sum_{i=1}^I \beta_i x_i + \varepsilon_t$$

where α, β_i are the factor loadings and x_i are the factors. Frequently, the error term ε_t is assumed to be independently and identically distributed. Recall that the resulting market impact cost of a trade of (dollar) size V is then given by $MIC_t = MI_t \cdot V$. However, extensions of this model including conditional volatility specifications are also possible. By analyzing both the mean and the volatility of the market impact, we can better understand and manage the trade-off between the two. For example, Bikker and Spierdijk use a specification where the error terms are jointly and serially uncorrelated with mean zero, satisfying

$$\text{Var}(\varepsilon_t) = \exp\left(\gamma + \sum_{i=1}^J \delta_j z_j\right)$$

where γ, δ_j , and z_j are the volatility, factor loadings, and factors, respectively.

Although the market impact function is linear, this of course does not mean that the dependent variables have to be. In particular, the factors in the previous specification can be nonlinear transformations of the descriptive variables.

Consider, for example, factors related to trade size (e.g., trade size and trade size to daily volume). It is well known that market impact is nonlinear in these trade size measures. One of the earliest studies in this regard was performed by Loeb (1983), who showed that for a large set of stocks the market impact is proportional to the square root of the trade size, resulting in a market impact cost proportional to $V^{3/2}$. Typically, a market impact function linear in trade size will underestimate the price impact of small- to medium-sized trades whereas larger trades will be overestimated.

Chen, Stanzl, and Watanabe (2002) suggest to model the nonlinear effects of trade size (dollar trade size V) in a market impact model by using the Box-Cox transformation; that is,

$$MI(V_t) = \alpha_b + \beta_b \frac{V_t^{\lambda_b} - 1}{\lambda_b} + \varepsilon_t$$

where t and τ represent the time of transaction for the buys and the sells, respectively. In their

specification, they assumed that ε_t and ε_τ are independent and identically distributed with mean zero and variance σ^2 . The parameters $\alpha_b, \beta_b, \lambda_b, \alpha_s, \beta_s$, and λ_s were then estimated from market data by nonlinear least squares for each individual stock. We remark that $\lambda_b, \lambda_s \in [0, 1]$ in order for the market impact for buys to be concave and for sells to be convex.

In their data sample (NYSE and Nasdaq trades between January 1993 and June 1993), Chen, Stanzl, and Watanabe report that for small companies the curvature parameters λ_b, λ_s are close to zero, whereas for larger companies they are not far away from 0.5. Observe that for $\lambda_b = \lambda_s = 1$ market impact is linear in the dollar trade size. Moreover, when $\lambda_b = \lambda_s = 0$ the impact function is logarithmic by the virtue of

$$\lim_{\lambda \rightarrow 0} \frac{V^\lambda - 1}{\lambda} = \ln(V)$$

As just mentioned, market impact is also a function of the characteristics of the particular exchange where the securities are traded as well as of the trading style of the investor. These characteristics can also be included in the general specification outlined previously. For example, Keim and Madhavan (1996, 1997) proposed the following two different market impact specifications

$$1. MI = \alpha + \beta_1 \chi_{OTC} + \beta_2 \frac{1}{p} + \beta_3 |q| + \beta_4 |q|^2 + \beta_5 |q|^3 + \beta_6 \chi_{Up} + \varepsilon$$

where

χ_{OTC} = a dummy variable equal to one if the stock is an OTC traded stock or zero otherwise.

p = the trade price.

q = the number of shares traded over the number of shares outstanding.

χ_{Up} = a dummy variable equal to one if the trade is done in the upstairs¹⁶ market or zero otherwise.

$$2. MI = \alpha + \beta_1 \chi_{\text{Nasdaq}} + \beta_2 q + \beta_3 \ln(MCap) + \beta_4 \frac{1}{p} + \beta_5 \chi_{\text{Tech}} + \beta_6 \chi_{\text{Index}} + \varepsilon$$

where

χ_{Nasdaq} = a dummy variable equal to one if the stock is traded on Nasdaq or zero otherwise.

q = the number of shares traded over the number of shares outstanding.

$MCap$ = the market capitalization of the stock.

p = the trade price.

χ_{Tech} = a dummy variable equal to one if the trade is a short-term technical trade or zero otherwise.

χ_{Index} = a dummy variable equal to one if the trade is done for a portfolio that attempts to closely mimic the behavior of the underlying index or zero otherwise.

These two models provide good examples for how nonlinear transformations of the underlying dependent variables can be used along with dummy variables that describe specific market or trade characteristics.

Several vendors and broker-dealers such as MSCI Barra¹⁷ and ITG¹⁸ have developed commercially available market impact models. These are sophisticated multimarket models that rely upon specialized estimation techniques using intraday data or tick-by-tick transaction-based data. However, the general characteristics of these models are similar to the ones described in this section.

We emphasize that in the modeling of transaction costs it is important to factor in the objective of the trader or investor. For example, one market participant might trade just to take advantage of price movement and hence will only trade during favorable periods. This investor's trading cost is different from that of an investor who has to rebalance a portfolio within a fixed time period and can therefore only partially use an opportunistic or liquidity searching strategy. In particular, this investor

has to take into account the risk of not completing the transaction within a specified time period. Consequently, even if the market is not favorable, this investor may decide to transact a portion of the trade. The market impact models described previously assume that orders will be fully completed and ignore this point.

KEY POINTS

- Trading and execution are integral components of the investment process. A poorly executed trade can eat directly into portfolio returns because of transaction costs.
- Transaction costs are typically categorized in two dimensions: fixed costs versus variable costs, and *explicit costs* versus *implicit costs*.
- In the first dimension, fixed costs include commissions and fees. Bid-ask spreads, taxes, delay cost, price movement risk, market impact costs, timing risk, and opportunity cost are variable trading costs.
- In the second dimension, explicit costs include commissions, fees, bid-ask spreads, and taxes. Delay cost, price movement risk, market impact cost, timing risk, and opportunity cost are implicit transaction costs.
- Implicit costs make up the larger part of the total transaction costs. These costs are not observable and have to be estimated.
- Liquidity is created by agents transacting in the financial markets by buying and selling securities.
- Liquidity and transaction costs are interrelated: In a highly liquid market, large transactions can be executed immediately without incurring high transaction costs.
- A limit order is an order to execute a trade only if the limit price or a better price can be obtained.
- A market order is an order to execute a trade at the current best price available in the market.
- In general, trading costs are measured as the difference between the execution price and

some appropriate fair market benchmark. The fair market benchmark of a security is the price that would have prevailed had the trade not taken place.

- Typical forecasting models for market impact costs are based on a statistical factor approach where the independent variables are trade-based factors or asset-based factors.

NOTES

1. See, for example, Domowitz, Glen, and Madhavan (2001) and Keim and Madhavan (1998).
2. Since the buyer buys at the ask and the seller sells at the bid, this definition of market impact cost ignores the bid-ask spread, which is an explicit cost.
3. Hu (2009).
4. Private communication, RAS Asset Management.
5. Hasbrouck and Saar (2008).
6. Note that even if it is possible to view the entire limit order book it does not give a *complete* picture of the liquidity in the market. This is because hidden and discretionary orders are not included. For a discussion on this topic, see Tuttle (2002).
7. Domowitz and Wang (2002) and Foucault, Kadan, and Kandel (2005).
8. NYSE and Securities Industry Automation Corporation, *NYSE OpenBook*[®], Version 1.1 (New York: 2004).
9. Collins and Fabozzi (1991) and Chan and Lakonishok (1993).
10. Strictly speaking, VWAP is not the benchmark here but rather the transaction type.
11. See Willoughby (1998) and McSherry (1998).
12. Keim and Madhavan (1997).
13. See Domowitz, Glen, and Madhavan (2001) and Chiyachantana, Jain, Jiang, and Wood (2004).
14. Foster and Viswanathan (1990).
15. Keim and Madhavan (1998) and Spierdijk, Nijman, and van Soest (2003).
16. A securities transaction not executed on the exchange but completed directly by a broker in-house is referred to as an upstairs market transaction. Typically, the upstairs market consists of a network of trading desks of the major brokerages and institutional investors. The major purpose of the upstairs market is to facilitate large block and program trades.
17. Torre and Ferrari (1999).
18. Investment Technology Group (2003).

REFERENCES

- Bikker, J. A., Spierdijk, L., and van der Sluis, P. L. (2007). Market impact costs of institutional equity trades. *Journal of International Money and Finance* 26: 974–1000.
- Chan, L. K. C., and Lakonishok, J. (1993). Institutional trades and intraday stock price behavior. *Journal of Financial Economics* 33: 173–199.
- Chan, L. K. C., and Lakonishok, J. (1997). Institutional equity trading costs: NYSE versus Nasdaq. *Journal of Finance* 52: 713–735.
- Chen, Z., Stanzl, W., and Watanabe, M. (2002). Price impact costs and the limit of arbitrage. Working paper, Yale School of Management, International Center for Finance.
- Chiyachantana, C. N., Jain, P. K., Jiang, C., and Wood, R. A. (2004). International evidence on institutional trading behavior and price impact. *Journal of Finance* 59: 869–895.
- Collins, B., and Fabozzi, F. J. (1991). A methodology for measuring transaction costs. *Financial Analysts Journal* 47: 27–36.
- Domowitz, I., Glen, J., and Madhavan, A. (1999). International equity trading costs: A cross-sectional and time-series analysis. Technical Report, Pennsylvania State University.
- Domowitz, I., Glen, J., and Madhavan, A. (2001). Liquidity, volatility, and equity trading costs across countries and over time. *International Finance* 4: 221–255.
- Domowitz, I., and Wang, X. (2002). Liquidity, liquidity commonality and its impact on portfolio theory. Working paper, Smeal College of Business Administration, Pennsylvania State University.
- Foster, F. D., and Viswanathan, S. (1990). A theory of the interday variations in volume, variance, and trading costs in securities markets. *Review of Financial Studies* 3: 593–624.

- Foucault, T., Kadan, O., and Kandel, E. (2005). Limit order book as a market for liquidity. *Review of Financial Studies* 18: 1171–1217.
- Hasbrouck, J., and Saar, G. (2008). Technology and liquidity provision: The blurring of traditional definitions. *Journal of Financial Markets* 12: 143–172.
- Hu, G. (2009). Measures of implicit trading costs and buy-sell asymmetry. *Journal of Financial Markets* 12: 418–437.
- Investment Technology Group, Inc. (2003). ITG ACE—Agency cost estimator: A model description. www.itginc.com.
- Keim, D. B., and Madhavan, A. (1996). The upstairs market for large-block transactions: Analysis and measurement of price effects. *Review of Financial Studies* 9: 1–36.
- Keim, D. B., and Madhavan, A. (1997). Transactions costs and investment style: An inter-exchange analysis of institutional equity trades. *Journal of Financial Economics* 46: 265–292.
- Keim, D. B., and Madhavan, A. (1998). The costs of institutional equity trades. *Financial Analysts Journal* 54: 50–69.
- Loeb, T. F. (1983). Trading costs: The critical link between investment information and results. *Financial Analysts Journal* 39: 39–44.
- McSherry, R. (1998). Global trading cost analysis. Mimeo, Elkins McSherry Co., Inc.
- Perold, A. F. (1998). The implementation shortfall: Paper versus reality. *Journal of Portfolio Management* 14: 4–9.
- Smith, B. F., Alasdair, D., Turnbull, S., and White, R. W. (2001). Upstairs market for principal and agency trades: Analysis of adverse information and price effects. *Journal of Finance* 56: 1723–1746.
- Spierdijk, L., Nijman, T., and van Soest, A. (2003). Temporary and persistent price effects of trades in infrequently traded stocks. Working paper, Tilburg University and Center.
- Torre, N. G., and Ferrari, M. J. (1999). The market impact model. Barra Research Insights.
- Tuttle, L. A. (2002). Hidden orders, trading costs and information. Working paper, Ohio State University.
- Willoughby, J. (1998). Executions song. *Institutional Investor* 32: 51–56.

Volatility

Monte Carlo Simulation in Finance

DESSISLAVA A. PACHAMANOVA, PhD

Associate Professor of Operations Research, Babson College

Abstract: Monte Carlo simulation has become an essential tool for pricing and risk estimation in financial applications. It allows finance professionals to incorporate uncertainty in financial models, and to consider additional layers of complexity that are difficult to incorporate in analytical models. The main idea of Monte Carlo simulation is to represent the uncertainty in market variables through scenarios, and to evaluate parameters of interest that depend on these market variables in complex ways. The advantage of such an approach is that it can easily capture the dynamics of underlying processes and the otherwise complex effects of interactions among market variables. A substantial amount of research in recent years has been dedicated to making scenario generation more accurate and efficient, and a number of sophisticated computational techniques are now available to the financial modeler.

This entry provides an introduction to Monte Carlo simulation and its applications to finance, from financial derivative *pricing* to portfolio *risk management*. We begin with a discussion of the main ideas behind *simulation* and a listing of several important areas in finance where simulation techniques are widely used. We then discuss technical issues that are important for understanding the advantages and limitations of the Monte Carlo simulation technique, such as how random numbers are actually generated, what techniques are used for increasing the accuracy of estimates from simulation, and what software can be helpful for applications.

MAIN IDEAS AND IMPORTANT CONCEPTS

Simulation can be most generally defined as imitation of real-life systems with the goal of

studying important characteristics of their behavior. Monte Carlo simulation is named after the main residential area of the Monaco principality, which was well known for its casino. The term alludes to randomness and process repetition, analogous to casino games such as roulette.

The idea of applying Monte Carlo simulation to finance arises naturally, given the inherent variability in markets and the need for finance professionals to evaluate strategies with uncertain outcomes. Consider, for example, a portfolio manager who would like to estimate the effect of a market downturn on the portfolio (e.g., if the market goes down by 10%). What would be the resulting portfolio value? If the portfolio beta is 1, the expected decline in the portfolio value will be 10% as well; if the portfolio beta is 0.9, the portfolio will decline 9% if the market declines by 10%. More generally,

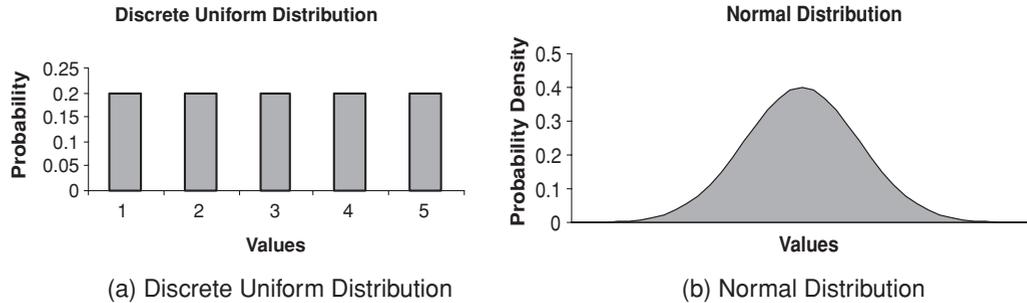


Figure 1 Examples of Probability Distributions

a portfolio manager may want to assess the exposure of a portfolio to a set of risk factors suggested by economic theory or empirical evidence such as interest rate changes, commodity price changes, exchange rate movements, and so on. These risk factors and their interactions with each other are not straightforward to evaluate. One can imagine that a portfolio manager would consider scenarios for possible joint realizations of market variables—for example, in a global recession or under favorable monetary policy changes—and would assess the change to the portfolio value in each of these scenarios. Taking it yet another step further, a portfolio manager may assign probabilities to the different scenarios, thus expressing a view on their likelihood of occurring. Assigning probabilities to outcomes produces probability distributions. Examples of probability distributions include the discrete uniform distribution (see Figure 1a), which assigns equal probabilities to all possible discrete outcomes, and the normal distribution (Figure 1b), which is continuous (defined on a range, as opposed to discrete values), and allocates more probability to outcomes close to the average than to those far from the average.

The example in the previous paragraph illustrates a Monte Carlo simulation system: Possibly random inputs (the risk factors) incorporating subjective or statistically estimated views via probability distributions are entered into an evaluation model (computation of change in portfolio value), and the resulting output (the portfolio change) is not a single

number, but a probability distribution of outcomes that incorporates characteristics of the input probability distributions and their complex interactions. The actual simulation process involves generating a certain number of scenarios, evaluating the portfolio change for each scenario, and obtaining a corresponding set of scenarios for the portfolio change. The latter set of scenarios can then be analyzed to determine most likely outcomes for portfolio change, variability of estimated portfolio change, range of possible outcomes, and the like. One can use the simulation output also to estimate any portfolio risk measure such as value-at-risk (VaR) or expected tail loss (ETL). Since VaR has been adopted by regulators and is commonly used by portfolio managers, we will use VaR in our illustrations. When generating scenarios for the factors influencing the future value of the portfolio, it is easy to collect information on possible portfolio losses relative to the current value of the portfolio in each scenario. Then, the 95% VaR, for example, can be computed as the 95th percentile of the distribution of portfolio losses (see Figure 2).

As another illustration of a simulation model, consider the problem of finding the fair price of a simple European call option on a stock with current stock price S_t . If the strike price is K and the option matures at time T , the option payoff at time T can be expressed as

$$V_T = \max \{S_T - K, 0\}$$

According to a fundamental theory in asset pricing, the fair price of a financial asset

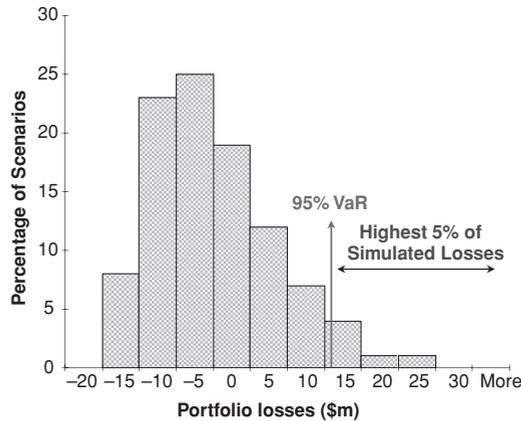


Figure 2 Determining Portfolio VaR from Simulation

under certain conditions can be meaningfully estimated as the expected value (equivalently, as a “probability-weighted average”) of the possible payoffs of the financial asset in different states of the world in the future. The fair value of the option at time t will therefore be the expected value of the discounted payoff:

$$V_t = E \left[e^{-r(T-t)} \max \{ S_T - K, 0 \} \right]$$

where r is the short-term risk-free rate.

The expected value in the expression above is meaningful only if one can specify a probability distribution of possible outcomes for the future price of the asset. For example, consider a European call option on a common stock with an exercise price of \$20. Assume that the short-term risk-free rate is 0%, and suppose that the stock price at time T can only take the values \$18, \$21, and \$23 with (risk-neutral) probabilities $3/6$, $2/6$, and $1/6$, respectively. Then the fair price of the option can be computed as the weighted average of the payoffs in the three possible states of the world:

$$\begin{aligned} V &= \frac{3}{6} \max \{ 18 - 20, 0 \} + \frac{2}{6} \max \{ 21 - 20, 0 \} \\ &\quad + \frac{1}{6} \max \{ 23 - 20, 0 \} \\ &= \frac{3}{6} \cdot 0 + \frac{2}{6} \cdot 1 + \frac{1}{6} \cdot 3 = \frac{5}{6} = 0.83 \end{aligned}$$

That is, the fair value of the option is \$0.83.

Typically, however, the stock price can take many more values, and the option price cannot be valued exactly. It therefore makes sense to generate a large number (e.g., 1,000) of scenarios for the future value of the stock price using the risk-neutral probabilities, and average out the payoffs to the option. The average obtained from the simulation will approximate the true expected value of the option.

The Black-Scholes formula for European options (Black and Scholes, 1973) is widely used in the financial industry. It provides a closed-form expression for computing the price of the option. The underlying assumption used in the derivation of the Black-Scholes formula is that the percentage changes in the asset price are increments of a Brownian motion.¹ The evolution of the stock price can then be described by the equation

$$dS_t = \mu S_t dt + \sigma S_t dW_t \tag{1}$$

where W_t is standard Brownian motion, and μ and σ are called “drift” and “volatility” of the process, respectively.

Equation (1) says that the change in the asset price at any time period is determined by two components: (1) a drift term that is a fraction of the current asset price level, and (2) a “random noise” term that assumes that volatility is proportional to the current price level. For technical reasons (namely, absence of arbitrage), when pricing an option on an asset whose movement is described by equation (1), the drift μ is replaced by the risk-free rate r . The technical details of equation (1) are not important for our purposes. The important result is that under the assumption for the random process followed by the stock price in (1), the value of the stock price S_T at time T can be computed as

$$S_T = S_t e^{(r - \frac{1}{2}\sigma^2)(T-t) + \sigma\sqrt{(T-t)}\tilde{w}} \tag{2}$$

where \tilde{w} is a random variable following a normal distribution with mean 0 and standard deviation 1 (see Figure 1b).

Hence, the option price obtained from the Black-Scholes formula can be approximated by simulation if a large number of values for the normal random variable \tilde{w} are generated, thus creating scenarios for the stock price S_T at time T and allowing for computing the discounted payoffs of the option. Suppose we generate n scenarios for \tilde{w} : w_1, \dots, w_n . Then, the price of the European option will be

$$V_t = e^{-r(T-t)} \cdot \sum_{i=1}^n \frac{1}{n} \max \left\{ S_t e^{(r - \frac{1}{2}\sigma^2)(T-t) + \sigma\sqrt{(T-t)}w_i} - K, 0 \right\}$$

Note that the expression above is still a weighted average of the payoffs of the option in each scenario: the “weight,” or the probability of each scenario, is assumed to be $1/n$, since the scenarios are picked at random, and the frequency of their occurrence already incorporates the probability distribution of \tilde{w} .

It appears unnecessarily complicated to price the option this way, and indeed, in practice simulation is rarely used for such simple problems. There are more complex derivative instruments and more sophisticated models for asset price behavior; in such cases, it may be simpler to generate scenarios and evaluate prices by simulation than to look for closed-form analytical expressions like the Black-Scholes formula. In addition, in the case of portfolios and baskets of multiple assets, generating joint scenarios for multiple securities in simulation can help capture the otherwise complicated effect of interactions among different risk factors influencing the future value of the portfolio or derivative instrument.

How Many Scenarios?

A simulation may not be able to capture all possible realizations of uncertainties in the model. For instance, consider the European option pricing example above. If the percentage change in the stock price is assumed to be the increment of a Brownian motion, the possible number of

values for the stock price S_T at time T is infinite. (This is because the number of values the normal random variable \tilde{w} can take is infinite—the normal distribution has an infinite range.) Thus, one could never obtain the exact value of the option price by simulation. One can, however, get close. The accuracy of the estimation will depend on the number of generated scenarios. If the scenario generation is truly random, then the standard error (the “variability”) in the estimate of the average will be

$$\frac{s}{\sqrt{n}}$$

where s is the standard deviation of the simulated discounted option payoffs, and n is the number of scenarios. This result follows from the central limit theorem (CLT). This theorem states that if a sample of n independent and identically distributed observations is drawn from a distribution with mean μ and standard deviation σ , then the sample mean (which is an estimate of the true distribution mean μ) will follow a normal distribution around the actual distribution mean μ with standard deviation σ/\sqrt{n} as the sample size n tends to infinity, regardless of the shape of the original distribution, as long as n is large. The fact that the distribution standard deviation σ in the CLT can be replaced by the sample standard deviation s follows from additional theoretical results on the convergence of s to σ in distribution as the number of observations grows large.

Hence, to double the accuracy of estimating the mean of the output distribution, one would have to quadruple the number of scenarios. This can get expensive computationally, especially in more complicated multistage situations. Fortunately, there are modern methods for generating random numbers and scenarios that can help reduce the computational burden.

While the average output from a simulation is important, it is often not the only quantity of interest, something that practitioners tend to forget when using simulation to value complex financial instruments. For example, as

mentioned earlier, in assessing the risk of a portfolio, a portfolio manager may be interested in the percentiles of the distribution of outputs (VaR for portfolios) or the worst-case and best-case scenarios. Unfortunately, it is not as straightforward to determine the accuracy of those estimates from a simulation. There are some useful results from probability theory that apply.² However, the general question of how many scenarios one should generate to get a good representation of the output distribution does not have an easy answer. This issue is complicated further by the fact that results from probability theory do not necessarily apply to many of the scenario-generating methods used in practice, which do not simulate “truly random” samples of observations, but instead use smarter methods that reduce the number of scenarios needed to achieve good estimate accuracy. We will discuss some such methods later in this entry.

Estimator Bias

The statistical concept of estimator bias is important in simulation applications because it shows whether an estimator estimates the “right thing” on average (that is, whether it approaches the true parameter one needs to estimate given a sufficient number of replications). For example, the average obtained from a sample of scenarios is an unbiased estimator of the true expected value. Depending on the way scenarios are generated, however, one may introduce a bias in the estimate of the parameter of interest.

Suppose, for example, that one generates scenarios for the future asset price in the option pricing example introduced earlier in this entry, but instead of the formula describing the evolution of the asset price in continuous time (equation (2)), one divides the time between now and the maturity of the option into small intervals of length h and uses a “discrete-time” formula [based on equation (1)] to approximate the stock price at each time period between t and T , com-

piling the changes to obtain the final asset price at the maturity of the option.

Simulating the asset price in this manner will generate a bias in the estimate of the expected present value of the option, because the simulated changes in the asset price along the way are not continuous or instantaneous, but happen over a fixed-length time interval. This kind of bias is referred to as “discretization error bias.” Of course, in the case of geometric Brownian motion with fixed drift and volatility described by equation (1) one can obtain an unbiased estimator of the average option payoff by simulating the future asset price with the continuous-time formula (2). However, in many instances it is not possible to find such a closed-form expression for the future asset price; for example, such a formula does not exist when the volatility σ in the random process for the asset price is time-dependent, or when one uses a mean-reversion process to describe the evolution of the underlying price. In such cases, one can reduce the time interval length h to reduce the bias, but it is important to keep in mind that reducing the time interval length increases the number of steps necessary to create a scenario for the future asset price, and becomes computationally expensive.

Estimator Efficiency

If there are two ways to obtain an estimate of a quantity of interest and the estimators are otherwise equivalent in terms of bias, which estimator should be preferable; that is, which estimator is more “efficient”? Statistical theory states that one should prefer the estimator with the smaller standard deviation, because it is more accurate. For example, consider two unbiased estimators, both of which are obtained as averages from a sample of independent replications. Their standard errors will be given by

$$\frac{s_1}{\sqrt{n_1}} \quad \text{and} \quad \frac{s_2}{\sqrt{n_2}}$$

where s_1 and s_2 are the standard deviations from the samples of scenarios, and n_1 and n_2 are the number of scenarios for each of the estimators.

In the case of simulation, statistical concepts frequently need to be extended to include numerical and computational considerations. For example, suppose that it takes longer to generate the scenarios for the estimator with the smaller standard deviation. Is that estimator still preferable, given that one can use the extra time to generate additional scenarios for the other estimator, thus reducing the latter estimator's standard error? It is natural (and theoretically justified) to modify the measure of variability and efficiency so that it includes a concept of time. If τ_1 and τ_2 are the times it takes to generate one scenario for each of the two estimators, then one should select the estimator with the smaller of the time-adjusted standard deviations $s_1\sqrt{\tau_1}$, $s_2\sqrt{\tau_2}$.

FINANCIAL APPLICATIONS OF SIMULATION

Simulation has become an important staple in a financial modeler's toolbox. This section lists some important examples of simulation applications in finance.

Financial Derivative Pricing

The use of Monte Carlo simulation in derivative pricing dates back to Boyle (1977). Although the technique is not widely used for pricing of European-style securities with a single underlying stochastic variable, it is helpful for pricing European-style securities with multiple underlying stochastic variables, path-dependent options, such as Asian and American options, as well as basket options, where correlations between assets need to be taken into consideration. Additional examples of Monte Carlo simulation applications in financial derivative pricing include options on the spread between

two assets, barrier options, and quantos, whose payoff depends both on a stock price and an exchange rate. We already described a simple example of pricing a European call by simulation. In this section, we discuss further simulation issues in the context of pricing Asian options.

The value of an Asian option is determined by the average price of the underlying asset either continuously over the time to maturity or at a prespecified set of monitoring dates t_1, \dots, t_T . In particular, the payoff of an Asian call option is

$$V_T = \max \{ S_{\text{average}} - K, 0 \}$$

Thus, to price the option, one needs information not only on the value of the asset at time T , but also on the possible paths the asset could take to reach its terminal value. If the percentage change in the underlying asset price S is assumed to be the increment of a Brownian motion and if the average is computed as a geometric (as opposed to an arithmetic) average, there are analytical formulas for pricing continuous-time Asian options. However, there are no exact formulas in the case of discrete monitoring dates or different assumptions on the process followed by the asset price.

To price the option by simulation, one would generate possible paths for the underlying asset price. Let $S_{t_i}(j)$ be the simulated asset price at time t_i , $i = 1, \dots, T$, for path j , $j = 1, \dots, n$. For example, if the percentage change in the underlying asset price S is assumed to be the increment of a Brownian motion, then the asset price at time t_1 can be simulated given the asset price at time 0 as

$$S_{t_1} = S_0 e^{(r - \frac{1}{2}\sigma^2)(t_1 - 0) + \sigma\sqrt{(t_1 - 0)}\tilde{w}_0}$$

where, as defined earlier, \tilde{w}_0 is a random variable following a normal distribution with mean 0 and standard deviation 1 (the subscript "0" stands for the fact that this realization of \tilde{w} is for the time period $(0, t_1]$). Having generated a realization of S_{t_1} , one can simulate a possible

value for S_{t_2} by using the formula

$$S_{t_2} = S_{t_1} e^{(r - \frac{1}{2}\sigma^2)(t_2 - t_1) + \sigma\sqrt{(t_2 - t_1)}\tilde{w}_1}$$

and generating a realization of the normal random variable \tilde{w}_1 . After repeating this T times, one has generated a path for the asset price. Averaging the (properly discounted) option payoff over n paths produces the fair price of the Asian option.

The simulation process makes it easy to calibrate model parameters to observed market factors and to incorporate additional layers of modeling complexity. For example, suppose that at time 0 one observes a term structure of zero-bond prices $B(0, t_1), \dots, B(0, t_T)$ that is not necessarily consistent with a single interest rate r . In other words, one cannot find a short rate r such that

$$B(0, t_i) = e^{-rt_i}$$

for all intermediate time periods t_i . It would be difficult to correct for this in a closed-form formula such as the Black-Scholes formula for European options. However, the correction can be easily implemented in the simulation: one only needs to simulate future asset prices at each intermediate time period as

$$S_{t_{i+1}} = S_{t_i} \frac{B(0, t_i)}{B(0, t_{i+1})} e^{-\frac{1}{2}\sigma^2(t_{i+1} - t_i) + \sigma\sqrt{(t_{i+1} - t_i)}\tilde{w}_i}$$

Similarly, if one observes forward prices $F(0, t_1), \dots, F(0, t_T)$ on the underlying asset, one can obtain a more accurate representation of the possible scenarios in the simulation by using the formula

$$S_{t_{i+1}} = S_{t_i} \frac{F(0, t_{i+1})}{F(0, t_i)} e^{-\frac{1}{2}\sigma^2(t_{i+1} - t_i) + \sigma\sqrt{(t_{i+1} - t_i)}\tilde{w}_i}$$

The complexity of the pricing model can be increased further by incorporating realistic models for the volatility σ . The simulation technique therefore has a tremendous modeling potential.

Estimating Sensitivities

For trading, hedging, and risk management purposes, the estimation of the sensitivity of derivative prices to different inputs is sometimes even more critical than the estimation of the prices themselves. These sensitivity measures are popularly referred to as the “Greeks” because each sensitivity measure is traditionally denoted by a Greek letter. A natural way to think of evaluating the sensitivity of a derivative price to a change in an underlying parameter is to use Monte Carlo simulation to compute the price of the derivative, and then use Monte Carlo simulation again to compute the price of the derivative if the input parameter is changed by a small amount h . This kind of estimation (referred to as a “finite-difference method”), however, presents both theoretical and practical challenges. On the theoretical side, finite difference methods frequently result in a large amount of bias. On the practical side, the amount of computation required for the estimation of the sensitivity is large (double the amount of computation used in the pricing of the derivative), and can become prohibitive if this computation is done in the context of evaluating the sensitivity of a whole portfolio of securities to changes in underlying factors.

In specific circumstances, the computational burden can be reduced by finding an expression for the Greek variable of interest that can be calculated as a by-product when paths are generated in a single simulation. Such expressions exist when computing the Black-Scholes delta or the delta of an Asian option.³ These methods are referred to as “pathwise methods”—namely, the evolution of the underlying model over paths is differentiated, and the parameter with respect to which the change is computed is treated as a parameter of that evolution. For example, consider the delta (denoted by Δ) for an option price calculated with the Black-Scholes formula, where delta is defined as the (mathematical) derivative of the option value with respect to the value of the underlying asset. To

Table 1 Scenarios for Portfolio VaR Estimation

Scenario	Market Variable 1	Market Variable 2	...	Market Variable m	Change in Portfolio Value (\$ million)
1	3.54	21.54	...	0.17	100.32
2	3.27	22.03	...	0.18	101.54
...
n	3.83	22.32	...	0.15	100.87

calculate the value of delta, one would generate n paths for the evolution of the asset price, and keep track of the paths in the simulation that end up in-the-money. Let the sum of the asset prices at the end of all in-the-money paths be Ω . Then, the delta at time t can be computed as

$$\Delta_t = e^{-r(T-t)} \cdot \frac{\Omega}{S_t \cdot n}$$

More recently, efficient estimators for sensitivity from simulation trials have been developed based on Malliavin calculus.⁴

Portfolio Risk Management

Earlier, we mentioned the importance of simulation for portfolio risk measurement and management. We now explain the simulation procedure in more detail.

To estimate the portfolio VaR, for example, one would generate n possible scenarios for the possible changes in m market variables that influence the change in the portfolio value, and compute the change in portfolio value in each scenario (see Table 1). Sometimes historical data are used to create the scenarios, but typically the scenarios are generated in a more sophisticated manner. The changes in the portfolio value are then sorted, and the 95% VaR, for example, can be computed as the 5th percentile of the so-obtained empirical distribution of portfolio value changes. (This is equivalent to computing the 95% VaR as the 95th percentile of the empirical distribution of future portfolio losses, as illustrated in Figure 2.)

While this standard Monte Carlo simulation procedure is comprehensive, it can be very slow, especially when the portfolio contains complex derivative securities whose changes in value

must be reevaluated in every scenario for the market variables. In fact, the portfolio VaR calculation by simulation involves a number of “subsimulations” evaluating the sensitivities of the securities in the portfolios to each of the market variables. For large portfolios, the computational cost of generating each scenario for the change in portfolio value can become prohibitive.

In practice, several approaches are used to speed up the calculation of VaR. One of the earliest approaches, popularized by JP Morgan’s RiskMetrics software in the 1990s, is to assume that all changes in market variables are normally distributed. If the portfolio value is a linear function of these market variables (this happens, e.g., when the portfolio contains equities and factor models are used to represent the changes in asset value relative to changes in market variable values), then the change in portfolio value is also normally distributed, and can be computed in closed form, by expressing the VaR as a multiple of the standard deviation of changes in the market variables. This approach does not necessarily have to involve simulation, and actually works reasonably well for large equity-only portfolios that contain liquid assets, because the empirical distributions of their returns can be indeed very close to normal. However, it can grossly underestimate the true portfolio VaR when the portfolio contains complex derivatives (which are nonlinear functions of the returns on the underlying market variables) or fixed income securities (which depend nonlinearly on interest rates).

The nonlinearity can be partially incorporated in the estimate of the change in portfolio value by using second-order information,

so-called “Delta-Gamma” or quadratic approximation to the change in portfolio value.⁵ In other words, the change in portfolio value is expressed not only through the changes in the values of the market variables, but also through the changes in the market variables squared and scaled by their so-called Hessian matrix. (From a mathematical perspective, this is a multidimensional Taylor expansion involving the Greeks of the different securities in the portfolio.) Since traders of complex derivatives often have to keep track of this information for their own risk management purposes, the portfolio risk management process amounts to disciplined accumulation of information that is already available. This method is only an approximation, but it can reduce substantially the time for computing the portfolio VaR.

Valuing Mortgage-Backed Securities

Monte Carlo methods are often used for valuing mortgage-backed securities (MBSs) such as collateralized mortgage obligations (CMOs) and stripped MBSs (mortgage strips). The cash flows for such products can be calculated using different pricing models. The highly uncertain terms in those cash flow models, such as the behavior of interest rates and the expected prepayments over the life of the MBS, are often simulated to determine the expected cash flows to the MBS holder, which then provide the sample average (“fair”) value for the MBS.

Valuing Credit-Risky Securities

Similar ideas to those for pricing CMOs are used for pricing collateralized debt obligations (CDOs), which employ securitization to package credit-risky debt obligations (bonds and loans) in ways analogous to the way mortgages are packaged in CMOs. In order to price the CDO, one needs to simulate the defaults of different bond issuers in the collection.⁶

Simulation is also used for pricing other credit-risky instruments, such as first-to-default baskets and basket default swaps.⁷ The simulation techniques applied in such cases can be quite advanced, as credit defaults are considered “rare events” and need to be modeled with care. We will discuss the main ideas of simulation modeling techniques for rare events, such as importance sampling, later in this entry.

RANDOM NUMBER GENERATION

At the core of Monte Carlo simulation is the *generation of random numbers*. In fact, however, generating random numbers from a wide variety of distributions reduces to generating random numbers on the unit interval from 0 to 1 uniformly, that is, generating random numbers on the interval $[0,1]$ in such a way that each value between 0 and 1 is equally likely to occur. Many computer languages and software packages have a command for generating a random number between 0 and 1: “=RAND()” in Microsoft Excel, “rand(1)” in MATLAB and FORTRAN, and “rand()” in C++.

From a Uniform Random Variable to a Variable from an Arbitrary Distribution

The most common method for converting a random number between 0 and 1 to a number from an arbitrary probability distribution is to evaluate the so-called “inverse” of the cumulative probability distribution function at the random number between 0 and 1. The idea works because the total mass for a probability distribution is always 1, and the cumulative probability for any value of the distribution (defined as the probability that this particular value or any value below it will occur) is always between 0 and 1. For example, suppose that one would like to generate a random number from the normal distribution in Figure 1b. Suppose

the =RAND() command in Excel returns the number 0.975. The next step is to look for a corresponding random number from a normal distribution so that 97.5% of the probability mass (the area under the probability density curve) is to the left of that number. In Excel in particular, the function '=NORMINV(RAND(), mean, standard deviation)' can be used to find that random number on the x-axis of a normal distribution with the specified mean and standard deviation.

"Inverting" the cumulating probability distribution is trickier for discrete probability distributions, but the idea still applies. For example, suppose that given a random number generator on the interval [0,1], one would like to simulate values for a random variable that takes the value 5 with probability 50%, the value 15 with probability 30%, and the value 35 with probability 20%. Let us split the unit interval [0,1] into three intervals based on the cumulative probabilities 50%, 80% and 100% for obtaining the values 5, 15, and 35: [0,0.5], (0.5,0.8], and (0.8,1]. If the random number that is drawn falls in the interval [0,0.5] (which happens 50% of the time if the number generator is truly random), then one records a value of 5 for that trial. If the random number is in the interval (0.05, 0.8] (which happens with probability 30%), then one records a value of 15 for that trial. Finally, if the random number is in the third interval (which happens with probability 20%), one records a value of 35. Thus, if many trials are run, the values 5, 15, and 35 are generated with the desired probabilities. In Excel, one can simulate these values with the corresponding probabilities by creating a table with the interval ranges in the first two columns, and the corresponding values (5, 15, and 35) in the third column, and using the Excel function

VLOOKUP(lookup_value, table_array, col_index_num)

to look up the range in which a number generated with RAND() falls.⁸

What Defines a "Good" Random Number Generator?

Given the discussion in the previous section, generating "good" uniform random numbers on [0,1] is critical for the performance of simulation algorithms. Interestingly, defining "good" random number generation is not as straightforward as it appears. Early random number generators tried to use "truly random" events for random number generation, such as the amount of background cosmic radiation. In practice, however, this kind of random number generation is time consuming and difficult. Moreover, it was realized that the ability to reproduce the random number sequence and to analyze the random number characteristics is actually a desirable property for random number generators. In particular, the ability to reproduce a sequence of random numbers allows for reducing the variance of estimates and for debugging computer code by rerunning experiments in the same conditions in which they were run in previous iterations of code development.

Most simulation software products employ random number generation algorithms that produce streams of numbers that appear to be random, but are in fact a result of a clearly defined series of calculation steps in which the next "random number" x_n in the sequence is a function of the previous "random number" x_{n-1} , that is, $x_n = f(x_{n-1})$. The sequence starts with a number called the seed, and if the same seed is used in several simulations, each simulation sequence will contain exactly the same numbers, which is helpful for code debugging and drawing fair comparisons between different strategies evaluated under uncertainty. It is quite an amazing statistical fact that some of these recursion formulas (named "pseudo-random number generators") define sequences of numbers that imitate random behavior well and appear to obey (roughly) some major laws of probability, such as the CLT and the Glivenko-Cantelli lemma.

In general, a pseudo-random number generator is considered “good” if it satisfies the following conditions:

1. The numbers in the generated sequence are uniformly distributed between 0 and 1. This can be tested by running a chi-square or a Kolmogorov-Smirnov test.
2. The sequence has a long cycle (that is, it takes many iterations before the sequence begins repeating itself).
3. The numbers in the sequence are not autocorrelated. This can be verified by running a Durbin-Watson test on the sequence of numbers. The Durbin-Watson test is widely used in statistics for identifying autocorrelation in time series of observations.

In the following section, we discuss briefly a couple of important types of pseudo-random number generators. The goal is not to provide comprehensive coverage of random number generators, but rather to give readers a flavor of the main ideas behind the method of producing apparently random numbers with deterministic algorithms.

Pseudo-Random Number Generators

One of the earliest pseudo-random number generators developed is called the midsquare technique. It takes a number (the seed), squares it, and takes the set of middle digits as the next random number. It is easy to predict when such an approach may run into difficulties. As soon as the “middle digits” become a small number such as 1 or 0, the sequence ends with the same numbers generated over and over again; that is, the sequence converges to a constant value (typically 0) or to a very short cycle of values.

A better, commonly used type of pseudo-random number generators is congruential pseudo-random number generators. They are based on sequences of numbers of the form

$$x_n = f(x_{n-1}) \bmod m$$

where $\bmod m$ stands for “modulus m ”. $f(x_{n-1}) \bmod m$ is the remainder after dividing $f(x_{n-1})$ by m . For example, $5 \bmod 3 = 2$, $15 \bmod 5 = 0$, etc. Note that $f(x_{n-1}) \bmod m$ will always be an integer between 0 and $m-1$. Thus, to create a good representation of randomness, one would want to make the range for the modulus as large as possible. For a 32-bit computer, for example, the maximum integer that can be stored is $2^{31} - 1$, which is large enough for practical purposes.

More advanced generators include matrix multiplicative congruential generators, multiple recursive generators, and shuffled generators. Most pseudo-random number generators used in popular software products nowadays have been thoroughly tested and are very good.

VARIANCE REDUCTION TECHNIQUES

Paradoxically, truly random numbers can be too random for all practical purposes. Recall that the error in the average estimate obtained from truly random Monte Carlo simulation is proportional to $1/\sqrt{n}$, where n is the number of scenarios for the random variable (this fact would be approximately true for good pseudo-random number generators as well). Much research has been dedicated in recent years to finding ways to reduce that error and to be computationally savvy when generating scenarios. Several methods for *variance reduction*, widely used in financial applications, are listed below.⁹

Antithetic Variables

Simulating a random number is computationally expensive. One technique that is used to reduce the error in the average estimate in derivative pricing without increasing the number of simulated values is to incorporate the generated random number twice in computing the derivative payoff: once as the original simulated number, and another as its “antithetic” number.

For example, recall from our earlier option pricing example that one possibility to model the value of the stock price S_T at time T is by using equation (2). In that expression, \tilde{w} is a random variable following a normal distribution with mean 0 and standard deviation 1. Suppose that n values for the normal random variable \tilde{w} are generated. With the *antithetic variable* method, the value of the derivative payoff in each of the n scenarios is computed as the average of two payoffs: one obtained by plugging in the simulated value for \tilde{w} , and another obtained by plugging in the negative of the simulated value for \tilde{w} . These n “adjusted” payoffs are otherwise treated in the same way as in the traditional simulation method described earlier in this entry: At the end, the n payoffs are averaged and properly discounted to obtain the “fair” estimate of the derivative price. The difference is that this approach substantially reduces the standard error in the average estimate, while keeping the number of simulation trials at n .

The antithetic variable approach does not apply only to normal random variables. As explained in the previous section, random number generation typically happens in two stages: First, a random number between 0 and 1 is generated, and then this random number is “inverted” to obtain a random number from the desired probability distribution. Thus, one can apply the antithetic technique at the first stage, and treat the randomly generated number U as two realizations: U and its “antithetic” variable $1-U$. For example, if the number generated on the interval $[0,1]$ is 0.7, then the antithetic number is 0.3. Both of these numbers can then be “inverted” to obtain a pair of *antithetic variables* from a prespecified distribution.

Stratified Sampling

Observations in the tails of input distributions that are typically less likely to be generated may never occur in a simulation, because the probability of their occurrence is small. Such observations, however, contain important information about extreme events which are of partic-

ular interest in financial applications. In order to ensure that they appear in the simulation, one would need to generate a huge number of scenarios.

This problem is often addressed by *stratified sampling*. Most generally, the term “stratified sampling” refers to any technique that divides the random values into ranges (called “strata” in statistics), and sampling from each range to ensure that a good representation of the distribution is obtained.

A simple way of stratifying the numbers in the $[0,1]$ interval to ensure that, when “inverted,” the generated random numbers cover well the whole range of a probability distribution of interest, is to divide the $[0,1]$ interval into k smaller intervals of equal length:

$$\left[0, \frac{1}{k}\right], \left(\frac{1}{k}, \frac{2}{k}\right], \dots, \left(\frac{k-1}{k}, 1\right]$$

Random numbers can then be drawn sequentially from each small interval. Therefore, values from the tails of the distribution of interest (which will be generated when uniform random numbers from the intervals $[0, \frac{1}{k}]$ and $(\frac{k-1}{k}, 1]$ are drawn) obtain better representation.

In multiple dimensions (that is, when simulating several random variables), this method extends to dividing a hypercube (as opposed to an interval) into smaller hypercubes, and drawing an observation along each dimension of the smaller hypercubes. An enhanced extension to the basic stratified sampling method is Latin hypercube sampling (an option in many advanced simulation software products), which permutes the coordinates of an initially generated random vector of observations—one observation within each small hypercube—to reduce the number of times an actual random number is generated while ensuring that all strata are sufficiently well represented.

Importance Sampling

Importance sampling is an alternative to stratified sampling for dealing with rare events, or

extreme observations, and for reducing the number of simulation trials necessary to achieve a particular level of accuracy. The method changes the underlying scenario probabilities so as to give more weight to important outcomes in the simulation. Such outcomes are generated with greater frequency than they otherwise would. At the end, the observations' weights are scaled back in the computation of the parameter of interest, so that the estimation is correct.

There is no single recipe for how to construct good importance sampling methods. The specific construction depends on the underlying random process dynamics. For example, when pricing a European call option in the Black-Scholes setting, generating paths that are out-of-the-money is wasteful. This is because only paths that are in-the-money count in the final computation of the option price—the contribution of out-of-the-money paths to the option price is 0. Although in practice one would not use importance sampling for pricing a European call option for which there is a closed-form formula, we will use European call option pricing as a context in which to explain the importance sampling method.

First, note that in-the-money paths will occur only if the asset price at expiration is greater than the strike price; that is, they will result from realizations of the standard normal random variable \tilde{w} such that

$$S_t e^{(r - \frac{1}{2}\sigma^2)(T-t) + \sigma\sqrt{T-t}\tilde{w}} > K$$

From this inequality, one can derive that only normal random numbers higher than

$$\frac{\ln(K/S_t) - (r - \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}}$$

will lead to in-the-money paths. Equivalently, this means that only random numbers between

$$N\left(\frac{\ln(K/S_t) - (r - \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}}\right) \text{ and } 1$$

on the unit interval [0,1], when “inverted” to obtain normal random numbers, will lead to in-

the-money paths ($N(\cdot)$ here denotes the cumulative normal distribution). Thus, one only needs to simulate random numbers in that range of the [0,1] interval. When computing the option price at the end, instead of weighing each payoff equally by multiplying it by $1/n$ as one would do in standard Monte Carlo sampling, one multiplies the sum of the payoffs obtained from the simulation by the probability that a particular random path would be in-the-money assuming truly random sampling, which is the standard Monte Carlo method. The latter probability is

$$\begin{aligned} &1 - N\left(\frac{\ln(K/S_t) - (r - \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}}\right) \\ &= N\left(\frac{\ln(S_t/K) + (r - \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}}\right) \end{aligned}$$

The call option price is then

$$\begin{aligned} V_t &= e^{-r(T-t)} \cdot N\left(\frac{\ln(S_t/K) + (r - \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}}\right) \\ &\cdot \sum_{i=1}^n \max\left\{S_t e^{(r - \frac{1}{2}\sigma^2)(T-t) + \sigma\sqrt{T-t}w_i} - K, 0\right\} \end{aligned}$$

where w_1, \dots, w_n are all random numbers generated from a normal distribution in the range higher than

$$\frac{\ln(K/S_t) - (r - \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}}$$

As mentioned above, this is only a simple example in order to illustrate the main idea of importance sampling. More practical (albeit more technically challenging) applications can be found, for instance, in Chapter 4.6 in Glasserman (2004).

Quasi-Random (Low-Discrepancy) Sequences

A truly random number generator may produce clustered observations (see Figure 3a), which necessitates generating many scenarios in order to obtain a good representation of the output distribution of interest. Recall from our earlier discussion that stratified sampling can

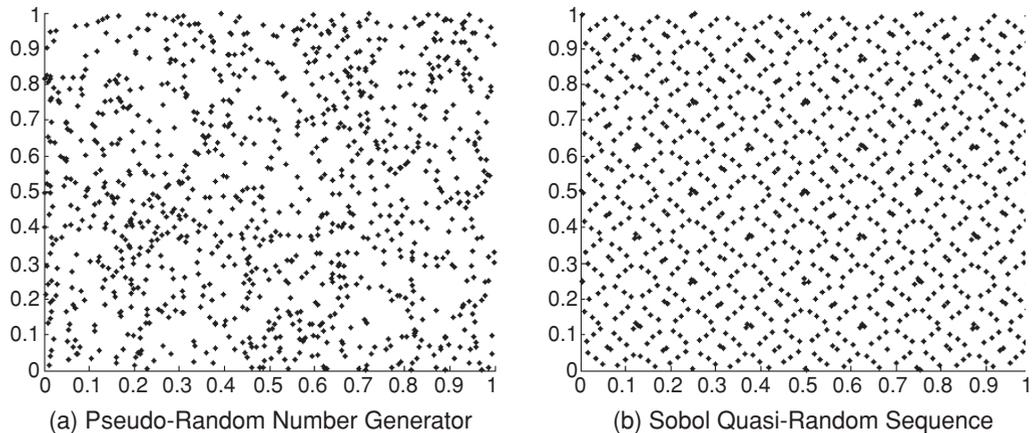


Figure 3 One Thousand Simulated Number Values for Two Uniform Random Variables on the Interval $[0,1]$

be used to deal with this problem—it divides the ranges of possible values into a fixed number of strata, so as to “disperse” observations more evenly over the range. Quasi-random sequences instead ensure a smooth representation of the range by continuously “filling in” gaps on the unit interval $[0,1]$ left by previously generated random numbers (see an example of 1,000 generated values of a quasi-random sequence in Figure 3b). The term “quasi-random” is actually a misnomer, because, unlike pseudo-random number sequences, quasi-random number sequences do not pretend to be random. They are deterministic on purpose, and their roots can be found in real analysis and abstract algebra rather than in simulation or probability theory. The term *low discrepancy sequences* is often used interchangeably with the term “quasi-random” sequences, and is more accurate.

Important examples of quasi-random sequences were suggested by Sobol (1967), Faure (1982), Halton (1960), and Hammersley (1960). These sequences build on a family of so-called Van der Korput sequences.¹⁰ For example, the Van der Korput sequence of base 2 is

$$0, \frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{5}{8}, \frac{3}{8}, \frac{7}{8}, \dots$$

The actual generation of Van der Korput sequences is somewhat technical, but the outcome

is intuitive. Note that as new points are added to the sequence, they appear on alternate sides of $\frac{1}{2}$ in a balanced way. The main idea is that as the number of generated values increases, the sequence covers uniformly the unit interval.

The values generated with quasi-random sequences are treated as “random” numbers for the purposes of simulation modeling. In particular, instead of generating random numbers between 0 and 1 and “inverting” them to obtain an arbitrary probability distribution, one would “invert” the numbers in the quasi-random sequence. Different sequences have different advantages for specific financial applications, but the Faure and Sobol sequences in particular have been proven to generate very accurate estimates for derivative pricing in tests.¹¹

Figure 4 illustrates the value of a European call option computed with three different methods: BS (the closed-form Black-Scholes price), MC (traditional Monte-Carlo), and QMC (quasi-random or quasi-Monte-Carlo using a Faure low discrepancy sequence to generate scenarios). The current asset price is assumed to be \$100, the exercise price for the option is assumed to be \$95, the asset volatility is 20%, the time to maturity of the option is 1 year, and the risk-free rate is 4% per annum. One can observe that as the number of scenarios generated increases, the quasi-Monte-Carlo method

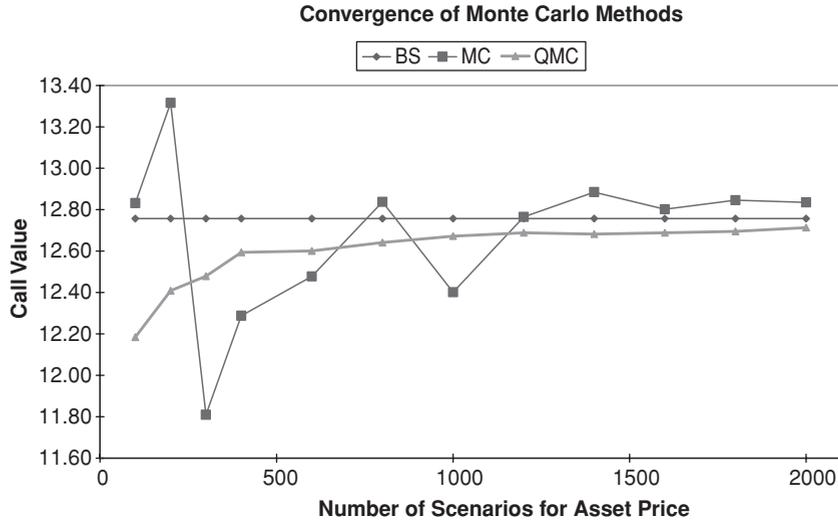


Figure 4 Value of a European Call Option Computed with Three Different Methods

results in a smoother and more consistent approximation to the true option price computed with the Black-Scholes formula than the traditional Monte Carlo method. In general, as the number of generated quasi-random numbers increases, so does the accuracy of estimation, although it is not easy to state the exact level of accuracy, because probability laws do not apply to deterministic sequences.

SIMULATION SOFTWARE

Today, good random number generators and user-friendly simulation software are easily available. Most computer languages have a “rand()” command that simulates a random number between 0 and 1. Microsoft Excel add-ins such as Crystal Ball and @RISK allow not only for simulating random numbers from a wide variety of probability distributions, but also for incorporating random number generation into larger models through macros and scripts. Computing environments such as Matlab and Mathematica contain commands for random number simulation, and the capability of generating low discrepancy sequences can be added through widely available libraries. In addition, a number of modules that allow

for simulating sophisticated probability distributions are available for open-source computer languages such as Perl (see the Comprehensive Perl Archive Network, <http://www.cpan.org>), Python (see <http://www.python.org>), and R (see <http://www.r-project.org>).

KEY POINTS

- The main idea behind the Monte Carlo simulation technique is to represent uncertainty in the form of scenarios and to evaluate variables of interest based on these scenarios.
- Monte Carlo simulation has widespread applications in pricing, hedging, and risk management. Examples include complex financial derivative pricing, assessment of sensitivity of prices to changes in market variables, portfolio risk measurement, and credit risk estimation and pricing.
- Despite great advances in computational power, Monte Carlo simulation can be expensive for large-scale problems, and a substantial amount of research in recent years has been dedicated to making it more efficient and accurate.
- Variance reduction methods such as antithetic variables, stratified sampling,

importance sampling, and carefully selected low discrepancy sequences are widely used in practice today.

NOTES

1. For an introduction to Brownian motion, see Hull (2005).
2. See, for example, Chapter 9 in Glasserman (2004).
3. See Chapter 7.2 in Glasserman (2004).
4. See Chen and Glasserman (2006a) for further details.
5. See Glasserman et al. (2000).
6. For example, see Duffie and Garleanu (2001).
7. See Chen and Glasserman (2006b).
8. See, for example, Chapter 2 in Evans and Olson (2002).
9. For a more detailed discussion of such methods, see Chapter 14 in Pachamanova and Fabozzi (2010).
10. See Chapter 5 in Glasserman (2004).
11. See the survey in Boyle et al. (1997).

REFERENCES

- Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 3: 637–654.
- Boyle, P. (1977). Options: A Monte Carlo approach. *Journal of Financial Economics* 4, 3: 323–338.
- Boyle, P., Broadie, M., and Glasserman, P. (1997). Monte Carlo methods for security pricing. *Journal of Economic Dynamics & Control* 21: 1267–1321.
- Chen, N., and Glasserman, P. (2006a). Malliavin Greeks without Malliavin calculus. Columbia Business School working paper.
- Chen, Z., and Glasserman, P. (2006b). Fast pricing of basket default swaps. Columbia Business School working paper.
- Duffie, D., and Garleanu, N. (2001). Risk and valuation of collateralized debt obligations. *Financial Analysts Journal* 57, 1: 41–59.
- Evans, J., and Olson, D. (2002). *Introduction to Simulation and Risk Analysis*, 2nd Edition. Upper Saddle River, NJ: Prentice Hall.
- Faure, H. (1982). Discr pance de suites associ es   un syst me de num ration (en dimension s). *Acta Arithmetica* 41: 337–351.
- Fishman, G. (2006). *A First Course in Monte Carlo*. Belmont, CA: Duxbury Press, Thomson Brooks/Cole.
- Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*, New York: Springer-Verlag.
- Glasserman, P., Heidelberger, P., and Shahabuddin, P. (2000). Variance reduction techniques for value-at-risk estimation. *Management Science* 46: 1349–1364.
- Halton, J. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numerische Mathematik* 2: 84–90.
- Hammersley, J. (1960). Monte Carlo methods for solving multivariable problems. *Annals of the New York Academy of Sciences* 86: 844–874.
- Hull, J. (2005). *Options, Futures and Other Derivatives*, 6th Edition. Upper Saddle River, NJ: Prentice Hall.
- Pachamanova, D. A., and Fabozzi, F. J. (2010). *Simulation and Optimization in Finance: Modeling with MATLAB, @RISK, and VBA*. Hoboken, NJ: John Wiley & Sons.
- Sobol, I. (1967). The distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics* 7, 4: 86–112.

Stochastic Volatility

ANATOLIY SWISHCHUK, PhD

Professor of Mathematics and Statistics, University of Calgary

Abstract: Volatility, as measured by the standard deviation, is an important concept in financial modeling because it measures the change in value of a financial instrument over a specific horizon. The higher the volatility, the greater the price risk of a financial instrument. There are different types of volatility: historical, implied volatility, level-dependent volatility, local volatility, and stochastic volatility (e.g., jump-diffusion volatility). Stochastic volatility models are used in the field of quantitative finance. Stochastic volatility means that the volatility is not a constant, but a stochastic process and can explain volatility smile and skew.

Volatility, typically denoted by the Greek letter σ , is the standard deviation of the change in value of a financial instrument over a specific horizon such as a day, week, month, or year. It is often used to quantify the price risk of a financial instrument over that time period. The price risk of a financial instrument is higher the greater its volatility.

Volatility is an important input in option pricing models. The Black-Scholes model for option pricing assumes that the volatility term is a constant. This assumption is not always satisfied in real-world options markets because probability distribution of common stock returns has been observed to have a fatter left tail and thinner right tail than the lognormal distribution (see Hull, 2000). Moreover, the assumption of constant volatility in a financial model, such as the original Black-Scholes option pricing model, is incompatible with option prices observed in the market.

As the name suggests, stochastic volatility means that volatility is not a constant, but a stochastic process. *Stochastic volatility* models are used in the field of quantitative finance and financial engineering to evaluate derivative securities, such as options and swaps. By assuming that volatility of the underlying price is a stochastic process rather than a constant, it becomes possible to more accurately model derivatives. In fact, stochastic volatility models can explain what is known as the volatility smile and volatility skew in observed option prices.

In this entry, we provide an overview of the different types of nonstochastic volatilities and the different types of stochastic volatilities. There are two approaches to introduce stochastic volatility: (1) changing the clock time t to a random time $T(t)$ (*subordinator*), and (2) changing constant volatility into a positive stochastic process.

NONSTOCHASTIC VOLATILITY MEASURES

We begin by providing an overview of the different types of nonstochastic volatility measure. These include

- Historical volatility
- Implied volatility
- Level-dependent volatility
- Local volatility

Historical Volatility

Historical volatility is the volatility of a financial instrument or a market index based on historical returns. It is a standard deviation calculated using historical (daily, weekly, monthly, quarterly, yearly) price data. The annualized volatility σ is the standard deviation of the instrument's logarithmic returns over a one-year period:

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (R_i - \bar{R})^2}$$

where $R_i = \ln \frac{S_i}{S_{i-1}}$, $\bar{R} = \frac{1}{n} \sum_{i=1}^n \ln \frac{S_i}{S_{i-1}}$, S_{t_i} is an asset price at time t_i , $i = 1, 2, \dots, n$.

Implied Volatility

Implied volatility is related to historical volatility. However, there are important differences. Historical volatility is a direct measure of the movement of the price (realized volatility) over recent history. Implied volatility, in contrast, is set by the market price of the derivative contract itself, and not the underlier. Therefore, different derivative contracts on the same underlier have different implied volatilities. Most derivative markets exhibit persistent patterns of volatilities varying by strike. The pattern displays different characteristics for different markets. In some markets, those patterns form a smile curve. In others, such as equity index options markets, they form more of a skewed curve.

This has motivated the name “volatility skew.” For markets where the graph is downward sloping, such as for equity options, the term “volatility skew” is often used. For other markets, such as FX options or equity index options, where the typical graph turns up at either end, the more familiar term “volatility smile” is used. In practice, either term may be used to refer to the general phenomenon of volatilities varying by strike.

The models by Black and Scholes (1973) (continuous-time (B,S)-security market) and Cox, Ross, and Rubinstein (1976) (discrete-time (B,S)-security market (binomial tree)) are unable to explain the negative skewness and leptokurticity (fat tail) commonly observed in the stock markets. The famous implied-volatility smile would not exist under their assumptions. Most derivatives markets exhibit persistent patterns of volatilities varying by strike. In some markets, those patterns form a smile. In others, such as equity index options markets, it is more of a skewed curve. This has motivated the name volatility skew. In practice, either the term volatility smile or volatility skew (or simply skew) may be used to refer to the general phenomenon of volatilities varying by strike. Another dimension to the problem of volatility skew is that of volatilities varying by expiration, known as volatility surface.

Given the prices of call or put options across all strikes and maturities, we may deduce the volatility that produces those prices via the full Black-Scholes equation.¹ This function has come to be known as *local volatility*. Local volatility-function of the spot price S_t and time t : $\sigma \equiv \sigma(S_t, t)$ (see Dupire's (1994) formulas for local volatility).

Level-Dependent Volatility

Level-dependent volatility (e.g., constant elasticity of variance (CEV) or firm model, see Beckers, 1980, Cox, 1975) is a function of the spot price alone. To have a smile across strike price,

we need σ to depend on $S : \sigma \equiv \sigma(S_t)$. In this case, the volatility and stock price changes are now perfectly negatively correlated (so-called “leverage effect”).

Local Volatility

Local volatility (LV) is a volatility function of the spot price and time. Volatility smile can be retrieved in this case from the option prices. Dupire (1994) derived the local volatility formula in continuous time and Derman and Kani (1994) used the binomial (or trinomial tree) framework instead of the continuous one to find the local volatility formula. The LV models are very elegant and theoretically sound. However, they present in practice many stability issues. They are ill-posed inversion problems and are extremely sensitive to the input data. This might introduce arbitrage opportunities and, in some cases, negative probabilities or variances.

Stochastic Volatility

Stochastic volatility means that volatility is not a constant, but a stochastic process. Black and Scholes (1973) made a major breakthrough by deriving pricing formulas for vanilla options written on the stock. The Black-Scholes model assumes that the volatility term is a constant. Stochastic volatility models are used in the field of quantitative finance to evaluate derivative securities, such as options and swaps (see Carr and Lee, 2009). By assuming that the volatility of the underlying price is a stochastic process rather than a constant, it becomes possible to more accurately model derivatives.

The above issues have been addressed and studied in several ways, such as:

1. Volatility is assumed to be a deterministic function of the time:² $\sigma \equiv \sigma(t)$, with the implied volatility for an option of maturity T given by $\hat{\sigma}_T^2 = \frac{1}{T} \int_0^T \sigma_u^2 du$;
2. Volatility is assumed to be a function of the time and the current level of the stock price

$S(t) : \sigma \equiv \sigma(t, S(t))$;³ the dynamics of the stock price satisfies the following stochastic differential equation:

$$dS(t) = \mu S(t)dt + \sigma(t, S(t))S(t)dW_1(t)$$

where $W_1(t)$ is a standard Wiener process;

3. The time variation of volatility involves an additional source of randomness, besides $W_1(t)$, represented by $W_2(t)$, and is given by

$$d\sigma(t) = a(t, \sigma(t))dt + b(t, \sigma(t))dW_2(t)$$

where $W_2(t)$ and $W_1(t)$ (the initial Wiener process that governs the price process) may be correlated;⁴

4. Volatility depends on a random parameter x such as $\sigma(t) \equiv \sigma(x(t))$, where $x(t)$ is some random process.⁵
5. Stochastic volatility, namely, uncertain volatility scenario. This approach is based on the uncertain volatility model developed in Avellaneda et al. (1995), where a concrete volatility surface is selected among a candidate set of volatility surfaces. This approach addresses the sensitivity question by computing an upper bound for the value of the portfolio under arbitrary candidate volatility, and this is achieved by choosing the local volatility $\sigma(t, S(t))$ among two extreme values σ_{\min} and σ_{\max} such that the value of the portfolio is maximized locally;
6. The volatility $\sigma(t, S_t)$ depends on $S_t = S(t + \theta)$ for $\theta \in [-\tau, 0]$, namely, stochastic volatility with delay.⁶

In approach (1), the volatility coefficient is independent of the current level of the underlying stochastic process $S(t)$. This is a deterministic volatility model, and the special case where σ is a constant reduces to the well-known Black-Scholes model that suggests changes in stock prices are lognormal. Empirical tests by Bollerslev (1986) seem to indicate otherwise. One explanation for this problem of a lognormal model is the possibility that the variance of $\log(S(t)/S(t - 1))$ changes randomly.

In approach (2), several ways have been developed to derive the corresponding Black-Scholes formula: One can obtain the formula by using stochastic calculus and, in particular, Ito's formula (see Shiryaev (1999), for example).

A generalized volatility coefficient of the form $\sigma(t, S(t))$ is said to be *level-dependent*. Because volatility and asset price are perfectly correlated, we have only one source of randomness given by $W_1(t)$. A time and level-dependent volatility coefficient makes the arithmetic more challenging and usually precludes the existence of a closed-form solution. However, the arbitrage argument based on portfolio replication and completeness of the market remains unchanged.

Approaches to Introduce Stochastic Volatility

The idea to introduce stochastic volatility is to make volatility itself a stochastic process. The aim with a stochastic volatility model is that volatility appears not to be constant and indeed varies randomly. For example, the situation becomes different if volatility is influenced by a second "nontradable" source of randomness, and we usually obtain a *stochastic volatility model*, introduced by Hull and White (1987). This model of volatility is general enough to include the deterministic model as a special case. Stochastic volatility models are useful because they explain in a self-consistent way why it is that options with different strikes and expirations have different Black-Scholes implied volatilities (the volatility smile). These cases are addressed in approaches 3, 4 and 5 above. Stochastic volatility is the main concept used in the fields of financial economics and mathematical finance to deal with the endemic time-varying volatility and codependence found in financial markets. Such dependence has been known for a long time; early comments include Mandelbrot (1963) and Officer (1973).

There are two approaches to introduce stochastic volatility: One approach is to change

the clock time t to a random time $T(t)$ (change of time). Another approach is to change constant volatility into a positive stochastic process. Continuous-time stochastic volatility models include:

- Ornstein-Uhlenbeck (OU) process (Ornstein-Uhlenbeck (1930))
- Geometric Brownian motion with zero correlation with respect to a stock price (Hull and White, 1987)
- Geometric Brownian motion with nonzero correlation with respect to a stock price (Wiggins, 1987)
- OU process, mean-reverting, positive with nonzero correlation with respect to a stock price (Scott, 1989)
- OU process, mean-reverting, negative, with zero correlation with respect to a stock price (Stein and Stein, 1991)
- Cox-Ingersoll-Ross process, mean-reverting, nonnegative with non zero correlation with respect to a stock price (Heston, 1993).

Heston and Nandi (1997) showed that the OU process corresponds to a special case of the GARCH model for stochastic volatility. Hobson and Rogers (1998) suggested a new class of nonconstant volatility models, which can be extended to include the aforementioned level-dependent model and share many characteristics with the stochastic volatility model. The volatility is nonconstant and can be regarded as an endogenous factor in the sense that it is defined in terms of the past behavior of the stock price. This is done in such a way that the price and volatility form a multidimensional Markov process.

Discrete Models for Stochastic Volatility

Another popular process is the continuous-time GARCH(1,1) process, developed by Engle (1982) and Bollerslev (1986) in a discrete framework. The generalized autoregressive conditional heteroskedasticity (GARCH) model

(see Bollerslev, 1986) is popular model for estimating stochastic volatility. It assumes that the randomness of the variance process varies with the variance, as opposed to the square root of the variance as in the Heston model. The standard GARCH(1,1) model has the following form for the variance differential:

$$d\sigma_t = \kappa(\theta - \sigma_t)dt + \gamma\sigma_t dB_t$$

The GARCH model has been extended via numerous variants, including the NGARCH, LGARCH, EGARCH, GJR-GARCH, and so on.

Continuous-time models provide the natural framework for an analysis of option pricing; discrete-time models are ideal for the statistical and descriptive analysis of the patterns of daily price changes. Volatility clustering, periods of high and low variance (large changes tend to be followed by small changes; see Mandelbrot, 1963), led to using discrete models, GARCH models. There are two main classes of discrete-time stochastic volatility models. First class—autoregressive random variance (ARV) or stochastic variance model—is a discrete time approximation to the continuous time diffusion models that we outlined above. Second class is the autoregressive conditional heteroskedastic (ARCH) model introduced by Engle (1982), and its descendants GARCH (Bollerslev, 1986), NARCH, NGARCH (Duan, 1996), LGARCH, EGARCH, GJR-GARCH. General class of stochastic volatility models, which includes many of the above-mentioned models, has been introduced by Ewald, Poulsen, and Schenk-Hoppe (2006). Gatheral (2006) introduce the Heston-like model for stochastic volatility that is more general than the Heston model.

Jump-Diffusion Volatility

Jump-diffusion volatility is essential as there is evidence that assumption of a pure diffusion for the stock return is not accurate. Fat tails have been observed away from the mean of the stock return. This phenomenon is called leptokurticity and could be explained in different

ways. One way to explain smile and leptokurticity is to introduce a jump-diffusion process for stochastic volatility (see Bates, 1996). Jump-diffusion is not a level-dependent volatility process, but can explain the leverage effect.

Multifactor Models for Stochastic Volatility

One-factor SV models (all above-mentioned): (1) incorporate the leverage between returns and volatility and (2) reproduce the skew of implied volatility. However, they fail to match either the high conditional kurtosis of returns (Chernov et al., 2003) or the full term structure of implied volatility surface (Cont et al., 2004). Two primary generalizations of one-factor SV models are: (1) adding jump components in returns and/or volatility process, and (2) considering multifactor SV models. Among multifactor SV models we mention here the following ones:

- Fouque et al. (2005) SV model, Chernov et al. (2003) model (used efficient method of moments to obtain comparable empirical-of-fit from affine jump-diffusion and two-factor SV family models).
- Molina et al. (2003) model (used a Markov chain Monte Carlo method to find strong evidence of two-factor SV models with well-separated time scales in foreign exchange data).
- Cont et al. (2004) (found that jump-diffusion models have a fairly good fit to the implied volatility surface).
- Fouque et al. (2000) model (found that two-factor SV models provide a better fit to the term structure of implied volatility than one-factor SV models by capturing the behavior at short and long maturities).
- Swishchuk (2006) (introduced two-factor and three-factor SV models with delay (incorporating mean-reverting level as a random process geometric Brownian model, OU, continuous-time GARCH(1,1) model).

We also mention the SABR model (see Hagan et al., 2002), describing a single forward under stochastic volatility, and Chen's (1996) three-factor model for the dynamics of the instantaneous interest rate.

Multifactor SV models have advantages and disadvantages. One of the disadvantages is that multifactor SV models do not admit in general explicit solutions for option prices. One of the advantages is that they have direct implications for hedges. As a comparison, a class of jump-diffusion models (Bates, 1996) enjoys closed-form solutions for option prices. But the complexity of hedging strategies increases due to jumps. In this way, there is no strong empirical evidence to judge the overwhelming position of jump-diffusion models over multifactor SV models or vice versa.

The probability literature demonstrates that stochastic volatility models are fundamental notions⁷ in financial markets analysis.

KEY POINTS

- Because it measures the change in value of a financial instrument over a specific horizon, volatility, as measured by the standard deviation, is an important concept in financial modeling.
- The different types of volatility are historical, implied, jump-diffusion, level-dependent, local, and stochastic volatilities.
- Stochastic volatility means that the volatility is not a constant, but a stochastic process. Stochastic volatility can explain the well-documented volatility smile and skew observed in option markets.
- Stochastic volatility is the main concept used in finance to deal with the endemic time-varying volatility and codependence found in financial markets and stochastic volatility models are used to evaluate derivative securities such as options and swaps.
- Two approaches to introduce stochastic volatility are: (1) changing the clock time to

a random time and (2) changing constant volatility into a positive stochastic process.

NOTES

1. Black and Scholes (1973), Dupire (1994), Derman and Kani (1994).
2. Wilmott et al. (1995), Merton (1976).
3. Dupire (1994), Hull (2000).
4. Hull and White (1987), Heston (1993).
5. Elliott and Swishchuk (2007), Swishchuk (2000, 2009), Swishchuk et al. (2010).
6. Kazmerchuk, Swishchuk, and Wu (2005), Swishchuk (2005, 2006, 2007, 2009a, 2010).
7. Barndor-Nielsen, Nicolato, and Shephard (1996), Shephard (2005).

REFERENCES

- Ahn, H., and Wilmott, P. (2006). *Stochastic Volatility and Mean-Variance Analysis*. New York: Wiley/Finance.
- Avellaneda, M., Levy, A., and Paras, A. (1995). Pricing and hedging derivative securities in markets with uncertain volatility. *Applied Mathematical Finance* 2: 73–88.
- Barndor-Nielsen, O. E., Nicolato, E., and Shephard, N. (1996). Some recent developments in stochastic volatility modeling. *Quantitative Finance* 2: 11–23.
- Bates, D. (1996) Jumps and stochastic volatility: The exchange rate processes implicit in Deutschemark options. *Review Finance Studies* 9: 69–107.
- Beckers, S. (1980) The constant elasticity of variance model and its implications for option pricing. *Journal of Finance* 35: 661–673.
- Black, F., and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81: 637–54.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Economics* 31: 307–27.
- Carr P. and Lee, R. (2009). Volatility derivatives. *Annual Review of Financial Economics* 1:1–21.
- Chen, L. (1996). Stochastic mean and stochastic volatility—A three-factor model of the term structure of interest rates and its application to the pricing of interest rate derivatives.

- Financial Markets, Institutions and Instruments* 5: 1–88.
- Chernov, R., Gallant, E., Ghysels, E., and Tauchen, G. (2003). Alternative models for stock price dynamics. *Journal of Econometrics* 116: 225–257.
- Cont, R., and Tankov, P. (2004). *Financial Modeling with Jump Processes*. New York: Chapman & Hall/CRC Fin. Math. Series.
- Cox, J. (1975) Notes on option pricing I: Constant elasticity of variance diffusions. Stanford, CA: Stanford University, class notes.
- Cox, J., Ingersoll, J., and Ross, S. (1985). A theory of the term structure of interest rate. *Econometrica* 53: 385–407.
- Derman, E., and Kani, I. (1994). Riding on a smile. *Risk* 7(2): 32–39.
- Duan, J. (1996). The GARCH option pricing model. *Mathematical Finance* 5, 1: 13–32.
- Dupire, B. (1994). Pricing with a smile. *Risk*, 7(1): 18–20.
- Elliott, R., and Swishchuk, A. (2007). Pricing options and variance swaps in Markov-modulated Brownian markets. In *Hidden Markov Models in Finance*. New York: Springer, pp. 45–68.
- Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 4: 987–1007.
- Fouque, J.-P., and Han, C.-H. (2003). A control variate method to evaluate option prices under multi-factor stochastic volatility models. Working paper. Santa Barbara, CA: University of California.
- Fouque, J.-P., Papanicolaou, G., and Sircar, K. R. (2000). *Derivatives in Financial Markets with Stochastic Volatilities*. New York: Springer.
- Gatheral, J. (2006). *The Volatility Surface. A Practitioner's Guide*. New York: Wiley.
- Hagan, P., Kumar, D., Lesniewski, S., and Woodward, D. (2002). Managing smile risk. *Wilmott Magazine*, July 26, 84–108.
- Heston, S. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies* 6: 327–343.
- Heston, S., and Nandi, S. (1998). Preference-free option pricing with path-dependent volatility: A closed-form approach. Discussion Paper. Atlanta: Federal Reserve Bank of Atlanta.
- Hobson, D., and Rogers, L. (1998). Complete models with stochastic volatility. *Mathematical Finance* 8, 1: 27–48.
- Hull, J. (2000). *Options, Futures and Other Derivatives*, 4th ed. Upper Saddle River, NJ: Prentice Hall.
- Hull, J., and White, A. (1987). The pricing of options on assets with stochastic volatilities. *Journal of Finance* 42: 281–300.
- Javaheri, A. (2005). *Inside Volatility Arbitrage*. New York: Wiley/Finance.
- Johnson, H., and Shanno, D. (1985) Option pricing when the variance is changing. Working paper 85-07. University of California, Davis: Graduate School of Administration.
- Kazmerchuk, Y., Swishchuk, A., and Wu, J. (2005). A continuous-time GARCH model for stochastic volatility with delay. *Canadian Applied Mathematics Quarterly* 13, 2: 123–149.
- Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business* 36: 394–419.
- Merton, R. (1973). Theory of rational option pricing. *Bell Journal of Economic Management Science* 4: 141–183.
- Molina, G., Han, C.-H., and Fouque, J.-P. (2003). MCMC estimation of multiscale stochastic volatility models. Preprint. Santa Barbara, CA: University of California.
- Ornstein, L., and Uhlenbeck, G. (1930). On the theory of Brownian motion. *Physical Review* 36: 823–841.
- Scott, L. (1987). Option pricing when the variance changes randomly: Theory, estimation and an application. *Journal of Financial Quantitative Analysis* 22: 419–438.
- Shephard, N. (2005). *Stochastic Volatility: Selected Readings*. Oxford: Oxford University Press.
- Shiryayev, A. (2008). *Essentials of Stochastic Finance: Facts, Models, Theory*. Singapore: World Scientific.
- Stein, E., and Stein, J. (1991). Stock price distribution with stochastic volatility. An analytic approach. *Review of Financial Studies* 4: 727–752.
- Swishchuk, A. (2000). *Random Evolutions and Their Applications*. New Trends. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Swishchuk, A. (2004). Modelling and valuing of variance and volatility swaps for financial markets with stochastic volatilities. *Wilmott Magazine* 2, September: 64–72.
- Swishchuk, A. (2005). Modeling and pricing of variance swaps for stochastic volatilities with delay. *Wilmott Magazine* 19, September: 63–73.
- Swishchuk, A. (2006). Modeling and pricing of variance swaps for multi-factor stochastic

- volatilities with delay. *Canadian Applied Mathematics Quarterly* 14, 4, Winter.
- Swishchuk, A. (2009a). Pricing of variance and volatility swaps with semi-Markov volatilities. *Canadian Applied Mathematics Quarterly* 18, 4.
- Swishchuk, A. (2009b). Variance swaps for local stochastic volatility with delay and jumps. Working Paper. Calgary: University of Calgary.
- Swishchuk, A., and Couch, M. (2010). Volatility and variance swaps for COGARCH(1,1) model. *Wilmott Journal* 2, 5: 231–246.
- Swishchuk, A., and Li, X. (2011). Variance swaps for stochastic volatility with delay and jumps. *International Journal of Stochastic Analysis* 2010: 1–27.
- Swishchuk, A., and Malenfant, K. (2010). Pricing of variance swaps for Levy-based stochastic volatility with delay. *International Review of Applied Financial Issues and Economics*. Paris: S.E.I.F. (accepted).
- Swishchuk, A., and Manca, R. (2010). Modeling and pricing of variance swaps for local semi-Markov volatility in financial engineering. In *Mathematical Models in Engineering*. New York: Hindawi, pp. 1–17.
- Wiggins, J. (1987). Option values under stochastic volatility. *Journal of Financial Economics* 19: 351–372.
- Wilmott, P., Howison, S., and Dewynne, J. (1995). *Option Pricing: Mathematical Models and Computations*. Oxford: Oxford Financial Press.

Index

- Absence of arbitrage principle, *I:99, I:127*. *See also* arbitrage, absence of
- ABS/MBS (asset-backed securities/mortgage-backed securities), *I:258–259, I:267*
- cash flow of, *III:4*
- comparisons to Treasury securities, *III:5*
- modeling for, *III:536*
- Accounting, *II:532, II:542–543*
- Accounting firms, watchdog function of, *II:542*
- Accounts receivable turnover ratio, *II:557–558*
- Active-passive decomposition model, *III:17, III:19–22, III:26*
- Activity ratios, *II:557–558, II:563*
- Adapted mesh, one year to maturity, *II:680f*
- Adjustable rate mortgages (ARMs). *See* ARMs (adjustable rate mortgages)
- Adjustments for changes in net working capital (ANWC), *II:25*
- Adverse selection, *III:76*
- Affine models, *III:554–557*
- Affine process, basic, *I:318–319, I:334n*
- Agency ratings, and model risk, *II:728–729*
- Airline stocks, *II:249–250, II:250f, II:250t, II:252t*
- Akaike Information Criterion (AIC), *II:703, II:717*
- Algorithmic trading, *II:117*
- Algorithms, *II:676–677, II:701–702, III:124*
- Allied Products Corp., cash flow of, *II:576*
- α -stable densities, *III:243f, III:244f*
- α -stable distributions
- defined, *II:738*
- discussion of, *III:233–238*
- fitting techniques for, *II:743–744*
- properties of, *II:739*
- simulations for, *II:750*
- subordinated representation of, *II:742–743*
- usefulness of, *III:242*
- and VaR, *II:748*
- variables with, *II:740*
- α -stable process, *III:499*
- Alternative risk measures proposed, *III:356–357*
- Amazon.com
- cash flows of, *II:568, II:568t*
- American International Group (AIG), stock prices of, *III:238*
- Amortization, *II:611, III:72–73*
- Analysis
- and Barra model, *II:244–248*
- bias in, *II:109*
- common-size, *II:561–563*
- crisis-scenario, *III:379–380*
- to determine integration, *II:514*
- formulas for quality, *II:239*
- fundamental, *II:243, II:248, II:253–254*
- interpretation of results, *III:42–44*
- mathematical, *I:18*
- model-generated, *III:41–42*
- multivariate, *II:48*
- statistical, *I:140, II:353–354*
- sum-of-the-parts, *II:43–44*
- vertical *vs.* horizontal common-size, *II:562*
- Analytics, aggregate, *II:269t*
- Anderson, Philip W., *III:275*
- Annual percentage rate (APR), *II:598, II:615–616*
- Annual standard deviation, *vs.* volatility, *III:534*
- Annuities
- balances in deferred, *II:610f*
- from bonds, *I:211–212*
- cash flows in, *II:604–607*
- future value factor, *II:605–606*
- ordinary, *II:605*
- present value factor, *II:605, II:606–607*
- valuation of due, *II:608–609*
- valuing deferred, *II:609–611*
- Anticipation, in stochastic integrals, *III:475*
- Approximation, quality of, *II:330–331*
- APT (arbitrage pricing theory), *I:116*
- Arbitrage
- absence of, *I:56, I:135, II:473*
- in continuous time, *I:121–123*
- convertible bond, *I:230*
- costless profits, *I:442*
- costless trades, *I:428t*
- defined, *I:99, I:119, I:123*
- in discrete-time, continuous state, *I:116–119*
- and equivalent martingale measures, *I:111–112*
- in multiperiod finite-state setting, *I:104–114*
- in one-period setting, *I:100–104*
- pricing of, *I:124, I:134–135, II:476*
- profit from, *I:221–222*
- and relative valuation models, *I:260*
- and state pricing, *I:55–56, I:102, I:130*
- test for costless profit, *I:441*
- trading strategy with, *I:105*
- types of, *I:55–56*
- using, *I:70–71*
- Arbitrage-free, *III:577, III:593–594*
- Arbitrage opportunities, *I:55, I:56, I:100, I:117, I:260–261, I:437*
- Arbitrage pricing theory (APT), *I:116*
- application of, *I:60–61*
- development of, *II:468, II:475–476*
- factors in, *II:138*
- key points on, *II:149–150*
- and portfolio optimization, *I:40*

- ARCH (autoregressive conditional heteroskedastic) models and behavior of errors, *II:362*
 defined, *I:176*
 in forecasting, *II:363*
 reasons for, *III:351*
 type of, *II:131*
 use of, *II:733–734*
- ARCH/GARCH models
 application to VaR, *II:365–366*
 behavior of, *II:361–362*
 discussion of, *II:362–366*
 generalizations of, *II:367–373*
 usefulness of, *II:366–367*
- ARCH/GARCH processes, *III:277*
- Area, approximation of, *II:589–590*, *II:589f*
- ARIMA (autoregressive integrated moving average) process, *II:509–510*
- ARMA (autoregressive moving average) models
 defined, *II:519*
 and Hankel matrices, *II:512*
 linearity of, *II:402*
 and Markov coefficients, *II:512*
 multivariate, *II:510–511*, *II:513–514*
 nonuniqueness of, *II:511*
 representations of, *II:508–512*
 and time properties, *II:733*
 univariate, *II:508–510*
- ARMA (autoregressive moving average) processes, *III:276–277*
- ARMs (adjustable rate mortgages), *III:25*, *III:71–72*, *III:72f*, *III:74*
- Arrays, in MATLAB and VBA, *III:420–421*, *III:457–458*, *III:466*
- Arrow, Kenneth, *II:467*, *II:699*
- Arrow-Debreu price, *I:53–55*. *See also* state price
- Arrow-Debreu securities, *I:458*, *I:463*
- Arthur, Bryan, *II:699*
- Artificial intelligence, *II:715*
- Asian fixed calls, with finite difference methods, *II:670t*
- Asian options, pricing, *III:642–643*
- Asset allocation
 advanced, *I:36*
 building blocks for, *I:38*
 modeling of, *I:42*
 standard approach to, *I:37–38*
- Asset-backed securities (ABS), *I:258*
- Asset-liability management (ALM), *II:303–304*, *III:125–126*
- Asset management, focus of, *I:35*
- Asset prices
 codependence of, *I:92*
 multiplicative model for, *I:86–87*, *I:88*
 negative, *I:84*, *I:88*
 statistical inference of models, *I:560*
- Asset pricing, *I:3*, *I:56–59*, *I:59–60*, *I:65–66*, *II:197*
- Asset return distributions, skewness of, *III:242*
- Asset returns
 characteristics of, *III:392*
 errors in estimation of, *III:140–141*
 generation of correlated, *I:380–381*
 log-normal distribution applied to, *III:223–225*
 models of, *III:381*
 normal distribution of, *I:40*
 real-world, *III:257*
 simulated vector, *I:380–381*
- Assets
 allocation of, *I:10*
 on the balance sheet, *II:533–534*
 carry costs, *I:424–425*
 correlation of company, *I:411*
 current *vs.* noncurrent, *II:533*
 deliverable, *I:483*
 discrete flows of, *I:425–426*
 expressing volatilities of, *III:396–397*
 financing of, *II:548*
 funding cost of, *I:531*
 future value of, *I:426t*, *I:427t*
 highly correlated, *I:192*
 intangible, *II:534*
 liquid, *II:551*
 management of, *II:558*
 market prices of, *I:486*
 new fixed, *II:25*
 prices of, *I:60*
 redundant, *I:51*
 representation of, *II:515*
 risk-free, *I:112–113*
 risky *vs.* risk-free, *I:5–6*
 shipping, *I:555*
 storage of physical, *I:439*, *I:442–443*, *I:560–561*
 values of after default events, *I:350*
- Asset swaps, *I:227–230*
- Assumptions
 about noise, *II:126*
 under CAPM, *I:68–69*
 errors in, *III:399*
 evaluation of, *II:696*
 homoskedasticity *vs.* heteroskedasticity, *II:360*
 importance of, *III:62*
 for linear models, *II:310–311*
 for linear regression models, *II:313*
 in scenario analysis, *II:289*
 simplification of, *III:397*
 using inefficient portfolio analysis, *I:288t*
 violations of, *I:475*
 zero mean return, *III:397*
- Attribution analysis, *II:188–189*
- AT&T stock, binomial experiment, *I:146–148*
- Audits, of financial statements, *II:532*
- Augmented Dickey-Fuller test (ADF), *II:387*, *II:389*, *II:390t*, *II:514*
- Autocorrelation, *II:328–329*, *II:503*, *II:733*
- Autoregressive conditional duration (ACD) model, *II:370*
- Autoregressive conditional heteroskedastic (ARCH) models. *See* ARCH (autoregressive conditional heteroskedastic) models
- Autoregressive integrated moving average (ARIMA) process, *II:509–510*
- Autoregressive models, *II:360–362*
- Autoregressive moving average (ARMA) models. *See* ARMA (autoregressive moving average) models
- AVaR. *See* average value at risk (AVaR)
- Average credit sales per day, calculation of, *II:553*
- Average daily volume (ADV), *II:63*
- Averages, equally weighted, *III:397–409*
- Average value at risk (AVaR) measure
 advantages of, *III:347*
 back-testing of, *III:338–340*
 boxplot of fluctuation of, *III:338f*
 and coherent risk measures, *III:333–334*
 computation of in practice, *III:336–338*
 computing for return distributions, *III:334–335*
 defined, *III:331–335*
 estimation from sample, *III:335–336*
 and ETL, *III:345–347*
 geometrically, *III:333f*
 graph of, *III:347f*
 higher-order, *III:342–343*
 historical method for, *III:336–337*
 hybrid method for, *III:337*
 minimization formula for, *III:343–344*
 Monte Carlo method for, *III:337–338*
 with the multivariate normal assumption, *III:336*
 of order one, *III:342–343*
 for stable distributions, *III:344–345*
 tail probability of, *III:332–333*
- Axiomatic systems, *III:152–153*

- Bachelier, Louis, *II:121–122, II:467, II:469–470, III:241–242, III:495*
- Back propagation (BP), *II:420*
- Back-testing
- binomial (Kupiec) approach, *III:363*
 - conditional testing (Christoffersen), *III:364–365*
 - diagnostic, *III:367–368*
 - example of, *II:748–751*
 - exceedance-based statistical approaches, *III:362–365*
 - in-sample *vs.* out-sample, *II:235–236*
 - need for, *III:361–362*
 - statistical, *III:362*
 - strengths/weaknesses of
 - exceedance-based, *III:365*
 - tests of independence, *III:363–364*
 - trading strategies, *II:236–237*
 - use of, *III:370*
 - using normal approximations, *III:363*
 - of VaRs, *III:365–367*
- Backward induction pricing technique, *III:26*
- Bailouts, *I:417*
- Balance sheets
- common-size, *II:562, II:562f*
 - information in, *II:533–536*
 - sample, *II:534t, II:546t*
 - structure of, *II:536*
 - XYZ, Inc. (example), *II:29t*
- Balls, drawing from urn, *III:174–177, III:175f, III:179–180*
- Bandwidth, *II:413–414, II:746*
- Bank accounts, and volatility, *III:472*
- Bank for International Settlements (BIS), definition of operational risk, *III:82*
- Bankruptcy, *I:350, I:366–369, II:577*
- Banks, use of VaR measures, *III:295*
- Barclays Global Risk Model, *II:173, II:193n, II:268*
- Barra models
- E3, *II:256, II:257t, II:261*
 - equity, *II:245–246*
 - fundamental data in, *II:246t*
 - fundamental factor, *II:244–248, II:248–250*
 - risk, *II:256*
 - use of, *II:254n*
- Barrier options, *II:683*
- Basel II Capital Accord, on operational risk, *III:86–87*
- Basic earning power ratio, *II:547, II:549*
- Bayes, Thomas, *I:140, I:196*
- Bayesian analysis
- empirical, *I:154–155*
 - estimation, *I:189*
 - hypothesis comparison, *I:156–157*
 - in parameter estimation, *II:78*
 - and probability, *I:140, I:148*
 - steps of decision making in, *I:141*
 - testing, *I:156–157*
 - use of, *I:18*
- Bayesian inference, *I:151, I:157–158, II:719*
- Bayesian Information Criterion (BIC), *II:703, II:717*
- Bayesian intervals, *I:156, I:170*
- Bayesian methods, and economic theory, *III:142*
- Bayes' theorem, *I:143–148, I:152*
- Behaviors, patterns of, *II:707–710, III:34–35*
- BEKK(1,1,K) model, *II:372*
- Beliefs
- about long-term volatility, *III:408–409*
 - posterior, *I:151–152*
 - prior, *I:152, I:159*
- Bellman's principle, *II:664–665*
- Benchmarks
- choice of, *II:114–115*
 - effect of taxes on, *II:74*
 - fair market, *III:626*
 - modeling of, *II:696*
 - portfolio, *II:272t*
 - for risk, *II:265, III:350, III:354–355*
 - risk in, *II:259*
 - tracking of, *II:67*
 - for trades, *II:117, III:624*
 - use of, *I:41–42, II:66–69*
- Benchmark spot rate curves, *I:222–223*
- Berkowitz transformation, application of, *III:366–367, III:368*
- Bernoulli model, parameter inference in, *II:726–727*
- Bernoulli trials, *I:81, III:170, III:174*
- Bessel function of the third kind, *III:232*
- Best bids/best asks, *II:449–450*
- Best practices, *I:416*
- Beta function, *III:222*
- Betas
- β_{1963} , *I:74–75*
 - β_{1964} , *I:75*
 - β_{1963} *vs.* β_{1964} , *I:76–77*
 - distribution of, *III:222*
 - meanings of, *I:74*
 - in portfolios, *II:273*
 - pricing model, *I:60–61, I:71–72*
 - propositions about, *I:75–77*
 - robust estimates of, *II:442–443*
 - in SL-CAPM models, *I:66–67*
 - two beta trap, *I:74–77*
- Bets, unintended, *II:261, II:263–264, II:264, II:265*
- Better building blocks, *I:36*
- Bias
- from data, *II:204*
 - discretization error, *III:641*
 - estimator, *III:641*
 - survivorship (look-ahead), *II:202, II:204, II:712–713, II:718*
- Bid-ask bounce, *II:455–457*
- Bid-ask spread
- aspects of, *III:597*
 - average hourly, *II:454f*
 - defined, *II:454*
 - under market conditions, *II:455f*
 - risk in, *III:372*
- Binomial experiment, *I:146–148*
- Black, Fischer, *II:468, II:476*
- Black and Scholes
- assumptions of, *I:510*
- Black-Derman-Toy (BDT) model
- defined, *I:492*
 - discussion of, *III:608–609*
 - features of, *III:549*
 - interest rate model, *III:616f*
 - as no arbitrage model, *III:604*
 - use of, *III:300*
- Black-Karasinski (BK) model, *III:548, III:607–608*
- binomial lattice, *III:611*
 - defined, *I:493*
 - features of, *III:604*
 - forms of, *III:600t*
 - interest rate trinomial lattice, *III:615f*
 - trinomial lattice, *III:616f*
- Black-Litterman model
- assumptions with, *I:196–197*
 - derivation of, *I:196–197*
 - discussion of, *I:195–201*
 - with investor's views and market equilibrium, *I:198–199*
 - mixed estimation procedure, *I:200*
 - use of for forecasting returns, *I:193–194, II:112*
 - use of in parameter estimation, *II:78*
 - variance of, *I:200*
- Black-Scholes formula
- for American options, *II:674*
 - with change of time, *III:522, III:524–525*
 - and diffusion equations, *II:654*
 - and Gaussian distribution, *II:732*
 - and Girsanov's theorem, *I:132–133*
 - statistical concepts for, *III:225*
 - use of, *I:126–127, I:136*
 - use of in MATLAB, *III:423–427, III:447*
 - use of with VBA, *III:462–463*
 - and valuation models, *I:271*
- Black-Scholes-Merton stock option pricing formula, *I:557*

- Black-Scholes model
 assumptions of, *I:512, III:655*
 and calibration, *II:681–682*
 for European options, *II:660–662, III:639–640*
 and hedging, *I:410*
 and Merton's model, *I:343*
 for pricing options, *I:487, I:509–510, I:522*
 usefulness of, *I:475*
 use of, *I:272*
 volatility in, *III:653*
- Black volatility, *III:548, III:550*
- Bohr, Niels, *I:123*
- Bond-price valuation model, *III:581–583*
- Bonds
 analytical models for, *I:271–273*
 annuities from, *I:211–212*
 calculating yields on, *II:618*
 callable, *I:24f, I:244–245, III:302–303, III:302f*
 capped floating rate, valuation of, *I:249f*
 changes in prices, *I:373–374*
 computing accrued interest and clean price of, *I:214–215*
 convertible, *I:230, I:271*
 corporate, *I:279, III:598–599*
 coupon-paying, *III:584–586*
 default-free, *I:223*
 determination of value of, *I:211–213*
 discount, *I:212*
 effective duration/convexity of, *I:255, I:256f*
 European convertible, *I:272*
 in European-style calls, *I:440*
 floating-coupon, *I:246–248, I:247f*
 floating-rate callable capped, *I:248*
 floating valuation, *I:253f*
 full (dirty) price, *I:214, I:370*
 futures contracts on, *I:498*
 general principles of valuation, *I:209–216*
 inflation-indexed, *I:278, I:279, I:283–290, I:290–294*
 input information for example, *III:613t*
 interest rate tree for, *I:244f*
 loading of specific, *II:279*
 modeling prices of, *I:490–494*
 and modified or effective duration, *III:299*
 nonpar, *I:232n*
 option-free, *I:241f, I:243*
 options on, *I:252–253, I:498–501, I:501–502*
 planned amortization class (PAC), *III:6*
 plot of convertible functions, *I:273f*
 prediction of yield spreads, *II:336–344*
 price/discount rate relationship, *I:215–216, I:215f*
 prices of, *I:213–214, I:278, I:382, II:727–728*
 prices with effective duration/convexity, *III:300t, III:301t*
 pricing for, *I:498–503, III:588*
 putable, effective duration of, *III:303–304, III:304f*
 regression data for spread application, *II:338–343t*
 relation to CDSs, *I:525–526*
 risk-free, *I:316*
 risk-neutral, *III:586*
 risk-neutral/equilibrium models for, *III:597–598*
 security levels of, *I:375t*
 spreads over time, *I:402f*
 straight, duration of, *III:301–302, III:301f*
 time path of, *I:216*
 valuation of, *I:213–215, I:216–223, I:223, II:730, III:576*
 valuing of, *I:213–214, I:244–246, I:246f*
 volatility of, *I:279*
- Book value, of companies, *II:535*
- Bootstrapping
 parametric, *II:428*
 of spot rate curve, *I:217–220*
 technique for, *II:711–712, II:746*
 usefulness of, *III:325*
 use of, *I:223, III:408*
- Borel functions, *III:508–509*
- Borel measures, *III:199, III:498*
- Borrowers, *III:5, III:70–71, III:74–75, III:598–600*
- Borrowing, *I:72–73, I:479–480*
- Boundary conditions, *II:660*
 need for, *II:661*
- Box-and-whiskers diagrams, use of, *III:329–330n*
- Boxplots, use of, *III:329–330n*
- Brennen-Schwartz model, *III:549*
- Brown, Robert, *III:476*
- Brownian motion, geometric (GBM), *I:95, III:656*
- Brownian motion (BM)
 arithmetic, *I:125, III:492, III:503*
 in binomial models, *I:114*
 bounds of, *III:473–474*
 canonical, *III:478*
 conditions defining, *III:483n*
 defined, *I:95, III:476–479*
 with drift, *III:491*
 early work on, *II:470*
 excursion of, *III:480*
 fractal properties of, *III:479–480, III:480f*
 generated by random walk, *III:479f*
 generating paths for in VBA, *III:463–465*
 geometric, *III:492–493, III:503, III:524*
 and Girsanov's theorem, *I:131, I:263*
 in Ito processes, *III:487–488*
 in Ito's formula, *III:488–489*
 and the Merton model, *I:306*
 one-dimensional standard, *III:477–478*
 paths of, *III:486, III:501–502, III:502f*
 path with deviation zones, *III:537f*
 process of, *I:269–270n*
 properties of, *III:479–481, III:501, III:536*
 in randomness calculations, *III:534–535*
 and stochastic integrals, *III:473*
 time-changed, *III:503–505*
 usefulness of, *III:495–496*
 use of, *I:262*
 variants of, *III:506*
- Bubbles, discovering, *II:396–399*
- Burmeister-Ibbotson-Roll-Ross (BIRR) model, *II:140*
- Burnout effect, *III:17–18, III:24, III:74*
- Burnout factor
 initializing of, *III:22*
- Business cycles, *I:351–352, I:408, II:430–431, II:432–433*
- Businesses, correlation within sectors, *I:411*
- Butterfly's wings, effect of, *II:645*
- Calculus, stochastic, *I:94–97*
- Calendarization, *II:43, II:487–488*
- Calibration
 of derivatives, *I:494*
 effect of, *III:619*
 under GIG model, *II:524*
 of local volatility, *II:681–685*
 need for, *III:604*
 to short forward curve, *III:545–546*
- Callable bonds, *I:462*
- Call options
 defined, *I:439*
 discrepancy measures across maturities of, *II:525t*
 early exercise of American-style, *I:442–443, I:449–450*
 European, *I:125, I:127–129, I:501, I:511, II:522–525, II:679*
 1998 prices of, *II:524t*
 value of, *I:259*

- Calls
 American-style, *I:441–442, I:449*
 error on value of, *II:668f*
 European-style, *I:440–441, I:448–449, I:448t*
- Canonical correlation analysis, *I:556*
- Capital asset pricing model (CAPM).
See CAPM (capital asset pricing model)
- Capital expenditures coverage ratio,
II:575–576
- Capital gains, taxes on, *II:73*
- Caplets, *I:249, III:589–590*
- CAPM
 multifactor, *II:475*
- CAPM (capital asset pricing model).
See also Roy CAPM; SL-CAPM
 application of, *I:60–61*
 areas of confusion, *I:67–68*
 for assessing operational risk,
III:92–93
 in asset pricing, *II:474*
 defined, *I:394*
 and discount factor model, *I:65–66*
 and investor risk, *I:73–74*
 using assumptions under, *I:68–69*
- Caps
 defined, *I:248–251*
 value of, *I:248, III:552–553*
 valuing of, with floors, *I:249–250, I:256*
- Carry, *I:423–426, I:481*
- Carry costs, *I:424–425, I:426, I:435, I:437–438, I:455n, I:481. See also* net cost of carry
- CART (classification and regression trees)
 defined, *II:375*
 example, input variables for, *II:379t*
 example, out-of-sample performance, *II:381t*
 fundamentals of, *II:376–377*
 in stock selection, *II:378–381*
 strengths and weaknesses of,
II:377–378
 uses of, *II:381*
- Cash-and-carry trade, *I:480, I:481, I:487*
- Cash concept, *II:567*
- Cash flows
 accounting for, *III:306*
 analysis of, *II:574–577, III:4–5*
 for bond class, *III:9t*
 of bonds, *I:211*
 cash flow at risk (CFaR), *III:376–378*
 classification of, *II:567*
 defined, *I:209–210, II:539, III:4*
 direct *vs.* indirect reporting method,
II:567
 discounted, *I:225*
 discrete, *I:429*
 distribution analysis *vs.* benchmark,
III:310
 estimation of, *I:209–210, II:21–23*
 expected, *I:211*
 factors in, *III:31–32, III:377*
 form residential mortgage loans,
III:62
 futures *vs.* forwards, *I:431t*
 future value of, *II:603f*
 influences on, *III:44*
 interest coverage ratio of, *II:561, II:575*
 interim, *I:482*
 for loan pool, *III:9t*
 measurement of, *II:565–566, III:14*
 monthly, *III:52–54, III:53t*
 net free (NFCF), *II:572–574, II:578*
 in OAS analysis, *I:259*
 perpetual stream of, *II:607–608*
 sources of, *II:540–541, II:569t*
 in state dependent models,
I:351–352
 statement of, *II:539–541, II:566–567*
 time patterns of, *II:607–611*
 and time value of money, *II:595–596*
 time value of series of, *II:602–607*
 for total return receivers, *I:542*
 for Treasuries, *I:219, III:564–565*
 types of in assessing liquidity risk,
III:378
 use of information on, *II:576–577*
 valuation of, *II:618–619*
vs. free cash flow, *II:22–23*
- Cash flow statements
 example of, *II:541*
 form of, *II:26t*
 information from, *II:577–578*
 reformatting of, *II:569t*
 restructuring of, *II:568*
 sample, *II:547t*
 use of, *II:24–26*
- Cash flow-to-debt ratio, *II:576*
- Cash-out refinancing, *III:66, III:69*
- Cash payments, *I:486–487, III:377*
- Categorizations, determining usefulness of, *II:335*
- Cauchy, Augustin, *II:655*
- Cauchy initial value problem, *II:655, II:656, II:656f, II:657*
- CAViAR (conditional autoregressive value at risk), *II:366*
- CDOs (collateralized debt obligations), *I:299, I:525, III:553, III:645*
- CDRs (conditional default rates)
 in cash flow calculators, *III:34*
 defaults measured by, *III:58–59*
 defined, *III:30–31*
 monthly, *III:62t*
 projections for, *III:35f*
 in transition matrices, *III:35f*
- CDSs (credit default swaps)
 basis, *I:232*
 bids on, *I:527*
 cash basis, *I:402*
 discussion of, *I:230–232*
 fixed premiums of, *I:530–531*
 hedging with, *I:418*
 illustration of, *I:527*
 initial value of, *I:538*
 maturity dates, *I:526*
 payoff and payment structure of,
I:534f
 premium payments, *I:231f, I:533–535*
 pricing models for, *I:538–539*
 pricing of by static replication,
I:530–532
 pricing of single-name, *I:532–538*
 quotations for, *I:413*
 risk and sensitivities of, *I:536–537*
 spread of, *I:526*
 unwinding of, *I:538*
 use of, *I:403, I:413, II:284*
 valuation of, *I:535–536*
 volume of market, *I:414*
- Central limit theorem
 defined, *I:149n, III:209–210, III:640*
 and the law of large numbers,
III:263–264
 and random number generation,
III:646
 and random variables, *II:732–733*
- Central tendencies, *II:353, II:354, II:355*
- Certainty equivalents, *II:723–724, II:724–725*
- CEV (constant elasticity of variance),
III:550, III:551f, III:654–655
- Chambers-Mallows-Stuck generator,
II:743–744
- Change of measures, *III:509–517, III:516t*
- Change of time methods (CTM)
 applications of, *III:522–527*
 discussion of, *III:519–522*
 general theory of, *III:520–521*
 main idea of, *III:519–520, III:527*
 in martingale settings, *III:522–523*
 in stochastic differential equation setting, *III:523*
- Chaos, defined, *II:653*
- Chaos: *Making a New Science* (Gleick),
II:714
- Characteristic function
vs. probability density function,
II:743

- Characteristic lines, *II:316, II:318t, II:344–348, II:345–347t*
- Chebyshev inequalities, *III:210, III:225*
- Chen model, *I:493*
- Chi-square distributions, *I:388–389, III:212–213*
- Cholesky factor, *I:380*
- Chow test, *II:336, II:343, II:344, II:350*
- CID (conditionally independent defaults) models, *I:320, I:321–322, I:333*
- CIR model, *I:498, I:500–501, I:502*
- Citigroup, *I:302, I:408f, I:409f*
- CLA (critical line algorithm), *I:73*
- Classes
criteria for, *II:494*
- Classical tempered stable (CTS) distribution, *II:741–742, II:741f, II:742f, II:743–744, III:512*
- Classification, and Bayes' Theorem, *I:145*
- Classification and regression trees (CART). *See* CART (classification and regression trees)
- Classing, procedure for, *II:494–498*
- Clearinghouses, *I:478*
- CME Group, *I:489–490*
- CMOs (collateralized mortgage obligations), *III:598, III:645*
- Coconut markets, *I:70*
- Coefficients
binomial, *III:171, III:187–191*
of determination, *II:315*
estimated, *II:336–337*
- Coherent risk measures, *III:327–329* and VaR, *III:329*
- Coins, fair/unfair, *III:169, III:326–327*
- Cointegrated models, *II:503*
- Cointegration
analysis of, *II:381t*
defined, *II:383*
empirical illustration of, *II:388–393*
technique of, *II:384–385*
testing for, *II:386–387*
test of, *II:394t, II:396t*
use of, *II:397*
- Collateralized debt obligations (CDOs), *I:299, I:525, III:553, III:645*
- Collateralized mortgage obligations (CMOs), *III:598, III:645*
- Collinearity, *II:329–330*
- Commodities, *I:279, I:556, I:566*
- Companies. *See* firms
- Comparison principals, *II:676*
- Comparisons vs. testing, *I:156*
- Complete markets, *I:103–104, I:119, I:133, I:461*
- Complexity, profiting from, *II:57–58*
- Complexity (Waldrop), *II:699*
- Complex numbers, *II:591–592, II:592f*
- Compounding. *See also* interest and annual percentage rates, *II:616*
continuous, *II:599, II:617*
determining number of periods, *II:602*
discrete vs. continuous, *III:570–571*
formula for growth rate, *II:8*
more than once per year, *II:598–599*
and present value, *II:618*
- Comprehensive Capital Analysis and Review, *I:300*
- Comprehensive Capital Assessment Review, *I:412*
- Computational burden, *III:643–644*
- Computers. *See also* various software applications
increased use of, *III:137–138*
introduction of into finance, *II:480*
modeling with, *I:511, II:695*
random walk generation of, *II:708*
in stochastic programming, *III:124, III:125–126*
- Concordance, defined, *I:327*
- Conditional autoregressive value at risk (CAViaR), *II:366*
- Conditional default rate (CDR). *See* CDRs (conditional default rates)
- Conditionally independent defaults (CID) models, *I:320, I:321–322, I:323*
- Conditioning/conditions, *I:24, II:307–308, II:361, II:645*
- Confidence, *I:200, I:201, II:723, III:319*
- Confidence intervals, *II:440, III:338t, III:399–400, III:400f*
- Conglomerate discounts, *II:43*
- Conseco, debt restructure of, *I:529*
- Consistency, notion of, *II:666–667*
- Constant elasticity of variance (CEV), *III:550, III:551f, III:654–655*
- Constant growth dividend discount model, *II:7–9*
- Constraints, portfolio
cardinality, *II:64–65*
common, *III:146*
commonly used, *II:62–66, II:84*
holding, *II:62–63*
minimum holding/transaction size, *II:65*
nonnegativity, *I:73*
real world, *II:224–225*
round lot, *II:65–66*
setting, *I:192*
turnover, *II:63*
on weights of, *I:191–192*
- Constraint sets, *I:21, I:28, I:29*
- Consumer Price Index (CPI), *I:277–278, I:291f, I:292, I:292f*
- Consumption, *I:59–60, II:360, III:570*
- Contagion, *I:320, I:324, I:333*
- Contingent claims
financial instruments as, *I:462*
incomplete markets for, *I:461–462*
unit, *I:458*
use of analysis, *I:463*
utility maximization in markets, *I:459–461*
value of, *I:458–459*
- Continuity, formal treatment of, *II:583–584*
- Continuous distribution function (c.d.f.), *III:167, III:196, III:205, III:345–346, III:345f*
- Continuous distribution function F(a), *III:196*
- Continuous time/continuous state, *III:578*
- Continuous-time processes, change of measure for, *III:511–512*
- Control flow statements in VBA, *III:458–460*
- Control methods, stochastic, *I:560*
- Convenience yields, *I:424, I:439*
- Convergence analysis, *II:667–668*
- Conversion, *I:274, I:445*
- Convexity
in callable bonds, *III:302–303*
defined, *I:258–259, III:309*
effective, *III:13, III:300–304, III:617t*
measurement of, *III:13–14, III:304–305*
negative, *III:14, III:49, III:303*
positive, *III:13*
use of, *III:299–300*
- Convex programming, *I:29, I:31–32*
- Cootner, Paul, *III:242*
- Copulas
advantages of, *III:284*
defined, *III:283*
mathematics of, *III:284–286*
usefulness of, *III:287*
visualization of bivariate independence, *III:285f*
visualization of Gaussian, *III:287f*
- Corner solutions, *I:200*
- Correlation coefficients
relation to R^2 , *II:316*
and Theil-Sen regression, *II:444*
use of, *III:286–287*
- Correlation matrices, *II:160t, II:163t, III:396–397*
- Correlations
in binomial distribution, *I:118*
computation of, *I:92–93*

- concept of, *III:283*
- drawbacks of, *III:283–284*
- between periodic increments, *III:540t*
- and portfolio risk, *I:11*
- robust estimates of, *II:443–446*
- serial, *II:220*
- undesirable, *I:293*
- use of, *II:271*
- Costs, net financing, *I:481*
- Cotton prices, model of, *III:383*
- Countable additivity, *III:158*
- Counterparts, robust, *II:81*
- Countries, low- vs. high inflation, *I:290*
- Coupon payments, *I:212, III:4*
- Coupon rates, computing of, *III:548–549*
- Courant-Friedrichs-Lewy (CFL) conditions, *II:657*
- Covariance
 - calculation of between assets, *I:8–9*
 - estimators for, *I:38–40, I:194–195*
 - matrix, *I:38–39, I:155, I:190*
 - relationship with correlation, *I:9*
 - reliability of sample estimates, *II:77*
 - use of, *II:370–371*
- Covariance matrices
 - decisions for interest rates, *III:406*
 - eigenvectors/eigenvalues, *II:160t*
 - equally weighted moving average, *III:402–403*
 - frequency of observations for, *III:404*
 - graphic of, *II:161t*
 - residuals of return process of, *II:162t*
 - of RiskMetrics™ Group, *III:412–413*
 - statistical methodology for, *III:398–399*
 - of ten stock returns, *II:159t*
 - use of, *II:158–159, II:169*
 - using EWMA in, *III:411*
- Coverage ratios, *II:560–561*
- Cox-Ingersoll-Ross (CIR) model, *I:260, I:491–492, I:547, I:548, III:546–547, III:656*
- Cox processes, *I:315–316, II:470–471*
- Cox-Ross-Rubenstein model, *I:510, I:522, II:678*
- CPI (Consumer Price Index), *I:277–278, I:291f, I:292, I:292f*
- CPRs (conditional prepayment rates). *See* prepayment, conditional
- CPR vector, *III:74. See also* prepayment, conditional
- Cramer, Harald, *II:470–471*
- Crank-Nicolson schemes, *II:666, II:669, II:674, II:680*
- Crank Nicolson-splitting (CN-S) schemes, *II:675*
- Crashmetrics, use of, *III:379, III:380*
- Credible intervals, *I:156*
- Credit-adjusted spread trees, *I:274*
- Credit crises
 - of 2007, *III:74*
 - of 2008, *III:381*
 - data from and DTS model, *I:396*
 - in Japan, *I:417*
- Credit curing, *III:73*
- Credit default swaps (CDSs). *See* CDSs (credit default swaps)
- Credit events
 - and credit loss, *I:379*
 - in default swaps, *I:526, I:528–530*
 - definitions of, *I:528*
 - descriptions of most used, *I:528t*
 - exchanges/payments in, *I:231f*
 - in MBS turnover, *III:66*
 - prepayments from, *III:49–50*
 - protection against, *I:230*
 - and simultaneous defaults, *I:323*
- Credit hedging, *I:405*
- Credit inputs, interaction of, *III:36–38*
- Credit loss
 - computation of, *I:382–383*
 - distribution of, *I:369f*
 - example of distribution of, *I:386f*
 - simulated, *I:389*
 - steps for simulation of, *I:379–380*
- Credit models, *I:300, I:302, I:303*
- Credit performance, evolution of, *III:32–36*
- Credit ratings
 - categories of, *I:362*
 - consumer, *I:302*
 - disadvantages of, *I:300–301*
 - implied, *I:381–382*
 - maturity of, *I:301*
 - reasons for, *I:300*
 - risks for, *II:280–281, II:280t*
 - use of, *I:309*
- Credit risk
 - common, *I:322*
 - counterparty, *I:413*
 - in credit default swaps, *I:535*
 - defined, *I:361*
 - distribution of, *I:377*
 - importance of, *III:81*
 - measures for, *I:386f*
 - modeling, *I:299–300, I:322, III:183*
 - quantification of, *I:369–372*
 - reports on, *II:278–281*
 - shipping, *I:566*
 - and spread duration, *I:391–392*
 - vs. cash flow risk, *III:377–378*
- Credit scores, *I:300–302, I:301–302, I:309, I:310n*
- Credit spreads
 - alternative models of, *I:405–406*
 - analysis with stock prices, *I:305t*
 - applications of, *I:404–405*
 - decomposition, *I:401–402*
 - drivers of, *I:402*
 - interpretation of, *I:403–404*
 - model specification, *I:403*
 - relationship with stock prices, *I:304*
 - risk in, *II:279t*
 - use of, *I:222–223*
- Credit support, evaluation of, *III:39–40*
- Credit value at risk (CVaR). *See* CVaR
- Crisis situations, estimating liquidity in, *III:378–380*
- Critical line algorithm (CLA), *I:73*
- Cross-trading, *II:85n*
- Cross-validation, leave-one-out, *II:413–414*
- Crude oil, *I:561, I:562*
- Cumulation, defined, *III:471*
- Cumulative default rate (CDX), *III:58*
- Cumulative frequency distributions, *II:493f, II:493t, II:498–499*
- formal presentation of, *II:492–493*
- Currency put options, *I:515*
- Current ratio, *II:554*
- Curve imbalances, *II:270–271*
- Curve options, *III:553*
- Curve risk, *II:275–278*
- CUSIPs/ticker symbols, changes in, *II:202–203*
- CVaR (credit value at risk), *I:384–385, I:385–386, II:68, II:85n, III:392t. See also* value at risk (VaR)
- Daily increments of volatility, *III:534*
- Daily log returns, *II:407–408*
- Dark pools, *II:450, II:454*
- Data. *See also* operational loss data
 - absolute, *II:487–488*
 - acquisition and processing of, *II:198*
 - alignment of, *II:202–203*
 - amount of, *I:196*
 - augmentation of, *I:186n*
 - availability of, *II:202, II:486*
 - backfilling of, *II:202*
 - bias of, *II:204, II:713*
 - bid-ask aggregation techniques for, *II:457f*
 - classification of, *II:499–500*
 - collection of, *II:102, II:103f*
 - cross-sectional, *II:201, II:488, II:488f*
 - in forecasting models, *II:230*
 - frequency of, *II:113, II:368, II:462–463, II:500*
 - fundamental, *II:246–247*
 - generation of, *II:295–296*

- Data (*Continued*)
- high-frequency (HFD) (*See* high-frequency data (HFD))
 - historical, *II:77–78, II:122, II:172*
 - housing bubble, *II:397–399*
 - importing into MATLAB, *III:433–434*
 - industry-specific, *II:105*
 - integrity of, *II:201–203*
 - levels and scale of, *II:486–487*
 - long-term, *III:389–390*
 - in mean-variance, *I:193–194*
 - misuse of, *II:108*
 - on operational loss, *III:99*
 - from OTC business, *II:486*
 - patterns in, *II:707–708*
 - pooling of, *III:96*
 - of precision, *I:158*
 - preliminary analysis of, *III:362*
 - problems in for operational risk, *III:97–98*
 - qualitative *vs.* quantitative, *II:486*
 - quality of, *II:204, II:211, II:452–453, II:486, II:695*
 - reasons for classification of, *II:493–494*
 - for relative valuation, *II:34–35*
 - restatements of, *II:202*
 - sampling of, *II:459f, II:711*
 - scarcity of, *II:699–700, II:703–704, II:718*
 - sorting and counting of, *II:488–491*
 - standardization of, *II:204, III:228*
 - structure/sample size of, *II:703*
 - types of, *II:486–488*
 - underlying signals, *II:111*
 - univariate, defined, *II:485*
 - working with, *II:201–206*
- Databases
- Compustat Point-In-Time, *II:238*
 - Factiva, *II:482*
 - Institutional Brokers Estimate System (IBES), *II:238*
 - structured, *II:482*
 - third-party, *II:198, II:211n*
- Data classes, criteria for, *II:500*
- Data generating processes (DGPs), *II:295–296, II:298f, II:502, II:702, III:278*
- Data periods, length of, *III:404*
- Data series, effect of large number of, *II:708–709*
- Data sets, training/test, *II:710–711*
- Data snooping, *II:700, II:710–712, II:714, II:717, II:718*
- Datini, Francesco, *II:479–480*
- Davis-Lo infectious defaults model, *I:324*
- Days payables outstanding (DPO), calculation of, *II:553–554*
- Days sales outstanding (DSO), calculation of, *II:553*
- DCF (discounted cash flow) models, *II:16, II:44–45*
- DDM (dividend discount models). *See* dividend discount models (DDM)
- Debt
- long-term, in financial statements, *II:542*
 - models of risky, *I:304–307*
 - restructuring of, *I:230*
 - risky, *I:307–308*
- Debt-to-assets ratio, *II:559*
- Debt-to-equity ratio, *II:559*
- Decomposition models
- active/passive, *III:19*
- Default correlation, *I:317–318*
- contagion, *I:353–354*
 - cyclical, *I:352, I:353*
 - linear, *I:320–321*
 - measures of, *I:320–321*
 - tools for modeling, *I:319–333*
- Default intensity, *III:225*
- Default models, *I:321–322, I:370f*
- Default probabilities
- adjustments in real time, *I:300–301*
 - between companies, *I:412–413*
 - cyclical rise and fall, *I:408f, I:409f*
 - defined, *I:299–300*
 - effect of business cycle on, *I:408*
 - effect of rating outlooks on, *I:365–366*
 - empirical approach to, *I:362–363*
 - five-year (Bank of America and Citigroup), *I:301f, I:302f*
 - merits of approaches to, *I:365*
 - Merton's approach to, *I:363–365*
 - probability of, *II:727, II:727f, II:728f*
 - and survival, *I:533–535*
 - and survival probability, *I:323–324*
 - term structure of, *I:303*
 - time span of, *I:302–303*
 - vs.* ratings and credit scores, *I:300–302*
 - for Washington Mutual, *I:415f, I:416f*
 - of Washington Mutual, *I:415f, I:416f*
- Defaults
- annual rates of, *I:363*
 - and Bernoulli distributions, *III:169–170*
 - calculation of monthly, *III:61t*
 - clustering of, *I:324–325*
 - contagion, *I:320*
 - copulas for times, *I:329–331*
 - correlation of between companies, *I:411*
 - cost of, *I:401, I:404f*
 - dollar amounts of, *III:59f*
 - effect of, *I:228, III:645*
 - event *vs.* liquidation, *I:349*
 - factors influencing, *III:74–75*
 - first passage model of, *I:349*
 - historical database of, *I:414*
 - intensity of, *I:330, I:414*
 - looping, *I:324–325*
 - measures of, *III:58–59*
 - in Merton approach, *I:306*
 - Moody's definition of, *I:363*
 - predictability of, *I:346–347*
 - and prepayments, *III:49–50, III:76–77*
 - process, relationship to recovery rate, *I:372*
 - pseudo intensities, *I:330*
 - rates of cumulative/conditional, *III:63*
 - recovery after, *I:316–317*
 - risk of, *I:210*
 - simulation of times, *I:322–324, I:325*
 - threshold of, *I:345–346*
 - times simulation of, *I:319*
 - triggers for, *I:347–348*
 - variables in, *I:307–308*
- Default swaps
- assumptions about, *I:531–532*
 - and credit events, *I:530*
 - digital, *I:537*
 - discussion of, *I:526–528*
 - market relationship with cash market, *I:530*
 - and restructuring, *I:528–529*
 - value of spread, *I:534*
- Default times, *I:332*
- Definite covariance matrix, *II:445*
- Deflators, *I:129, I:136*
- Degrees, in ordinary differential equations, *II:644–645*
- Degrees of freedom (DOF)
- across assets and time, *II:735–736*
 - in chi-square distribution, *III:212*
 - defined, *II:734*
 - for Dow Jones Industrial Average (DJIA), *II:735–737, II:737f*
 - prior distribution for, *I:177*
 - range of, *I:187n*
 - for S&P 500 index stock returns, *II:735–736, II:736f*
- Delinquency measures, *III:57–58*
- Delivery date, *I:478*
- Delta, *I:509, I:516–518, I:521*
- Delta-gamma approximation, *I:519, III:644–645*
- Delta hedging, *I:413, I:416, I:418, I:517*

- Delta profile, *I:518f*
- Densities
- beta, *III:108f*
 - Burr, *III:110f*
 - closed-form solutions for, *III:243*
 - exponential, *III:105–106, III:105f*
 - gamma, *III:108f*
 - Pareto, *III:109f*
 - posterior, *I:170f*
 - two-point lognormal, *III:111f*
- Density curves, *I:147f*
- Density functions
- asymmetric, *III:205f*
 - of beta distribution, *III:222f*
 - chi-square distributions, *III:213f*
 - common means, different variances, *III:203f*
 - computing probabilities from, *III:201*
 - discussion of, *III:197–200*
 - of *F*-distribution, *III:217f*
 - histogram of, *III:198f*
 - of log-normal distribution, *III:223f*
 - and normal distribution, *II:733*
 - and probability, *III:206*
 - rectangular distributions, *III:220*
 - requirements of, *III:198–200*
 - symmetric, *III:204f*
 - of *t*-distribution, *III:214f*
- Dependence, *I:326–327, II:305–308*
- Depreciation, *II:22*
- accumulated, *II:533–534*
 - expense *vs.* book value, *II:539f*
 - expense *vs.* carrying value, *II:540f*
 - in financial statements, *II:537–539*
 - on income statements, *II:536*
 - methods of allocation, *II:537–538*
- Derivatives
- construction of, *II:586–587*
 - described, *II:585–586*
 - embedded, *I:462*
 - energy, *I:558*
 - exotic, *I:558, I:559–560*
 - of functions, defined, *II:593*
 - and incomplete markets, *I:462*
 - interest rate, *III:589–590*
 - nonlinearity of, *III:644–645*
 - OTC, *I:538*
 - pricing of, *I:58, III:594–596*
 - pricing of financial, *III:642–643*
 - relationship with integrals, *II:590*
 - for shipping assets, *I:555, I:558, I:565–566*
 - use of instruments, *I:477*
 - valuation and hedging of, *I:558–560*
 - vanilla, *I:559*
- Derman, Emanuel, *II:694*
- Descriptors, *II:140, II:246–247, II:256*
- Determinants, *II:623*
- Deterministic methods
- usefulness of, *II:685*
- Diagonal VEC model (DVEC), *II:372*
- Dice, and probability, *III:152, III:153, III:155–156, III:156t*
- Dickey-Fuller statistic, *II:386–387*
- Dickey-Fuller tests, *II:514*
- Difference, notation of, *I:80*
- Differential equations
- classification of, *II:657–658*
 - defined, *I:95, II:644, II:657*
 - first-order system of, *II:646*
 - general solutions to, *II:645*
 - linear, *II:647–648*
 - linear ordinary, *II:644–645*
 - partial (PDE), *II:643, II:654–657*
 - stochastic, *II:643–644*
 - systems of ordinary, *II:645–646*
 - usefulness of, *II:658*
- Diffusion, *III:539, III:554–555*
- Diffusion invariance principle, *I:132*
- Dimensionality, curse of, *II:673, III:127*
- Dirac measures, *III:271*
- Directional measures, *II:428, II:429*
- Dirichlet boundary conditions, *II:666*
- Dirichlet distribution, *I:181–183, I:186–187n*
- Discounted cash flow (DCF) models, *II:16, II:44–45*
- Discount factors, *I:57–58, I:59–62, I:60, II:600–601*
- Discount function
- calculation of, *III:571*
 - defined, *III:563*
 - discussion of, *III:563–565*
 - forward rates from, *III:566–567*
 - graph of, *III:563f*
 - for on-the-run Treasuries, *III:564–565*
- Discounting, defined, *II:596*
- Discount rates, *I:211, I:212, I:215–216, II:6*
- Discovery heuristics, *II:711*
- Discrepancies, importance of small, *II:696*
- Discrete law, *III:165–169*
- Discrete maximum principle, *II:668*
- Discretization, *I:265, II:669f, II:672*
- Disentangling, *II:51–56*
- complexities of, *II:55–56*
 - predictive power of, *II:54–55*
 - return revelation of, *II:52–54*
 - usefulness of, *II:52, II:58*
- Dispersion measures, *III:352, III:353–354, III:357*
- Dispersion parameters, *III:202–205*
- Distress events, *I:351*
- Distributional measures, *II:428*
- Distribution analysis, cash flow, *III:310*
- Distribution function, *III:218f, III:224f*
- Distributions
- application of hypergeometric, *III:177–178*
 - beliefs about, *I:152–153*
 - Bernoulli, *III:169–170, III:185t*
 - beta, *I:148, III:108*
 - binomial, *I:81f, III:170–174, III:185t, III:363*
 - Burr, *III:109–110*
 - categories for extreme values, *II:752*
 - common loss, *III:112t*
 - commonly used, *III:225*
 - conditional, *III:219*
 - conditional posterior, *I:178–179, I:182–183, I:184–185*
 - conjugate prior, *I:154*
 - continuous probability, *III:195–196*
 - discrete, *III:185t*
 - discrete cumulative, *III:166*
 - discrete uniform, *III:183–184, III:185t, III:638f*
 - empirical, *II:498, III:104–105, III:105f*
 - exponential, *III:105–106*
 - finite-dimensional, *II:502*
 - of Fréchet, Gumbel and Weibull, *III:267f*
 - gamma, *III:107–108, III:221–222*
 - Gaussian, *III:210–212*
 - Gumbel, *III:228, III:230*
 - heavy-tailed, *I:186n, II:733, III:109, III:260*
 - hypergeometric, *III:174–178, III:185t*
 - indicating location of, *III:235*
 - infinitely divisible, *III:253–256, III:253t*
 - informative prior, *I:152–153*
 - inverted Wishart, *I:172*
 - light- *vs.* heavy-tailed, *III:111–112*
 - lognormal, *III:106, III:106f, III:538–539*
 - mixture loss, *III:110–111*
 - for modeling applications, *III:257*
 - multinomial, *III:179–182, III:185t*
 - non-Gaussian, *III:254*
 - noninformative prior, *I:153–154*
 - normal (*See* normal distributions)
 - parametric, *III:201*
 - Poisson, *I:142, III:182–183, III:185t, III:217–218*
 - Poisson probability, *III:187t*
 - posterior, *I:147–148, I:165, I:166–167, I:169–170, I:177, I:183–184*
 - power-law, *III:262–263*
 - predictive, *I:167*
 - prior, *I:177, I:181–182, I:196*
 - proposal, *I:183–184*
 - representation of stable and CTS, *II:742–743*

- Distributions (*Continued*)
 spherical, II:310
 stable, III:238, III:242, III:264–265, III:384 (*See also* α -stable distributions)
 subexponential, III:261–262
 tails of, III:112*f*, III:648
 tempered stable, III:257, III:382
 testing applied to truncated, III:367
- Diversification, II:57–58
 achieving, I:10
 and cap weighting, I:38
 and credit default swaps, I:413–414
 example of, I:15
 international, II:393–396
 Markowitz's work on, II:471
- Diversification effect, III:321
- Diversification indicators, I:192
- Dividend discount models (DDM)
 applied to electric utilities, II:12*t*
 applied to stocks, II:16–17
 basic, II:5
 constant growth, II:7–9, II:17–18
 defined, II:14
 finite life general, II:5–7
 free cash flow model, II:21–23
 intuition behind, II:18–19
 multiphase, II:9–10
 non-constant growth, II:18
 predictive power of, II:54
 in the real world, II:19–20
 stochastic, II:10–12, II:12*t*
- Dividend payout ratio, II:4, II:20
- Dividends
 expected growth in, II:19
 forecasting of, II:6
 measurement of, II:3–4, II:14
 per share, II:3–4
 reasons for not paying, II:27
 required rate of return, II:19
 and stock prices, II:4–5
- Dividend yield, II:4, II:19
- Documentation
 of model risk, II:696, II:697
- Dothan model, I:491, I:493
- Dow Jones Global Titans 500 (DJGTI), II:490*t*, II:491*t*
- Dow Jones Industrial Average (DJIA)
 in comparison of risk models, II:747–751
 components of, II:489*t*
 fitted stable tail index for, II:740*f*
 frequency distribution in, II:489*t*
 performance (January 2004 to June 2011), II:749*f*
 relative frequencies, II:491*t*
 stocks by share price, II:492*t*
- Drawing without replacement, III:174–177
- Drawing with replacement, III:170, III:174, III:179–180
- Drift
 effects of, III:537
 of interest rates, I:263
 in randomness calculations, III:535
 in random walks, I:84, I:86
 time increments of, I:83
 of time series, I:80
 as variable, III:536
- DTS (duration times spread), I:392, I:393–394, I:396–398
- Duffie-Singleton model, I:542–543
- Dupire's formula, II:682–683, II:685
- DuPont system, II:548–551, II:551*f*
- Duration
 calculations of real yield and inflation, I:286
 computing of, I:285
 defined, I:284, III:309
 effective, III:300–304, III:617*t*
 effective/option adjusted, III:13
 empirical, of common stock, II:318–322, II:319–322*t*
 estimation of, II:323*t*
 measurement of, III:12–13, III:304–305
 models of, II:461
 modified *vs.* effective, III:299
- Duration/convexity, effective, I:255, I:256*f*
- Duration times spread (DTS). *See* DTS (duration times spread)
- Durbin-Watson test, III:647
- Dynamical systems
 equilibrium solution of, II:653
 study of, II:651
- Dynamic conditional correlation (DCC) model, II:373
- Dynamic term structures, III:576–577, III:578–579, III:591
- Early exercise, I:447, I:455. *See* calls, American-style; options
- Earnings before interest, taxes, depreciation and amortization (EBITDA), II:566
- Earnings before interest and taxes (EBIT), II:23, II:547, II:556
- Earnings growth factor, II:223
- Earnings per share (EPS), II:20–21, II:38–39, II:537
- Earnings revisions factor, II:207, II:209*f*
- EBITDA/EV factor
 correlations with, II:226
 examples of, II:203, II:203*f*, II:207, II:208*f*
 in models, II:232, II:238–239
 use of, II:222–223
- Econometrics
 financial, II:295, II:298–300, II:301–303
 modeling of, II:373, II:654
- Economic cycles, I:537, II:42–43
- Economic intuition, II:715–716
- Economic laws, changes in, II:700
- Economy
 states of, I:49–50, II:518–519, III:476
 term structures in certain, III:567–568
 time periods of, II:515–516
- Economy as an Evolving Complex System*, *The* (Anderson, Arrow, & Pines), II:699
- Educated guesses, use of, I:511
- EE (explicit Euler) scheme, II:674, II:677–678
- Effective annual rate (EAR), interest, II:616–617
- Efficiency
 in estimation, III:641–642
- Efficient frontier, I:13–14, I:17*f*, I:289*f*
- Efficient market theory, II:396, III:92
- Eggs, rotten, I:457–458
- Eigenvalues, II:627–628, II:705, II:706–707*f*, II:707*t*
- Einstein, Albert, II:470
- Elements, defined, III:153–154
- Embedding problem, and change of time method, III:520
- Emerging markets, transaction costs in, III:628
- EM (expectation maximization) algorithm, II:146, II:165
- Empirical rule, III:210, III:225
- Endogenous parameterization, III:580–581
- Energy
 cargoes of, I:561–562
 commodity price models, I:556–558
 forward curves of, I:564–565
 power plants and refineries, I:563
 storage of, I:560–561, I:563–564
- Engle-Granger cointegration test, II:386–388, II:391–392, II:395
- Entropy, III:354
- EPS (earnings per share), II:20–21, II:38–39, II:537
- Equally weighted moving average, III:400–402, III:406–407, III:408–409
- Equal to earnings before interest and taxes (EBIT), II:23, II:547, II:556
- Equal-variance assumption, I:164, I:167
- Equations
 difference, homogenous *vs.* nonhomogenous, II:638

- difference *vs.* differential, *II:629*
 diffusion, *II:654–656, II:658n*
 error-correction, *II:391, II:395t*
 homogeneous linear difference,
 II:639–642, II:641f
 homogenous difference, *II:630–634,*
 II:631–632f, II:633–634f, II:642
 linear, *II:623–624*
 linear difference, systems of,
 II:637–639
 matrix characteristics of, *II:628*
 no arbitrage, *III:612, III:617–619*
 nonhomogeneous difference,
 II:634–637, II:635f, II:637–638f
 stochastic, *III:478*
- Equilibrium**
 and absolute valuation models,
 I:260
 defined, *II:385–386*
 dimensions of, *III:601*
 in dynamic term structure models,
 III:576
 expectations for, *II:112*
 expected returns from, *II:112*
 modeling of, *III:577, III:594*
 in supply and demand, *III:568*
- Equilibrium models**
 use of, *III:603–604*
- Equilibrium term structure models,**
 III:601
- Equities, I:279**
 investing in, *II:89–90*
- Equity**
 on the balance sheet, *II:535*
 changes in homeowner, *III:73*
 in homes, *III:69*
 as option on assets, *I:304–305*
 shareholders', *II:535*
- Equity markets, II:48**
- Equity multipliers, II:550**
- Equity risk factor models, II:173–178**
- Equivalent probability measures,**
 I:111, III:510–511
- Ergodicity, defined, II:405**
- Erlang distribution, III:221–222**
- Errors. See also estimation error;**
 standard errors
 absolute percentages of, *II:525f,*
 II:526f
 estimates of, *II:676*
 in financial models, *II:719*
 a posteriori estimates, *II:672–673*
 sources of, *II:720*
 terms for, *II:126*
 in variables problem, *II:220*
- Esscher transform, III:511, III:514**
- Estimates/estimation**
 confidence in, *I:199*
 consensus, *II:34–35*
- equations for, *I:348–349*
 in EVT, *III:272–274*
 factor models in, *II:154*
 with GARCH models, *II:364–365*
 in-house from firms, *II:35*
 maximum likelihood, *II:311–313*
 methodology for, *II:174–176*
 and PCA, *II:167f*
 posterior, *I:176*
 posterior point, *I:155–156*
 processes for, *I:193, II:176*
 properties of for EWMA, *III:410–411*
 robust, *I:189*
 techniques of, *II:330*
 use of, *II:304*
- Estimation errors**
 accumulation of, *II:78*
 in the Black-Litterman model, *I:201*
 covariance matrix of, *III:139–140*
 effect of, *I:18*
 pessimism in, *III:143*
 in portfolio optimization, *II:82,*
 III:138–139
 sensitivity to, *I:191*
 and uncertainty sets, *III:141*
- Estimation risk, I:193**
 minimizing, *III:145*
- Estimators**
 bias in, *III:641*
 efficiency in, *III:641–642*
 equally weighted average,
 III:400–402
 factor-based, *I:39*
 terms used to describe, *II:314*
 unbiased, *III:399*
 variance, *II:313*
- ETL (expected tail loss), III:355–356**
- Euler approximation, II:649–650,**
 II:649f, II:650f
- Euler constant, III:182**
- Euler schemes, explicit/implicit, II:666**
- Europe**
 common currency for, *II:393*
 risk factors of, *II:174*
- European call options**
 Black-Scholes formula for,
 III:639–640
 computed by different methods,
 III:650–651, III:651f
 explicit option pricing formula,
 III:526–527
 pricing by simulation in VBA,
 III:465–466
 pricing in Black-Scholes setting,
 III:649
 simulation of pricing, *III:444–445,*
 III:462–463
 and term structure models,
 III:544–545
- European Central Bank, I:300**
- Events**
 defined, *III:85, III:162, III:508*
 effects of macroeconomic, *II:243–244*
 extreme, *III:245–246, III:260–261,*
 III:407
 identification of, *II:516*
 mutually exclusive, *III:158*
 in probability, *III:156*
 rare, *III:645*
 rare *vs.* normal, *I:262*
 tail, *III:88n, III:111, III:118*
 three- δ , *III:381–382*
- EVT (extreme value theory). See**
 extreme value theory (EVT)
- EWMA (exponentially weighted**
 moving averages), III:409–413
- Exceedance observations, III:362–363**
- Exceedances, of VaR, III:325–326,**
 III:339
- Excel**
 accessing VBA in, *III:477*
 add-ins for, *I:93, III:651*
 data series correlation in, *I:92–93*
 determining corresponding
 probabilities in, *III:646*
 Excel Link, *III:434*
 Excel Solver, *II:70*
 interactions with MATLAB, *III:448*
 macros in, *III:449, III:454–455*
 notations in, *III:477n*
 random number generation in,
 III:645–646
 random walks with, *I:83, I:85, I:87,*
 I:90
 @RISK in, *II:12f*
 syntax for functions in, *III:456*
- Exchange-rate intervention, study on,**
 III:177–178
- Exercise prices, I:452, I:484, I:508**
- Expectation maximization (EM)**
 algorithm, *II:146, II:165*
- Expectations, conditional, I:122,**
 II:517–518, III:508–509
- Expectations hypothesis, III:568–569,**
 III:601n
- Expected shortfall (ES), I:385–386,**
 III:332. See also average value at
 risk (AVaR)
- Expected tail loss (ETL), III:291,**
 III:293f, III:345–347, III:347f,
 III:355–356
- Expected value (EV), I:511**
- Expenses, noncash, II:25**
- Experiments, possibility of, II:307**
- Explicit costs, defined, III:623**
- Explicit Euler (EE) scheme, II:674,**
 II:677–678
- Exponential density function, III:218f**

- Exponential distribution, *III*:217–219
 applications in finance, *III*:219
- Exponentially weighted moving averages (EWMA)
 discussion of, *III*:409–413
 forecasting model of, *III*:411
 properties of the estimates, *III*:410–411
 standard errors for, *III*:411–412
 statistical methodology in, *III*:409
 usefulness of, *III*:413–414
 volatility estimates for, *III*:410*f*
- Exposures
 calculation of, *II*:247*t*
 correlation between, *II*:186
 distribution of, *II*:250*f*, *II*:251*f*, *II*:254
 management of, *II*:182–183
 monitoring of portfolio, *II*:249–250
 name-specific, *II*:188
- Extrema, characterization of local, *I*:23
- Extremal random variables, *III*:267
- Extreme value distributions, generalized, *III*:269
- Extreme value theory (EVT), *II*:744–746, *III*:95, *III*:228
 defined, *III*:238
 for IID processes, *III*:265–274
 in IID sequences, *III*:275
 role of in modeling, *II*:753*n*
- Factor analysis
 application of, *II*:165
 based on information coefficients, *II*:222
 defined, *II*:141, *II*:169
 discussion of, *II*:164–166
 importance of, *II*:238
vs. principal component analysis, *II*:166–168
- Factor-based strategies
vs. risk models, *II*:236
- Factor-based trading, *II*:196–197
 model construction for, *II*:228–235
 performance evaluation of, *II*:225–228
- Factor exposures, *II*:247–248, *II*:275–283
- Factorials, computing of, *III*:456
- Factorization, defined, *II*:307
- Factor mimicking portfolio (FMP), *II*:214
- Factor model estimation, *II*:142–147, *II*:150
 alternative approaches and extensions, *II*:145–147
 applied to bond returns, *II*:144–145
 computational procedure for, *II*:142–144
 fixed N, *II*:143
 large N, *II*:143–144
- Factor models
 in the Black-Litterman framework, *I*:200
 commonly used, *II*:150
 considerations in, *II*:178
 cross-sectional, *II*:220–221
 defined, *II*:153
 fixed income, *II*:271–272
 in forecasting, *II*:230–231
 linear, *II*:154–156, *II*:168
 normal, *II*:156
 predictive, *II*:142
 static/dynamic, *II*:146–147, *II*:155
 in statistical methodology, *II*:141
 strict, *II*:155–156
 types of, *II*:138–142
 usefulness of, *II*:154, *II*:503
 use of, *I*:354, *II*:137, *II*:150, *II*:168, *II*:219–225
- Factor portfolios, *II*:224–225
- Factor premiums, cross-sectional methods for evaluation of, *II*:214–219
- Factor returns, *II*:191*t*, *II*:192*t*
 calculation of, *II*:248
- Factor risk models, *II*:113, *II*:119
- Factors
 adjustment of, *II*:205–206
 analysis of data of, *II*:206–211
 categories of, *II*:197
 choice of, *II*:232–235
 defined, *II*:196, *II*:211
 desirable properties of, *II*:200
 development of, *II*:198
 estimation of types of, *II*:156
 graph of, *II*:166*f*
 known, *II*:138–139
 K systematic, *II*:138–139
 latent, *II*:140–141, *II*:150
 loadings of, *II*:144, *II*:145*t*, *II*:155, *II*:166*t*, *II*:167*f*, *II*:168*t*
 market, *II*:176
 orthogonalization of, *II*:205–206
 relationship to time series, *II*:168*f*
 sorting of, *II*:215
 sources for, *II*:200–201
 statistical, *II*:197
 summary of well-known, *II*:196*t*
 transformations applied to, *II*:206
 use of multiple, *II*:141–142
- Failures, probability of, *II*:726–727
- Fair equilibrium, between multiple accounts, *II*:76
- Fair value
 determination of, *III*:584–585
- Fair value, assessment of, *II*:6–7
- Fama, Eugene, *II*:468, *II*:473–474
- Fama-French three-factor model, *II*:139–140, *II*:177
- Fama-MacBeth regression, *II*:220–221, *II*:224, *II*:227–228, *II*:228*f*, *II*:237, *II*:240*n*
- Fannie Mae/Freddie Mac, writedowns of, *III*:77*n*
- Fast Fourier transform algorithm, *II*:743
- Fat tails
 of asset return distributions, *III*:242
 in chaotic systems, *II*:653
 class \mathcal{L} , *III*:261–263
 comparison between risk models, *II*:749–750
 effects of, *II*:354
 importance of, *II*:524
 properties of, *III*:260–261
 in Student's *t* distribution, *II*:734
- Favorable selection, *III*:76–77
- F*-distribution, *III*:216–217
- Federal Reserve
 effects of on inflation risk premium, *I*:281
 study by Cleveland Bank, *III*:177–178
 timing of interventions of, *III*:178
- Feynman-Kac formulas, *II*:661
- FFAs (freight forward agreements), *I*:566
- Filtered probability spaces, *I*:314–315, *I*:334*n*
- Filtration, *II*:516–517, *III*:476–477, *III*:489–490, *III*:508
- Finance, three major revolutions in, *III*:350
- Finance companies, captive, *I*:366–369
- Finance theory
 development of, *II*:467–468
 effect of computers on, *II*:476
 in the nineteenth century, *II*:468–469, *II*:476
 in the 1960s, *II*:476
 in the 1970s, *II*:476
 stochastic laws in, *III*:472
 in the twentieth century, *II*:476
- Financial assets, price distribution of, *III*:349–350
- Financial crisis (2008), *III*:71
- Financial date, pro forma, *II*:542–543
- Financial distress, defined, *I*:351
- Financial institutions, model risk of, *II*:693
- Financial leverage ratios, *II*:559–561, *II*:563
- Financial modelers, mistakes of, *II*:707–710

- Financial planning, *III*:126–127, *III*:128, *III*:129
- Financial ratios, *II*:546, *II*:563–564
- Financial statements
 assumptions used in creating, *II*:532
 data in, *II*:563
 information in, *II*:533–542, *II*:543
 pro forma, *II*:22–23
 time statements for, *II*:532
 usefulness of, *II*:531
 use of, *II*:204–205, *II*:246
- Financial time series, *I*:79–80, *I*:386–387, *II*:415–416, *II*:503–504
- Financial variables, modeling of, *III*:280
- Find, in MATLAB, *III*:422
- Finite difference methods, *II*:648–652, *II*:656–657, *II*:665–666, *II*:674–675, *II*:676–677, *III*:19
- Finite element methods, *II*:669–670, *II*:672, *II*:679–681
- Finite element space, *II*:670–672
- Finite life general DDM, *II*:5–7
- Finite states, assumption of, *I*:100–101
- Firms
 assessment of, *II*:546–547
 and capital structure, *II*:473
 characteristics of, *II*:94, *II*:176–177, *II*:201
 clientele of, *II*:36
 comparable, *II*:34, *II*:35–36
 geographic location of, *II*:36
 history *vs.* future prospects, *II*:92
 phases of, *II*:9–10
 retained earnings of, *II*:20
 valuation of, *II*:26–27, *II*:473
 value of, *II*:27–31, *II*:39
vs. characteristics of group, *II*:90–91
- First boundary problem, *II*:655–656, *II*:657f
- First Interstate Bancorp, *I*:304
 analysis of credit spreads, *I*:305t
 debt ratings of, *I*:410
- First passage models (FPMs), *I*:342, *I*:344–348
- Fischer-Tippett theorem, *III*:266–267
- Fisher, Ronald, *I*:140
- Fisherian, defined, *I*:140
- Fisher's information matrix, *I*:160n
- Fisher's law, *II*:322–323
- Fixed-asset turnover ratio, *II*:558
- Fixed-charge coverage ratio, *II*:560–561
- Flesaker-Hughston (FH) model, *III*:548–549
- Flows, discrete, *I*:448–453
- FMP (factor mimicking portfolio), *II*:214
- Footnotes, in financial statements, *II*:541–542
- Ford Motor Company, *I*:408f, *I*:409f
- Forecastability, *II*:132
- Forecastability, concept of, *II*:123
- Forecast encompassing
 defined, *II*:230–231
- Forecasts
 of bid-ask spreads, *II*:456–457
 comparisons of, *II*:420–421
 contingency tables, *II*:429t
 development of, *II*:110–114
 directional, *II*:428
 effect on future of, *II*:122–123
 errors in, *II*:422f
 evaluation of, *II*:428–430, *III*:368–370
 machine-learning approach to, *II*:128
 measures of, *II*:429–430, *II*:430
 need for, *II*:110–111
 in neural networks, *II*:419–420
 one-step ahead, *II*:421f
 parametric bootstraps for, *II*:428–430
 response to macroeconomic shocks, *II*:55f
 usefulness of, *II*:131–132
 use of models for, *II*:302
 of volatility, *III*:412
- Foreclosures, *III*:31, *III*:75
- Forward contracts
 advantages of, *I*:430
 buying assets of, *I*:439
 defined, *I*:426, *I*:478
 equivalence to futures prices, *I*:432–433
 hedging with, *I*:429, *I*:429t
 as OTC instruments, *I*:479
 prepaid, *I*:428
 price paths of, *I*:428t
 short *vs.* long, *I*:437–438, *I*:438f
 valuing of, *I*:426–430
vs. futures, *I*:430–431, *I*:433
vs. options, *I*:437–439
- Forward curves
 graph of, *I*:434f
 modeling of, *I*:533, *I*:557–558, *I*:564–565
 normal *vs.* inverted, *I*:434
 of physical commodities, *I*:555
- Forward freight agreements (FFAs), *I*:555, *I*:558, *I*:566
- Forward measure, use of, *I*:543–544
- Forward rates
 calculation of, *I*:491, *III*:572
 defined, *I*:509–510
 from discount function, *III*:566–567
 implied, *III*:565–567
 models of, *III*:543–544
 from spot yields, *III*:566
 of term structure, *III*:586
- Fourier integrals, *II*:656
- Fourier methods, *I*:559–560
- Fourier transform, *III*:265
- FPMs (first passage models), *I*:342, *I*:344–348
- Fractals, *II*:653–654, *III*:278–280, *III*:479–480
- Franklin Tempelton Investment Funds, *II*:496t, *II*:497t, *II*:498t
- Fréchet distribution, *II*:754n, *III*:228, *III*:230, *III*:265, *III*:267, *III*:268
- Fréchet-Hoeffding copulas, *I*:327, *I*:329
- Freddie Mac, *II*:77n, *II*:754n, *III*:49
- Free cash flow (FCF), *II*:21–23
 analysis of, *II*:570–571
 calculation of, *II*:23–24, *II*:571–572
 defined, *II*:569–571, *II*:578
 expected for XYZ, Inc., *II*:30t
 financial adjustments to, *II*:25–26
 statement of, direct method, *II*:24–25, *II*:24t
 statement of, indirect method, *II*:24–25, *II*:24t
vs. cash flow, *II*:22–23
- Freedman-Diaconis rule, *II*:494, *II*:495, *II*:497
- Frequencies
 accumulating, *II*:491–492
 distributions of, *II*:488–491, *II*:499f
 empirical cumulative, *II*:492
 formal presentation of, *II*:491
- Frequentist, *I*:140, *I*:148
- Frictions, costs of, *II*:472–473
- Friedman, Milton, *I*:123
- Frontiers, true, estimated and actual efficient, *I*:190–191
- F_SCORE, use of, *II*:230–231
- F-test, *II*:336, *II*:337, *II*:344, *II*:425, *II*:426
- FTSE 100, volatility in, *III*:412–413
- Fuel costs, *I*:561, *I*:562–563. *See also* energy
- Full disclosure, defined, *II*:532
- Functional, defined, *I*:24
- Functional-coefficient autoregressive (FAR) model, *II*:417
- Functions
 affine, *I*:31
 Archimedean, *I*:329, *I*:330–331, *I*:331
 Bessel, of the third kind, *II*:591
 beta, *II*:591
 characteristic, *II*:591–592, *II*:593
 choosing and calibrating of, *I*:331–333
 Clayton, Frank, Gumbel, and Product, *I*:329

- Functions (*Continued*)
 continuous, *II:581–584, II:582f, II:583, II:592–593*
 continuous / discontinuous, *II:582f*
 convex, *I:24–27, I:25, I:25f, I:26f*
 convex quadratic, *I:26, I:31f*
 copula, *I:320, I:325–333, I:407–408*
 for default times, *I:329–331*
 defined, *I:24, I:333*
 density, *I:141*
 with derivatives, *II:585f*
 elementary, *III:474*
 elliptical, *I:328–329*
 empirical distribution, *III:270*
 factorial, *II:590–591*
 gamma, *II:591, II:591f, III:212*
 gradients of, *I:23*
 Heaviside, *II:418–419*
 hypergeometric, *III:256, III:257*
 indicator, *II:584–585, II:584f, II:593*
 likelihood function, *I:141–143, I:143f, I:144f, I:148, I:176, I:177*
 measurable, *III:159–160, III:160f, III:201*
 minimization and maximization of values, *I:22, I:22f*
 monotonically increasing, *II:587–588, II:588f*
 nonconvex quadratic, *I:26–27*
 nondecreasing, *III:154–155, III:155f*
 normal density, *III:226f*
 optimization of, *I:24*
 parameters of copulas, *I:331–332*
 properties of quasi-convex, *I:28*
 quasi-concave, *I:27–28, I:27f*
 right-continuous, *III:154–155, III:155f*
 surface of linear, *I:33f*
 with two local maxima, *I:23f*
 usefulness of, *I:411–412*
 utility, *I:4–5, I:14–15, I:461*
 Fund management, art of, *I:273*
 Fund separation theorems, *I:36*
 Futures
 Eurodollar, *I:503*
 hedging with, *I:433*
 market for housing, *II:396–397*
 prices of, and interest rates, *I:435n*
 telescoping positions of, *I:431–432*
 theoretical, *I:487*
 valuing of, *I:430–433*
 vs. forward contracts, *I:430–431*
 Futures contracts
 defined, *I:478*
 determining price of, *I:481*
 pricing model for, *I:479–481*
 theoretical price of, *I:481–484*
 vs. forward contracts, *I:433, I:478–479*
 Futures options, defined, *I:453*
 Future value, *II:618*
 determining of money, *II:596–600*
 Galerkin methods, principle of, *II:671*
 Gamma, *I:509, I:518–520*
 Gamma process, *III:498*
 Gamma profile, *I:519f*
 Gapping effect, *I:509*
 GARCH (generalized autoregressive conditional heteroskedastic) models
 asymmetric, *II:367–368*
 exponential (EGARCH), *II:367–368*
 extensions of, *III:657*
 factor models, *II:372*
 GARCH-M (GARCH in mean), *II:368*
 Markov-switching, *I:180–184*
 time aggregation in, *II:369–370*
 type of, *II:131*
 usefulness of, *III:414*
 use of, *I:175–176, I:185–186, II:371, II:733–734, III:388*
 and volatility, *I:179*
 weights in, *II:363–364*
 GARCH (1,1) model
 Bayesian estimation of, *I:176–180*
 defined, *II:364*
 results from, *II:366, II:366f*
 skewness of, *III:390–391*
 strengths of, *III:388–389*
 Student's *t*, *I:182*
 use of, *I:550–551, III:656–657*
 GARCH (1,1) process, *I:551t*
 Garman-Kohlhagen system, *I:510–511, I:522*
 Gaussian density, *III:98f*
 Gaussian model, *III:547–548*
 Gaussian processes, *III:280, III:504*
 Gaussian variables, and Brownian motion, *III:480–481*
 Gauss-Markov theorem, *II:314*
 GBM (geometric Brownian motion), *I:95, I:97*
 GDP (gross domestic product), *I:278, I:282, II:138, II:140*
 General inverse Gaussian (GIG) distribution, *II:523–524*
 Generalized autoregressive conditional heteroskedastic (GARCH) models. *See* GARCH (generalized autoregressive conditional heteroskedastic) models
 Generalized central limit theorem, *III:237, III:239*
 Generalized extreme value (GEV) distribution, *II:745, III:228–230, III:272–273*
 Generalized inverse Gaussian distribution, use of, *II:521–522*
 Generalized least squares (GLS), *I:198–199, II:328*
 Generalized tempered stable (GTS) processes, *III:512*
 Generally accepted accounting principles (GAAP), *II:21–22, II:531–532, II:542–543*
 Geometric mean reversion (GMR) model, *I:91–92*
 computation of, *I:91*
 Gibbs sampler, *I:172n, I:179, I:184–185*
 GIG models, calibration of, *II:526–527*
 Gini index of dissimilarity (Gini measure), *III:353–354*
 Ginnie Mae/Fannie Mae/Freddie Mac, actions of, *III:49*
 Girsanov's theorem
 and Black-Scholes option pricing formula, *I:132–133*
 with Brownian motion, *III:511*
 and equivalent martingale measures, *I:130–133*
 use of, *I:263, III:517*
 Glivenko-Cantelli theorem, *III:270, III:272, III:348n, III:646*
 Global Economy Workshop, Santa Fe Institute, *II:699*
 Global Industry Classification Standard (GICS®), *II:36–37, II:248*
 Global minimum variance (GMV) portfolios, *I:39*
 GMR (geometric mean reversion) model, *I:91–92*
 GMV (global minimum variance) portfolios, *I:15, I:194–195*
 GNP, growth rate of (1947–1991), *II:410–411, II:410f*
 Gradient methods, use of, *II:684*
 Granger causality, *II:395–396*
 Graphs, in MATLAB, *III:428–433*
 Greeks, the, *I:516–522*
 beta and omega, *I:522*
 delta, *I:516–518*
 gamma, *I:518–520*
 rho, *I:521–522*
 theta, *I:509, I:520–521*
 use of, *I:559, II:660, III:643–644*
 vega, *I:521*
 Greenspan, Alan, *I:140–141*
 Growth, *I:283f, II:239, II:597–598, II:601–602*
 Gumbel distribution, *III:265, III:267, III:268–269*

- Hamilton-Jacobi equations, *II:675*
- Hankel matrices, *II:512*
- Hansen-Jagannathan bound, *I:59, I:61–62*
- Harrison, Michael, *II:476*
- Hazard, defined, *III:85*
- Hazard (failure) rate, calculation of, *III:94–95*
- Heat diffusion equation, *II:470*
- Heath-Jarrow-Morton framework, *I:503, I:557*
- Heavy tails, *III:227, III:382*
- Hedge funds, and probit regression model, *II:349–350*
- Hedge ratios, *I:416–417, I:509*
- Hedges
 - importance of, *I:300*
 - improvement using DTS, *I:398*
 - in the Merton context, *I:409*
 - rebalancing of, *I:519*
 - risk-free, *I:532f*
- Hedge test, *I:409, I:411*
- Hedging
 - costs of, *I:514, II:725*
 - and credit default swaps, *I:413–414*
 - determining, *I:303–304*
 - with forward contracts, *I:429, I:429t*
 - of fuel costs, *I:561*
 - with futures, *I:433*
 - gamma, *I:519*
 - portfolio-level, *I:412–413*
 - of positions, *II:724–726*
 - ratio for, *II:725*
 - with swaps, *I:434–435*
 - transaction-level, *I:412*
 - usefulness of, *I:418*
 - use of, *I:125–126*
 - using macroeconomic indices, *I:414–417*
- Hessian matrix, *I:23–24, I:25, I:186n, III:645*
- Heston model, *I:547, I:548, I:552, II:682*
 - with change of time, *III:522*
- Heteroskedasticity, *II:220, II:359, II:360, II:403*
- HFD (high-frequency data). *See* high-frequency data (HFD)
- Higham's projection algorithm, *II:446*
- High-dimensional problems, *II:673*
- High-frequency data (HFD)
 - and bid-ask bounce, *II:454–457*
 - defined, *II:449–450*
 - generalizations to, *II:368–370*
 - Level I, *II:451–452, II:452f, II:453t*
 - Level II, *II:451*
 - properties of, *II:451, II:453t*
 - recording of, *II:450–451*
 - time intervals of, *II:457–462*
 - use of, *II:300, II:481*
 - volume of, *II:451–454*
- Hilbert spaces, *II:683*
- Hill estimator, *II:747, III:273–274*
- Historical method
 - drawbacks of, *III:413*
 - weighting of data in, *III:397–398*
- Hit rate, calculation of, *II:240n*
- HJM framework, *I:498*
- HJM methodology, *I:496–497*
- Holding period return, *I:6*
- Ho-Lee model
 - continuous variant for, *I:497*
 - defined, *I:492*
 - in history, *I:493*
 - interest rate lattice, *III:614f*
 - as short rate model, *III:23*
 - for short rates, *III:605*
 - as single factor model, *III:549*
- Home equity prepayment (HEP) curve, *III:55–56, III:56f*
- Homeowners, refinancing behavior of, *III:25*
- Home prices, *I:412, II:397f, II:399t, III:74–75*
- Homoskedasticity, *II:360, II:373*
- Horizon prices, *III:598*
- Housing, *II:396–399, III:48*
- Howard algorithm (policy iteration algorithm), *II:676–677, II:680*
- Hull-White (HW) models
 - binomial lattice, *III:610–611*
 - for calibration, *II:681*
 - defined, *I:492*
 - interest rate lattice, *III:614f*
 - and short rates, *III:545–546*
 - for short rates, *III:605*
 - trinomial lattice, *III:613, III:616f*
 - usefulness of, *I:503*
 - use of, *III:557, III:604*
 - valuing zero-coupon bond calls with, *I:500*
- Hume, David, *I:140*
- Hurst, Harold, *II:714*
- Hypercubes, use of, *III:648*
- IBM stock, log returns of, *II:407f*
- Ignorance, prior, *I:153–154*
- Implementation risk, *II:694*
- Implementation shortfall approach, *III:627*
- Implicit costs, *III:631*
- Implicit Euler (IE) scheme, *II:674, II:677–678*
- Implied forward rates, *III:565–567*
- Impurity, measures of, *II:377*
- Income, defined for public corporation, *II:21–22*
- Income statements
 - common-size, *II:562–563, II:562t*
 - defined, *II:536*
 - in financial statements, *II:536–537*
 - sample, *II:537t, II:547t*
 - structure of, *II:536*
 - XYZ Inc. (example), *II:28t*
- Income taxes. *See* taxes
- Independence, *I:372–373, II:624–625, III:363–364, III:368*
- Independence function, in VaR models, *III:365–366*
- Independently and identically distributed (IDD) concept, *I:164, I:171, II:127, III:274–280, III:367, III:414*
- Indexes
 - characteristics of efficient, *I:42t*
 - defined, *II:67*
 - of dissimilarity, *III:353–354*
 - equity, *I:15t, II:190t, II:262–263*
 - tail, *II:740–741, II:740f, III:234*
 - tracking of, *II:64, II:180*
 - use of weighted market cap, *I:38*
 - value weighted, *I:76–77*
 - volatility, *III:550–552, III:552f*
- Index returns, scenarios of, *II:190t, II:191t*
- Indifference curves, *I:4–5, I:5f, I:14*
- Industries, characteristics of, *II:36–37, II:39–40*
- Inference, *I:155–158, I:169t*
- Inflation
 - effect on after-tax real returns, *I:286–287*
 - and GDP growth, *I:282*
 - indexing for, *I:278–279*
 - in regression analysis, *II:323*
 - risk of, *II:282*
 - risk premiums for, *I:280–283*
 - seasonal factors in, *I:292*
 - shifts in, *I:285f*
 - volatility of, *I:281*
- Information
 - anticipation of, *III:476*
 - from arrays in MATLAB, *III:421*
 - completeness of, *I:353–354*
 - contained in high volatility stocks, *III:629*
 - and filtration, *III:517*
 - found in data, *II:486*
 - and information propagation, *II:515*
 - insufficient, *III:44*
 - integration of, *II:481–482*
 - overload of, *II:481*
 - prior in Bayesian analysis, *I:151–155, I:152*
 - propagation of, *I:104*

- Information (*Continued*)
 structures of, I:106f, II:515–517
 unstructured *vs.* semistructured,
 II:481–482
- Information coefficients (ICs), II:98–99,
 II:221–223, II:223f, II:227f, II:234
- Information ratios
 defined, II:86n, II:115, II:119, II:237
 determining, II:100f
 for portfolio sorts, II:219
 use of, II:99–100
- Information sets, II:123
- Information structures
 defined, II:518
- Information technology, role of,
 II:480–481
- Ingersoll models, I:271–273, I:275f
- Initial conditions, fixing of, II:502
- Initial margins, I:478
- Initial value problems, II:639
- Inner quartile range (IQR), II:494
- Innovations, II:126
- Insurance, credit, I:413–414
- Integrals, II:588–590, II:593. *See also*
 stochastic integrals
- Integrated series, and trends,
 II:512–514
- Integration, stochastic, III:472, III:473,
 III:483
- Intelligence, general, II:154
- Intensity-based frameworks, and the
 Poisson process, I:315
- Interarrival time, III:219, III:225
- Intercepts, treatment of, II:334–335
- Interest
 accumulated, II:604–605, II:604f
 annual *vs.* quarterly compounding,
 II:599f
 compound, II:597, II:597f
 computing accrued, and clean price,
 I:214–215
 coverage ratio, II:560
 defined, II:596
 determining unknown rates,
 II:601–602
 effective annual rate (EAR),
 II:616–617
 mortgage, II:398
 simple *vs.* compound, II:596
 terms of, II:619
 from TIPS, I:277
- Interest rate models
 binomial, III:173–174, III:174f
 classes of, III:600
 confusions about, III:600
 importance of, III:600
 properties of lattices, III:610
 realistic, arbitrage-free, III:599
 risk-neutral/arbitrage-free, III:597
- Interest rate paths, III:6–9, III:7, III:8t
- Interest rate risk, III:12–14
- Interest rates
 absolute *vs.* relative changes in,
 III:533–534
 approaches in determining future,
 III:591
 binomial model of, III:173–174
 binomial trees, I:236, I:236f, I:237f,
 I:240f, I:244, I:244f, III:174f
 borrowing *vs.* lending, I:482–483
 calculation of, II:613–618
 calibration of, I:495
 caps/caplets of, III:589–590
 caps on, I:248–249
 categories of term structure, III:561
 computing sensitivities, III:22–23
 continuous, I:428, I:439–488
 derivatives of, III:589–590
 determination of appropriate,
 I:210–211
 distribution of, III:538–539
 dynamic of process, I:262
 effect of, I:514–515
 effect of shocks, III:23
 effect on putable bonds, III:303–304
 future course of, III:567, III:573
 and futures prices, I:435n
 importance of models, III:600
 jumps of, III:539–541
 jumpy and continuous, III:539f
 long *vs.* short, III:538
 market spot/forward, I:495t
 mean reversion of, III:7
 modeling of, I:261–265, I:267, I:318,
 I:491, I:503, III:212–213
 multiple, II:599–600
 negative, III:538
 nominal, II:615–616
 and option prices, I:486–487
 and prepayment risk, III:48
 risk-free, I:442
 shocks/shifts to, III:585–596
 short-rate, I:491–494, III:595
 simulation of, III:541
 stochastic, I:344, I:346
 structures of, III:573, III:576
 use of for control, I:489
 volatility of, III:405, III:533
- Intermarket relations, no-arbitrage,
 I:453–455
- Internal consistency rule, in OAS
 analysis, I:265
- Internal rate of return (IRR), II:617–618
 in MBSs, III:36
- International Monetary Fund
 Global Stability Report, I:299
- International Swap and Derivatives
 Association (ISDA). *See* ISDA
- Interpolated spread (I-spread), I:227
- Interrate relationship, arbitrage-free,
 III:544
- Intertemporal dependence, and risk,
 III:351
- Intertrade duration, II:460–461,
 II:462t
- Intertrade intervals, II:460–461
- Intervals, credible, I:170
- Interval scales, data on, II:487
- Intrinsic value, I:441, I:511, I:513,
 II:16–17
- Invariance property, III:328–329
- Inventory, II:542, II:557
- Inverse Gaussian process, III:499
- Investment, goals of, II:114–115
- Investment management, III:146
- Investment processes
 activities of integrated, II:61
 evaluation of results of, II:117–118
 model creation, II:96
 monitoring of performance, II:104
 quantitative, II:95, II:95f
 quantitative equity, II:95f, II:96f,
 II:105
 research, II:95–102
 sell-structured, II:108
 steps for equity investment, II:119
 testing of, II:109
- Investment risk measures, III:350–351
- Investments, I:77–78n, II:50–51,
 II:617–618
- Investment strategies, II:66–67,
 II:198
- Investment styles, quantamental,
 II:93–94, II:93f
- Investors
 behavior of, II:207, II:504
 comfort with risk, I:193
 completeness of information of,
 I:353–354
 focus of, I:299, II:90–91
 fundamental *vs.* quantitative,
 II:90–94, II:91f, II:92f, II:105
 goals/objectives of, II:114–115,
 II:179, III:631
 individual accounts of, II:74
 monotonic preferences of, I:57
 number of stocks considered, II:91
 preferences of, I:5, I:260, II:48, II:56,
 II:92–93
 prior beliefs of, II:727
 real-world, II:132
 risk aversion of, II:82–83, II:729
 SL-CAPM assumptions about, I:66
 sophistication of, II:108
 in uncertain markets, II:54
 views of, I:197–199
- Invisible hand, notion of, II:468–469

- ISDA (International Swap and Derivatives Association)
 Credit Derivative Definitions (1999), *I:230, I:528*
 Master Agreement, *I:538*
 organized auctions, *I:526–527*
 supplement definition, *I:230*
- I-spread (interpolated spread), *I:227*
- Ito, Kiyosi, *II:470*
- Ito definition, *III:486–487*
- Ito integrals, *I:122, III:475, III:481, III:490–491*
- Ito isometry, *III:475*
- Ito processes
 defined, *I:95*
 generic univariate, *I:125*
 and Girsanov's theorem, *I:131*
 under HJM methodology, *I:497*
 properties of, *III:487–488*
 and smooth maps, *III:493*
- Ito's formula, *I:126, III:488–489*
- Ito's lemma
 defined, *I:98*
 discussion of, *I:95–97*
 in estimation, *I:348*
 and the Heston model, *I:548*
- James-Stein shrinkage estimator, *I:194*
- Japan, credit crisis in, *I:417*
- Jarrow-Turnbull model, *I:307*
- Jarrow-Yu propensity model, *I:324–325*
- Jeffreys' prior, *I:153, I:160n, I:171–172*
- Jensen's inequality, *I:86, III:569*
- Jevons, Stanley, *II:468*
- Johansen-Juselius cointegration tests, *II:391–393, II:395*
- Joint jumps/defaults, *I:322–324*
- Joint survival probability, *I:323–324*
- Jordan diagonal blocks, *II:641–642*
- Jorion shrinkage estimator, *I:194, I:202*
- Jump-diffusion, *III:554–557, III:657*
- Jumps
 default, *I:322–324*
 diffusions, *I:559–560*
 downward, *I:347*
 idiosyncratic, *I:323*
 incorporation of, *I:93–94*
 in interest rates, *III:539–541*
 joint, *I:322–324*
 processes of, *III:496*
 pure processes, *III:497–501, III:506*
 size of, *III:540*
- Kalotay-Williams-Fabozzi (KWF) model, *III:604, III:606–607, III:615f*
- Kamakura Corporation, *I:301, I:307, I:308–309, I:310n*
- Kappa, *I:521*
- Karush-Kuhn-Tucker conditions (KKT conditions), *I:28–29*
- Kendall's tau, *I:327, I:332*
- Kernel regression, *II:403, II:412–413, II:415*
- Kernels, *II:412, II:413f, II:746*
- Kernel smoothers, *II:413*
- Keynes, John Maynard, *II:471*
- Key rate durations (KRD), *II:276, III:311–315, III:317*
- Key rates, *II:276, III:311*
- Kim-Rachev (KR) process, *III:512–513*
- KKT conditions (Karush-Kuhn-Tucker conditions), *I:28–29, I:31, I:32*
- KoBoL distribution, *III:257n*
- Kolmogorov extension theorem, *III:477–478*
- Kolmogorov-Smirnov (KS) test, *II:430, III:366, III:647*
- Kolmogorov equation, use of, *III:581*
- Kreps, David, *II:476*
- Krispy Kreme Doughnuts, *II:574–575, II:574f*
- Kronecker product, *I:172, I:173n*
- Kuiper test, *III:366*
- Kurtosis, *I:41, III:234*
- Lag operator L , *II:504–506, II:507, II:629–630*
- Lagrange multipliers, *I:28, I:29–31, I:30, I:32*
- Lag times, *II:387, III:31*
- Laplace transforms, *II:647–648*
- Last trades, price and size of, *II:450*
- Lattice frameworks
 bushy trees in, *I:265, I:266f*
 calibration of, *I:238–240*
 fair, *I:235*
 interest rate, *I:235–236, I:236–238*
 one-factor model, *I:236f*
 for pricing options, *I:487*
 usefulness of, *I:235*
 use of, *I:240, I:265–266, III:14*
 value at nodes, *I:237–238*
 1-year rates, *I:238f, I:239f*
- Law of iterated expectations, *I:110, I:122, II:308*
- Law of large numbers, *I:267, I:270n, III:263–264, III:275*
- Law of one α , *II:50*
- Law of one price (LOP), *I:52–55, I:99–100, I:102, I:119, I:260*
- LCS (liquidity cost score), *I:402*
 use of, *I:403*
- LDIs (liability-driven investments), *I:36*
- LD (loss on default), *I:370–371*
- Leases, in financial statements, *II:542*
- Least-square methods, *II:683–685*
- Leavens, D. H., *I:10*
- Legal loss data
 Cruz study, *III:113, III:115t*
 Lewis study, *III:117, III:117t*
- Lehman Brothers, bankruptcy of, *I:413*
- Level (parallel) effect, *II:145*
- Lévy-Khinchine formula, *III:253–254, III:257*
- Lévy measures, *III:254, III:254t*
- Lévy processes
 and Brownian motion, *III:504*
 in calibration, *II:682*
 change of measure for, *III:511–512*
 conditions for, *III:505*
 construction of, *III:506*
 from Girsanov's theorem, *III:511*
 and Poisson process, *III:496*
 as stochastic process, *III:505–506*
 as subordinators, *III:521*
 for tempered stable processes, *III:512–514, III:514t*
 and time change, *III:527*
- Lévy stable distribution, *III:242, III:339, III:382–386, III:392*
- LGD (loss given default), *I:366, I:370, I:371*
- Liabilities, *II:533, II:534–535, III:132*
- Liability-driven investments (LDIs), *I:36*
- Liability-hedging portfolios (LHPs), *I:36*
- LIBOR (London Interbank Offered Rate)
 and asset swaps, *I:227*
 changes in, by type, *III:539–540*
 curve of, *I:226*
 interest rate models, *I:494*
 market model of, *III:589*
 spread of, *I:530*
 in total return swaps, *I:541*
 use of in calibration, *III:7*
- Likelihood maximization, *I:176*
- Likelihood ratio statistic, *II:425*
- Limited liability rule, *I:363*
- Limit order books, use of, *III:625, III:632n*
- Lintner, John, *II:474*
- Lipschitz condition, *II:658n, III:489, III:490*
- Liquidation
 effect of, *II:186*
 procedures for, *I:350–351*
 process models for, *I:349–351*
 time of, *I:350*
 vs. default event, *I:349*
- Liquidity
 assumption of, *III:371*
 in backtesting, *II:235*
 changes in, *I:405*

- Liquidity (*Continued*)
 cost of, I:401
 creation of, III:624–625, III:631
 defined, III:372, III:380
 effect of, II:284
 estimating in crises, III:378–380
 in financial analysis, II:551–555
 and LCS, I:404
 and market costs, III:624
 measures of, II:554–555
 premiums on, I:294, I:307
 ratios for, II:555
 in risk modeling, II:693
 shortages in, I:347–348
 and TIPS, I:293, I:294
 and transaction costs, III:624–625
- Liquidity-at-risk (LAR), III:376–378
- Liquidity cost, III:373–374, III:375–376
- Liquidity cost score (LCS), I:402, I:403
- Liquidity preference hypothesis, III:570
- Liquidity ratios, II:563
- Liquidity risk, II:282, III:380
- Ljung-Box statistics, II:407, II:421, II:422, II:427–428
- LnMix models, calibration of, II:526–527
- Loading, standardization of, II:177
- Loan pools, III:8–9
- Loans
 amortization of, II:606–607, II:611–613
 amortization table for, II:612t
 delinquent, III:63
 fixed rate, fully amortized schedule, II:614t
 floating rate, II:613
 fully amortizing, II:611
 modified, III:32
 nonperforming, III:75
 notation for delinquent, III:45n
 recoverability of, III:31–32
 refinancing of, III:68–69
 repayment of, II:612f, II:613f
 term schedule, II:615t
- Loan-to-value ratios (LTVs), III:31–32, III:69, III:73, III:74–75
- Location parameters, I:160n, III:201–202
- Location-scale invariance property (Gaussian distribution), II:732
- Logarithmic Ornstein-Uhlenbeck (log-OU) processes, I:557–558
- Logarithmic returns, III:211–212, III:225
- Logistic distribution, II:350
- Logistic regression, I:307, I:308, I:310
- Logit regression models, II:349–350, II:350
- Log-Laplace transform, III:255–256
- Lognormal distribution, III:222–225, III:392
- Lognormal mixture (LnMix) distribution, II:524–525
- Lognormal variables, I:86
- Log returns, I:85–86, I:88
- London Interbank Offered Rate (LIBOR). *See* LIBOR
- Lookback options, I:114, III:24
- Lookback periods, III:402, III:407
- LOP (law of one price). *See* law of one price (LOP)
- Lorenz, Edward, II:653
- Loss distributions, conditional, III:340–341
- Losses. *See also* operational losses
 allocation of, III:32
 analysis of in backtesting, III:338
 collateral *vs.* tranche, III:36
 computation of, I:383
 defined, III:85
 estimation of cumulative, III:39–40
 expected, I:369–370, I:373–374
 expected *vs.* unexpected, I:369, I:375–376
 internal *vs.* external, III:83–84
 median of conditional, III:348n
 projected, III:37f
 restricting severity of, I:385–386
 severity of, III:44
 unexpected, I:371–372, I:374–375
- Loss functions, I:160n, III:369
- Loss given default (LGD), I:366, I:370, I:371
- Loss matrix analysis, III:40–41
- Loss on default (LD), I:370–371
- Loss severity, III:30–31, III:60–62, III:97–99
- Lottery tickets, I:462
- Lower partial moment risk measure, III:356
- Lundbert, Filip, II:467, II:470–471
- Macroeconomic influences, defined, II:197
- Magnitude measures, II:429–430
- Maintenance margins, I:478
- Major indexes, modeling return distributions for, III:388–392
- Malliavin calculus, III:644
- Management, active, II:115
- Mandelbrot, Benoit, II:653, II:738, III:234, III:241–242
- Manufactured housing prepayment (MHP) curve, III:56
- Marginalization, II:335
- Marginal rate of growth, III:197–198
- Marginal rate of substitution, I:60
- Margin calls, exposure to, III:377
- Market cap *vs.* firm value, II:39
- Market completeness, I:52, I:105
- Market efficiency, I:68–73, II:121, II:473–474
- Market equilibrium
 and investor's views, I:198–199
- Market impact
 costs of, III:623–624, III:627
 defined, II:69
 forecasting/modeling of, III:628–631
 forecasting models for, III:632
 forecasting of, III:628–629, III:629–631
 measurement of, III:626–628
 between multiple accounts, II:75–76
 in portfolio construction, II:116
 and transaction costs, II:70
- Market model regression, II:139
- Market opportunity, two state, I:460f
- Market portfolios, I:66–67, I:72–73
- Market prices, I:57, III:372
- Market risk
 approaches to estimation of, III:380
 in bonds, III:595
 in CAPM, I:68–69, II:474
 importance of, III:81
 models for, III:361–362
 premium for, I:203n, I:404
- Markets
 approach to segmented, II:48–51
 arbitrage-free, I:118
 complete, I:51–52, III:578
 complex, II:49
 effect of uncertainty in on bid-ask spreads, II:455–456
 efficiency of, II:15–16
 frictionless, I:261
 incomplete, I:461–462
 liquidity of, III:372
 models of, III:589
 for options and futures, I:453–454
 perfect, II:472
 properties of modern, III:575–576
 sensitivities to value-related variables, II:54t
 simple, I:70
 systematic fluctuations in, II:172–173
 unified approach to, II:49
 up/down, defined, II:347
- Market sectors, defined, III:560
- Market standards, I:257
- Market structure, and exposure, II:269–270
- Market timing, II:260
- Market transactions, upstairs, III:630–631, III:632n

- Market weights, *II:269t*
- Markov chain approximations, *II:678*
- Markov chain Monte Carlo (MCMC)
methods, *II:410f, II:417–418*
- Markov coefficients, *II:506–507, II:512*
- Markov matrix, *I:368*
- Markov models, *I:114*
- Markov processes
in dynamic term structures, *III:579*
hidden, *I:182*
use of, *III:509, III:517*
- Markov property, *I:82, I:180–181, I:183, II:661, III:193n*
- Markov switching (MS) models
discussion of, *I:180–184*
and fat tails, *III:277–278*
stationarity with, *III:275*
usefulness of, *II:433*
use of, *II:409–411, II:411t*
- Markowitz, Harry M., *I:38, I:140, II:467, II:471–472, III:137, III:351–352*
- Markowitz constraint sets, *I:69, I:72*
- Markowitz diversification, *I:10–11, I:11*
- Markowitz efficient frontiers, *I:191f*
- Markowitz model
in financial planning, *III:126*
- Mark-to-market (MTM)
calculation of value, *I:535–536, I:536t*
defined, *I:535*
and telescoping futures, *I:431–432*
- Marshall and Siegel, *II:694*
- Marshall-Olkin copula, *I:323–324, I:329*
- Martingale measures, equivalent
and arbitrage, *I:111–112, I:124*
and complete markets, *I:133*
defined, *I:110–111*
and Girsanov's theorem, *I:130–133*
and state prices, *I:133–134*
use of, *I:130–131*
working with, *I:135*
- Martingales
with change of time methods
(CTM), *III:522–523*
defined, *II:124, II:126, II:519*
development of concept, *II:469–470*
equivalent, *II:476*
measures of, *I:110–111*
use of conditions, *I:116*
use of in forward rates, *III:586*
- Mathematical theory, importance of
advances in, *III:145*
- Mathworks, website of, *III:418*
- MATLAB
array operations in, *III:420–421*
basic mathematical operations in,
III:419–420
- construction of vectors/matrices,
III:420
- control flow statements in,
III:427–428
- desktop, *III:419f*
- European call option pricing with,
III:444–445
- functions built into, *III:421–422*
- graphs in, *III:428–433, III:429–430f, III:431f*
- interactions with other software,
III:433–434
- M-files in, *III:418–419, III:423, III:447*
- operations in, *III:447*
- optimization in, *III:434–444, III:435t*
- Optimization Tool, *III:435–436, III:436f, III:440f, III:441f*
- overview of desktop and editor,
III:418–419
- quadprog function, *II:70*
- quadratic optimization with,
III:441–444
- random number generation,
III:444
- for simulations, *III:651*
- Sobol sequences in, *III:445–446*
- for stable distributions, *III:344*
- surf function in, *III:432–433*
- syntax of, *III:426–427*
- toolboxes in, *III:417–418*
- user-defined functions in,
III:423–427
- Matrices
augmented, *II:624*
characteristic polynomial of, *II:628*
coefficient, *II:624*
companion, *II:639–640*
defined, *II:622*
diagonal, *II:622–623, II:640*
eigenvalues of random, *II:704–705*
eigenvectors of, *II:640–641*
in MATLAB, *III:422, III:432*
operations on, *II:626–627*
ranks of, *II:623, II:628*
square, *II:622–623, II:626–627*
symmetric, *II:623*
traces of, *II:623*
transition, *III:32–33, III:32t, III:33t, III:35f*
types of, *II:622, II:628*
- Matrix differential equations, *III:492*
- Maturity value (lump sum), from
bonds, *I:211*
- Maxima, *III:265–269, III:266f*
- Maximum Description Length
principle, *II:703*
- Maximum eigenvalue test, *II:392–393*
- Maximum likelihood (ML)
approach, *I:141, I:348*
methods, *II:348–349, II:737–738, III:273*
principal, *II:312*
- Maximum principle, *II:662, II:667*
- Max-stable distributions, *III:269, III:339–340*
- MBA (Mortgage Bankers Association)
refi index, *III:70, III:70f*
- MBS (mortgage-backed securities),
I:258
agency vs. nonagency, *III:48*
cash flow characteristics of, *III:48*
default assumptions about, *III:8*
negative convexity of, *III:49*
performance of, *III:74*
prices of, *III:26*
projected long-term performance of,
III:34f
time-related factors in, *III:73–74*
valuation of, *III:62*
valuing of, *III:645*
- MBS (mortgage-backed securities),
nonagency
analysis of, *III:44–45*
defined, *III:48*
estimation of returns, *III:36–44*
evaluation of, *III:29*
factors impacting returns of,
III:30–32
yield tables for, *III:41t*
- Mean absolute deviation (MAD),
III:353
- Mean absolute moment (MAM(q)),
III:353
- Mean colog (M-colog), *III:354*
- Mean entropy (M-entropy), *III:354*
- Mean excess function, *II:746–747*
- Mean/first moment, *III:201–202*
- Mean residual life function, *II:754n*
- Mean reversion
discussion of, *I:88–92*
geometric, *I:91–92*
in HW models, *III:605*
and market stability, *III:537–538*
models of, *I:97*
parameter estimation, *I:90–91*
risk-neutral asset model, *III:526*
simulation of, *I:90*
in spot rate models, *III:580*
stabilization by, *III:538*
within a trinomial setting, *III:604*
- Mean-reverting asset model (MRAM),
III:525–526
- Means, *I:148, I:155, I:380, III:166–167*
- Mean-variance
efficiency, *I:190–191*
efficient portfolios, *I:13, I:68, I:69–70*

- Mean-variance (*Continued*)
 nonrobust formulation, *III:139–140*
 optimization, *I:192*
 constraints on, *I:191*
 estimation errors and, *I:17–18*
 practical problems in, *I:190–194*
 risk aversion formulation, *II:70*
 Mean variance analysis, *I:3, I:15f, I:201, II:471–472, III:352*
- Measurement levels, in descriptive statistics, *II:486–487*
- Media effects, *III:70*
- Median, *I:155, I:159n, II:40*
- Median tail loss (MTL), *III:341*
- Mencken, H. L., *II:57*
- Menger, Carl, *II:468*
- Mercurio-Moraleda model, *I:493–494*
- Merton, Robert, *I:299, I:310, II:468, II:475, II:476*
- Merton model
 advantages and criticisms of, *I:344*
 applied to probability of default, *I:363–365*
 with Black-Scholes approach, *I:305–306*
 default probabilities with, *I:307–308*
 discussion of, *I:343–344*
 drawbacks of, *I:410*
 with early default, *I:306*
 evidence on performance, *I:308–309*
 as first modern structural model, *I:313, I:341*
 in history, *I:491*
 with jumps in asset values, *I:306*
 portfolio-level hedging with, *I:411–413*
 with stochastic interest rates, *I:306*
 and transaction-level hedging, *I:408–410*
 usefulness of, *I:410, I:411–412, I:417–418*
 use of, *I:304, I:305, I:510*
 variations on, *I:306–307*
- Methodology, equally weighted, *III:399*
- Methods
 quantile, *II:354–356*
- Methods pathwise, *III:643*
- Metropolis-Hastings (M-H) algorithm, *I:178*
- M-H algorithm, *I:179*
- MIB 30, *III:402–403, III:402f, III:403f*
- Microsoft, *II:722f*. *See also* Excel
- Midsquare technique, *III:647*
- Migration mode
 calculation of expected/unexpected losses under, *I:376t*
 expected loss under, *I:373–374*
- Miller, Merton, *II:467, II:473*
- MiniMax (MM) risk measure, *III:356*
- Minimization problems, solutions to, *II:683–684*
- Minimum-overall-variance portfolio, *I:69*
- Minority interest, on the balance sheet, *II:536*
- Mispricing, risk of, *II:691–692*
- Model creep, *II:694*
- Model diagnosis, *III:367–368*
- Model estimation, in non-IDD framework, *III:278*
- Modeling
 calibration of structure, *III:549–550*
 changes in mathematical, *II:480–481*
 discrete *vs.* continuous time, *III:562*
 dynamic, *II:105*
 issues in, *II:299*
 nonlinear time series, *II:427–428, II:430–433*
 quantitative, *II:481*
- Modeling techniques
 non-parametric/nonlinear, *II:375*
- Model risk
 of agency ratings, *II:728–729*
 awareness of, *I:145, II:695–696*
 with computer models, *II:695*
 consequences of, *II:729–730*
 contribution to bond pricing, *II:727–728*
 defined, *I:331, II:691, II:697*
 discussion of, *II:714–715*
 diversification of, *II:378*
 endogenous, *II:694–695, II:697*
 in financial institutions, *II:693*
 guidelines for institutions, *II:696–697*
 management of, *II:695–697, II:697*
 misspecification of, *II:199*
 and robustness, *II:301*
 of simple portfolio, *II:721–726*
 sources of, *II:692–695*
- Models. *See also* operational risk models
 accuracy in, *III:321*
 adjustment, *II:502*
 advantages of reduced-form, *I:533*
 analytical tractability of, *III:549–550*
 APD, *III:18, III:20–22, III:21f, III:26*
 application of, *II:694*
 appropriate use of classes of, *III:597–598*
 arbitrage-free, *III:600*
 autopredictive, *II:502*
 averages across, *II:715*
 bilinear, *II:403–404*
 binomial, *I:114–116, I:119*
 binomial stochastic, *II:10–11*
- block maxima, *II:745*
 choosing, *III:550–552*
 comparison of, *III:617*
 compatibility of, *III:373*
 complexity of, *II:704, II:717*
 computer, *I:511, II:695*
 conditional normal, *II:733–734*
 conditional parametric fat-tailed, *II:744*
 conditioning, *II:105*
 construction of, *II:232–235*
 for continuous processes, *I:123*
 creation of, *II:100–102*
 cross-sectional, *II:174–175, II:175t*
 cumulative return of, *II:234*
 defined, *II:691, II:697*
 to describe default processes, *I:313*
 description and estimation of, *II:256–257*
 designing the next, *III:590–591*
 determining, *II:299–300*
 disclosure of, *I:410*
 documentation of, *II:696*
 dynamic factor, *II:128, II:131, III:126–127*
 dynamic term structure, *III:591*
 econometric, *II:295, II:304*
 equilibrium forms of, *III:599–600*
 equity risk, *II:174, II:178–191, II:192*
 error correction, *II:381t, II:387–388, II:394–395*
 evidence of performance, *I:308–309, II:233*
 examples of multifactor, *II:139–140*
 financial, *I:139, II:479–480*
 forecasting, *II:112, II:303–304*
 for forecasting, *III:411*
 formulation of, *III:128–131*
 fundamental factor, *II:244, II:248*
 generally, *II:360–362*
 Gordon-Shapiro, *II:17–18*
 Heath-Jarrow-Morton, *III:586–587, III:589*
 hidden-variable, *II:128, II:131*
 linear, *II:264, II:310–311, II:348, II:507–508*
 linear autoregressive, *II:128, II:130–131*
 linear regression, *I:91, I:163–170, II:360, II:414–415*
 liquidation process, *I:342*
 martingale, *II:127–128, III:520–521*
 MGARCH, *II:371–372*
 model-vetting procedure, *II:696–697*
 moving average, *III:414*
 multifactor, *II:231–232, III:92*
 multivariate extensions of, *II:370–373*
 no arbitrage, *III:604*

- nonlinear, *II*:402–421, *II*:417–418
 penalty functions in, *II*:703
 performance measurement of, *II*:301
 predictive regressive, *II*:130
 predictive return, *II*:128–131
 for pricing, *II*:127–128
 pricing errors in, *I*:322
 principals for engineering,
 II:482–483
 probabilistic, *II*:299
 properties of good, *I*:320
 ranking alternative, *III*:368–370
 recalibration of, *II*:713–714
 reduced form default, *I*:310, *I*:313
 regressive, *II*:128, *II*:129–130
 relative valuation, *I*:260
 return forecasting, *II*:119
 returns of, *II*:233*t*
 robustness of, *II*:301
 selection of, *I*:145, *II*:298, *II*:692–693,
 II:699–701
 short-rate, *I*:494
 single-index market, *II*:317–318
 static, *II*:297, *III*:573
 static regressive, *II*:129–130
 static *vs.* dynamic, *II*:295–296, *II*:304
 statistical, *II*:175, *II*:175*t*
 stochastic, *I*:557, *III*:124–125
 structural, *I*:305, *I*:313–314, *I*:341–342
 structural *vs.* reduced, *I*:532–533
 subordinated, *II*:742–743
 temporal aggregation of, *II*:369
 testing of, *II*:126–127, *II*:696–697
 time horizon of, *II*:300–301
 time-series, *II*:175, *II*:175*t*
 tree, *II*:381, *III*:22–23
 tuning of, *III*:580–581
 two-factor, *I*:494
 univariate regression, *I*:165
 usefulness of, *II*:122
 use of in practice, *I*:494–496, *III*:600*t*
- Models, lattice**
 binomial, *III*:610, *III*:610*f*
 Black-Karasinski (BK) lattice, *III*:611
 Hull White binomial, *III*:610–611
 Hull White trinomial, *III*:613
 trinomial, *III*:610, *III*:610*f*,
 III:611–612
- Models, selection of**
 components of, *II*:717
 generally, *II*:715–717
 importance of, *II*:700
 machine learning approach to,
 II:701–703, *II*:717
 uncertainty/noise in, *II*:716–717
 use of statistical tools in, *II*:230
- Modified Accelerated Cost Recovery
 System (MACRS), *II*:538**
- Modified Restructuring clause, *I*:529**
- Modified tempered stable (MTS)
 processes, *III*:513**
- Modigliani, Franco, *II*:467, *II*:473
- Modigliani-Miller theorem, *I*:343,
 I:344, *II*:473, *II*:476
- Moment ratio estimators, *III*:274
- Moments**
 exponential, *III*:255–256
 first, *III*:201–202
 of higher order, *III*:202–205
 integration of, *II*:367–368
 raw, *II*:739
 second, *III*:202
 types of, *II*:125
- Momentum**
 formula for analysis of, *II*:239
 portfolios based on, *II*:181
- Momentum factor, *II*:226–227
- Money, future value of, *II*:596–600
- Money funds, European options on,
 I:498–499
- Money markets, *I*:279, *I*:282, *I*:314,
 II:244
- Monotonicity property, *III*:327
- Monte Carlo methods**
 advantages of, *II*:672
 approach to estimation, *I*:193
 defined, *I*:273
 examples of, *III*:637–639
 foundations of, *I*:377–378
 for interest rate structure, *I*:494
 main ideas of, *III*:637–642
 for nonlinear state-space modeling,
 II:417–418
 stochastic content of, *I*:378
 usefulness of, *I*:389
 use of, *I*:266–268, *III*:651
 of VaR calculation, *III*:324–325
- Monte Carlo simulations**
 for credit loss, *I*:379–380
 effect of sampling process, *I*:384
 in fixed income valuation modeling,
 III:6–12
 sequences in, *I*:378–379
 speed of, *III*:644
 use of, *III*:10–11, *III*:642
- Moody's diversity score, use of,
 I:332
- Moody's Investors Service, *I*:362
- Moody's KMV, *I*:364–365
- Mortgage-backed securities (MBS). *See*
 MBS (mortgage-backed
 securities)
- Mortgage Bankers Association (MBA)
 method, *III*:57–58
- Mortgagee pools
 composition of, *III*:52
 defined, *III*:23, *III*:65
 nonperforming loans and, *III*:75
- population of, *III*:19
 seasoning of, *III*:20, *III*:22
- Mortgages, *III*:48–49, *III*:65, *III*:69,
 III:71
- Mosaic Company, distribution of price
 changes of, *II*:723*f*
- Mossin, Jan, *II*:468, *II*:474
- Moving averages, infinite, *II*:504–508
- MSCI Barra model, *II*:140
- MSCI EM, historical distributions of,
 III:391*f*
- MSCI-Germany Index, *I*:143
- MSCI World Index, *I*:15–17
 analysis of 18 countries, *I*:16*t*
- MS GARCH model, *I*:185–186
 estimation of, *I*:182
 sampling algorithm for, *I*:184
- MSR (maximum Sharpe Ratio), *I*:36–37
- MS-VAR models, *II*:131
- Multiaccount optimization, *II*:75–77
- Multicollinearity, *II*:221
- Multilayer perceptrons, *II*:419
- Multinomial/polynomial coefficients,
 III:191–192
- Multivariate normal distribution, in
 MATLAB, *III*:432–433, *III*:433*f*
- Multivariate random walks, *II*:124
- Multivariate stationary series,
 II:506–507
- Multivariate *t* distribution, loss
 simulation, *I*:388–389
- Nadaraya-Watson estimator, *II*:412,
 II:415
- Natural conjugate priors, *I*:160*n*
- Navigation, fuel-efficient, *I*:562–563
- Near-misses, management of,
 III:84–85
- Net cash flow, defined, *II*:541
- Net cost of carry, *I*:424–425, *I*:428,
 I:437, *I*:439–440, *I*:455
- Net free cash flow (NFCF), *II*:572–574,
 II:578
- Net profit margin, *II*:556
- Net working capital-to-sales ratio,
 II:554–555
- Network investment models,
 III:129–130, *III*:129*f*
- Neumann boundary condition, *II*:666,
 II:671
- Neural networks, *II*:403, *II*:418–421,
 II:418*f*, *II*:701–702
- Newey-West corrections, *II*:220
- NIG distribution, *III*:257*n*
- 9/11 attacks, effects of, *III*:402–403
- No-arbitrage condition, in certain
 economy, *III*:567–568
- No arbitrage models, use of, *III*:604
- No-arbitrage relations, *I*:423

- Noise
 continuous-time, *III:486*
 in financial models, *II:721–722*
 in model selection, *II:716–717*
 models for, *II:726*
 reduction of, *II:51–52*
- Noise, white
 defined, *I:82, II:297*
 qualities of, *II:127*
 sequences, *II:312, II:313*
 in stochastic differential equations,
III:486
 strict, *II:125*
vs. colored noise, *III:275*
- Nonlinear additive AR (NAAR)
 model, *II:417*
- Nonlinear dynamics and chaos, *II:645, III:652–654*
- Nonlinearity, *II:433*
 in econometrics, *II:401–403*
 tests of, *II:421–427*
- Non-normal probability distributions,
II:480
- Nonparametric methods, *II:411–416*
- Normal distributions, *I:81, I:82f, I:177–178, III:638f*
 and AVaR, *III:334*
 comparison with α -stable, *III:234f*
 fundamentals of, *II:731–734*
 inverse Gaussian, *III:231–233, III:232f, III:233f* (*See also* Gaussian distribution)
 likelihood function, *I:142–143*
 for logarithmic returns, *III:211–212*
 mixtures of for downside risk estimation, *III:387–388*
 for modeling operational risk, *III:98–99*
 multivariate, and tail dependence, *I:387*
 properties of, *II:732–733, III:209–210*
 relaxing assumption of, *I:386–387*
 standard, *III:208*
 standardized residuals from, *II:751*
 use of, *II:752n*
 using to approximate binomial distribution, *III:211*
 for various parameter values, *III:209f*
vs. normal inverse Gaussian distribution, *III:232–233*
- Normal mean, and posterior tradeoff,
I:158–159
- Normal tempered stable (NTS)
 processes, *III:513*
- Normative theory, *I:3*
- Notes, step-up callable, *I:251–252, I:251f, I:252f*
- Novikov condition, *I:131–132*
- NTS distribution, *III:257n*
- Null hypothesis, *I:157, I:170, III:362*
- Numeraire, change of, *III:588–589*
- Numerical approximation, *I:265*
- Numerical models for bonds,
I:273–275
- OAS (option-adjusted spread). *See* option-adjusted spread
- Obligations, deliverable, *I:231, I:526*
- Observations, frequency of, *III:404*
- Occam's razor, in model selection,
II:696
- Odds ratio, posterior, *I:157*
- Office of Thrift Supervision (OTS)
 method, *III:57–58*
- Oil industry, free cash flows of, *II:570*
- OLS (ordinary least squares). *See* ordinary least squares (OLS)
- Open classes, *II:493–494*
- Operating cash flow (OCF), *II:23*
- Operating cycles, *II:551–554*
- Operating profit margin, *II:556*
- Operational loss data
 de Fontnouvelle, Rosengren, and Jordan study, *III:116–117, III:116t*
 empirical evidence with, *III:112–118*
 Moscadelli study, *III:113, III:116, III:116t*
 Müller study, *III:113, III:114f, III:115t*
 Reynolds-Syer study, *III:117–118*
 Rosenberg-Schuermann study, *III:118*
- Operational losses
 and bank size, *III:83*
 definitions of types, *III:84t*
 direct *vs.* indirect, *III:84–85*
 expected *vs.* unexpected, *III:85*
 histogram of, *III:104f*
 histogram of severity distribution, *III:95f*
 historical data on, *III:96*
 near-miss, *III:84–85*
 process of arriving at data, *III:96–97*
 process of occurrence, *III:86f*
 recording of, *III:97*
 severity of, *III:104f*
 time lags in, *III:96–97*
 types of, *III:81, III:88*
- Operational loss models
 approaches to, *III:103–104*
 assumptions in, *III:104*
 nonparametric approach,
III:103–104, III:104–105, III:118
 parametric approach, *III:104, III:105–110, III:118*
 types of, *III:118*
- Operational risk
 classifications of, *III:83–88, III:87–88, III:87f, III:88*
 defined, *III:81–83, III:88*
 event types with descriptions,
III:86t
 indicators of, *III:83*
 models of, *III:91–96*
 nature of, *III:99*
 and reputational risk, *III:88*
 sources of, *III:82*
- Operational risk/event/loss types,
 distinctions between, *III:85–87*
- Operational risk models
 actuarial (statistical) models, *III:95*
 bottom-up, *III:92f, III:94–96, III:99*
 causal, *III:94*
 expense-based, *III:93*
 income-based, *III:93*
 multifactor causal models, *III:95*
 operating leverage, *III:93*
 process-based, *III:94–95*
 proprietary, *III:96*
 reliability, *III:94–95*
 top down, *III:92–94, III:99*
 types of, *III:91–92*
- Operations
 addition, *II:625, II:626*
 defined, *II:628*
 inverse and adjoint, *II:626–627*
 multiplication, *II:625–626, II:626*
 transpose, *II:625, II:626*
 vector, *II:625–626*
- Operators in sets, defined, *III:154*
- Ophelimity, concept of, *II:469*
- Opportunity cost, *I:435, I:438, I:439, II:596, III:623*
- Optimal exercise, *I:515–516*
- Optimization
 algorithms for, *III:124*
 complexity of, *II:82*
 constrained, *I:28–34*
 defined, *III:434–435*
 local *vs.* global, *II:378*
 in MATLAB, *III:434–444*
 unconstrained, *I:22–28*
- Optimization theory, *I:21*
- Optimization Toolbox, in MATLAB,
III:435–436, III:436f
- Optimizers, using, *II:115–116, II:483*
- Option-adjusted spread (OAS)
 calculation of, *I:253–255*
 defined, *I:254, III:11*
 demonstrated, *I:254f*
 determination of, *I:259*
 implementation of, *I:257*
 and market value, *I:258*
 results from example, *III:617t*
 and risk factors, *III:599*

- rules-of-thumb for analysis, I:264–265
- usefulness of, III:3
- values of, I:267, I:268
- variance between dealers, I:257–258
- Option premium, I:508–509
 - time/intrinsic values of, I:513
- Option premium profiles, I:512, I:512f
- Option prices
 - components of, I:484–485, I:511–512
 - factors influencing, I:486–487, I:486t, I:487–488, I:522–523
 - models for, I:490
- Options
 - American, II:664–665, II:669–670, II:674–679, II:679–681
 - American-style, I:444, I:454–455, I:490
 - Asian, II:663–664, II:668–669, II:642–643
 - on the average, II:663–664
 - barrier, II:662–663
 - basic properties of, I:507–508
 - basket, II:662, II:672
 - Bermudean, II:663–664, III:597
 - buying assets of, I:439
 - costs of, I:441–442, III:11–12
 - difference from forwards, I:437–439
 - early exercise of, I:442–443, I:447
 - Eurodollar, I:489
 - European, I:125, I:127–129, II:660–664, II:665–674
 - European-style, I:444–445, I:454
 - European-style *vs.* American-style, I:453t, I:455n, I:508, I:515–516
 - and expected volatility, I:486
 - expiration/maturity dates of, I:484
 - factors affecting value of, I:474
 - formulas for pricing, III:522, III:527
 - in/out of/at-the-money, I:485
 - long *vs.* short call, I:437–439, I:438f
 - lookback, II:663, II:672, II:673f
 - on the maximum, II:663
 - models of, I:510–511
 - no-arbitrage futures, I:453
 - price relations for, I:448t
 - pricing of, I:124–129, I:455t, I:484–488, I:507, III:408
 - theoretical valuation of, I:508–509
 - time premiums of, I:485
 - time to expiration of, I:486
 - types of, I:484
 - valuing of, I:252–253, III:639
 - vanilla, II:661, III:655
 - volatility of, I:488
- Orders
 - in differential equations, II:643, II:644–645
 - fleeting limit, III:625
 - limit, III:625, III:631
 - market, III:625, III:631
- Order statistics, III:269–270
 - bivariate, III:293–295
 - joint probability distributions for, III:291–292
 - use of, III:289
 - for VaR and ETL, III:292t
 - in VaR calculations, III:291
- Ordinary differential equations (ODE), II:644–645, II:646–648, II:648–652, II:649f
- Ordinary least squares (OLS)
 - alternate weighting of, II:438–439
 - estimation of factor loadings matrix with, II:165
 - in maximum likelihood estimates, II:313–314
 - pictorial representations of, II:437–438, II:438f
 - squared errors in, II:439–440
 - use of, I:165, I:172n, II:353
 - vs.* Theil-Sen estimates of beta, II:442f
 - vs.* Theil-Sen regression, II:441t
- Ornstein-Uhlenbeck process
 - with change of time, III:523
 - and mean reversion, I:263, I:264f
 - solutions to, III:492
 - use of, I:89, I:95
 - and volatility, III:656
- Outcomes, identification and
 - evaluation of worst-case, III:379–380
- Outliers
 - in data sets, II:200
 - detection and management of, II:206
 - effect of, II:355f, II:442–443
 - and market crashes, II:503
 - in OLS methods, II:354
 - in quantile methods, II:355–356
 - and the Thiel-Sen regression algorithm, II:440
- Out-of-sample methodology, II:238
- Pair trading, II:710
- P-almost surely (P-a.s.) occurring events, III:158
- Parallel yield curve shift assumption, III:12–13
- Parameters
 - calibration of, II:693
 - density functions for values, III:229f, III:230f, III:231f
 - distributions of, II:721
 - estimation of for random walk, I:83
 - robust estimation of, II:77–78
 - stable, III:246f
- Parametric methods, use of, II:522
- Parametric models, II:522–523, II:526–527
- Par asset swap spreads, I:530, I:531
- Par CDS spread, I:531
- Par-coupon curve, III:561
- Pareto, Vilfredo, II:467, II:468–469, II:474
- Pareto(2) distribution, II:441
- Pareto distributions
 - density function of, II:738
 - generalized (GPD), II:745–746, II:747, III:230–231
 - in loss distributions, III:108–109
 - parameters for determining, II:738
 - stable, II:738–741
 - stable/varying density, II:739f
 - tails of, II:751
- Pareto law, II:469
- Pareto-Lévy stable distribution, III:242
- Partial differential equations (PDEs)
 - for American options, II:664–665
 - equations for option pricing, II:660–665
 - framework for, I:261, I:265, II:675, III:555
 - pricing European options with, II:665–674
 - usefulness of, II:659–660
 - use of, III:18–19
- Partitioning, binary recursive, II:376–377, II:376f
- Paths
 - in Brownian motion, III:501, III:502f
 - dependence, III:18–19
 - stochastic, II:297
- Payments, I:229, II:611–612
- Payment shock, III:72
- Payoff-rate process, I:121–122
- Payoffs, III:466, III:638–639
- PCA (principal components analysis). *See* principal component analysis (PCA)
- Pearson skewness, III:204–205
- Pension funds, constraints of, II:62
- Pension plans, II:541, III:132
- P/E (price/earnings) ratio, II:20–21, II:38
- Percentage rates, annual *vs.* effective, II:615–617
- Percolation models, III:276
- Performance attribution, II:57, II:58, II:104, II:188–189, II:252–253, II:253t
- Performance-seeking portfolios (PSPs), I:36, I:37
- Perpetuities, II:607–608
- Pharmaceutical companies, II:7–8, II:11, II:244

- Phillips-Perron statistic, *II:386, II:398*
- Pickand-Balkema-de Haan theorem, *II:746*
- Pickand estimator, *III:273*
- Pliska, Stankey, *II:476*
- Plot function, in MATLAB, *III:428–432*
- P-null sets, *III:197*
- Pochhammer symbol, *III:256*
- Poincaré, Henri, *II:469*
- POINT®
- features of, *II:193n, II:291n*
 - modeling with, *II:182*
 - screen shot of, *II:287f, II:288f*
 - use of, *II:179, II:189, II:286–287*
- Point processes, *III:270–272*
- Poisson-Merton jump process,
 - distribution tails for, *III:540–541*
- Poisson-Merton jump variable, *III:540*
- Poisson processes
 - compounded, *III:497*
 - homogeneous, *III:270–271*
 - and jumps, *I:93, III:498, III:540*
 - for modeling durations, *II:461*
 - as stochastic process, *III:496, III:497, III:506*
 - use of, *I:262, I:315–316*
- Poisson variables, distribution of, *III:271f*
- Policy iteration algorithm (Howard algorithm), *II:676–677*
- Polyhedral sets, *I:33, I:33f*
- Polynomial fitting of trend stationary process, *II:702–703, II:702f*
- Population profiles, in transition matrices, *III:32–34*
- Portfolio allocation, example using MATLAB, *III:436–441*
- Portfolio management
 - approaches to, *II:108–110*
 - checklist for robust, *III:144*
 - for credit risk, *I:416–417*
 - of large portfolios, *III:325*
 - and mean-variance framework, *I:196*
 - real world, *I:190*
 - software for, *II:75 (See also Excel)*
 - tax-aware, *II:74–75*
 - using Bayesian techniques, *I:196*
- Portfolio managers, *III:444–445*
 - approaches used by, *II:108–109*
 - enhanced indexers, *II:268*
 - example of, *III:436–441, III:437t*
 - questions considered by, *II:277*
 - specialization of, *II:48–49*
 - traditional vs. quantitative, *II:109, II:110t*
 - types of, *II:179, II:286*
- Portfolio optimization
 - for American options, *II:678*
 - classical mean-variance problem, *III:441–444*
 - constraints on, *II:62*
 - defined, *I:36*
 - formulation of theory, *II:476*
 - max-min problem, *III:139*
 - models of, *II:84–85n*
 - robust, *III:146*
 - techniques of, *II:115–116*
 - uncertainty in, *I:192–193, II:82–83*
- Portfolios. *See* constraints, portfolio allocation of, *I:192–193, II:72*
 - assessment of risk factors of, *III:637–638*
 - benchmark, *I:41–42, II:180*
 - building efficient, *II:115*
 - bullet vs. barbell, *III:308t, III:309t*
 - bullet vs. barbell (hypothetical), *III:308*
 - cap-weighted, *I:38f*
 - centering optimal, *I:199*
 - considerations for rebalancing of, *II:75*
 - construction of, *I:37–38, II:56–57, II:102–104, II:102f, II:114–116, II:179–184, II:261–264, II:286–287, II:301–303*
 - cor-plus, and DTS, *I:398*
 - credit bond, hedging of, *I:405*
 - data on, *II:365t*
 - diversification of, *I:10–12*
 - efficient, *I:12, I:77, I:288f, I:289f, I:290f*
 - efficient set of, *I:13*
 - efficient vs. feasible, *I:13*
 - efficient vs. optimal, *I:5*
 - examples of, *II:261t, II:262t*
 - expected returns from, *I:6–7, I:7, I:12t, I:69t, I:195*
 - factor exposures in, *II:183t, II:184t, II:263t, II:264t*
 - factor model approach to, *II:224*
 - feasible and efficient, *I:12–14*
 - feasible set of, *I:12–13, I:13f*
 - index-tracking, *II:186*
 - information content of, *I:192*
 - long-short, *II:181–182, II:226f*
 - management of fixed-income, *I:391*
 - and market completeness, *I:50–52*
 - mean-variance efficient, *I:66, I:69f*
 - mean-variance optimization of, *II:79*
 - momentum, *II:182f*
 - monitoring of, *II:106*
 - MSR (maximum Sharpe Ratio), *I:36–37*
 - normalized, *II:157*
 - optimal, *I:14–15, I:14f, I:15–17, II:181t*
 - optimization-based approach to, *II:224–225*
 - optimization of, *I:17–18, I:40, II:56–57, II:301–303*
 - optimized, *II:116*
 - performance-seeking, *I:36*
 - quadratic approximation for value, *III:644–645*
 - rebalancing of, *II:287–288*
 - replication of, *II:476*
 - resampling of, *I:189, II:78–80, II:84*
 - returns of, *I:6–7*
 - risk control in, *II:181–182*
 - riskless, *I:509*
 - with risky assets, *I:12–17*
 - robust optimization of, *II:80–84*
 - rule-based, *II:116*
 - selection of, *I:3–19, III:351–353, III:356*
 - self-financing, *II:660–661*
 - stress tests for, *I:412*
 - tangency, *I:36–37*
 - tilting of, *II:263–264*
 - tracking, *II:187t*
 - weighting in, *I:50–51, II:64–65*
 - weights of, *I:191–192*
 - yield simulations of, *I:284–285*
- Portfolio sorts
 - based on EBITDA/EV factor, *II:216–217, II:216f*
 - based on revisions factor, *II:217–218, II:217f*
 - based on share repurchase factor, *II:218, II:218f*
 - information ratios for, *II:219*
 - results from, *II:225f*
 - use of, *II:214–219*
- Portfolio trades, arbitrage, *I:440t*
- Position distribution and likelihood function, *I:142–143*
- Positive homogeneity property, *III:327–328*
- Posterior distribution, *I:159, I:165*
- Posterior odds ratio, *I:157*
- Posterior tradeoff, and normal mean, *I:158–159*
- Power conditional value at risk measure, *III:356*
- Power law, *III:234–235*
- Power plants/refineries, valuation and hedging of, *I:563*
- Power sets, *III:156, III:156t*
- Precision, *I:158, II:702*
- Predictability, *II:122–127*
- Predictions, *I:167, II:124*
- Predictive return modes, adoption of, *II:128–129*

- Preferred habitat hypothesis, III:569–570
- Prepayments
 burnout, III:19
 calculating speeds of, III:50–56
 in cash-flow yields, III:4
 conditional rate of (CPR), III:30, III:50–51, III:58–59
 defaults and involuntary, III:59, III:74–77
 defined, III:50
 disincentives for, III:7–8
 drivers of, III:77
 effect of time on rates of, III:73–74
 evaluation of, III:62
 factors influencing speeds of, III:69–74
 fundamentals of, III:66–69
 for home equity securities, III:55–56
 interactions with defaults, III:76–77
 interest rate path dependency of, III:6
 lag in, III:24–25
 levels of analysis, III:50
 lock-ins, III:73
 modeling of, I:258, I:267, I:268, III:63n, III:598–600
 practical interpretations of, III:20
 rates of, III:74
 reasons for, III:48
 risk of, II:281, II:281t
 S-curves for, III:67–68, III:67f
 sources of, III:23–24
 voluntary, III:38
 voluntary *vs.* involuntary, III:30, III:75–76
- Prepay modeling, III:19–20
 rational exercise, III:25
- Present value, I:268n, II:19, II:603–604, II:609, III:9–10
- Price/earnings (P/E) ratio, II:20–21, II:38
- Price patterns, scaling in, III:279
- Price processes, bonds, I:128
- Prices
 bid/ask, III:625
 Black-Scholes, II:673–674
 changes in, II:722f, II:723f, II:742, III:305–306, III:305t
 compression of, III:303
 computing clean, I:214–215
 dirty, I:382
 distribution of, I:510
 estimating changes in bond, I:373–374
 flexible and sticky in CPI basket, I:292
 formula for discounted, I:110
 marked-to-market, I:430
 modeling realistic, I:93–94
 natural logarithm of, I:85
 path-dependent, III:193n
 strike, I:484–485, I:486
 truncation of, III:304
vs. value, I:455n
- Price time series, autocorrelation in, III:274
- Pricing
 backward induction, III:18
 formulas for relationships, I:105–110
 grids for, III:18–19
 linear, I:52–55
 models for, II:127–128
 rational, I:53
 risk-neutral, I:533, I:544
 rule representation, I:260–261
 use of trees, III:22–23
- Principal component analysis (PCA)
 compared to factor analysis, II:166–168
 concept of, II:157
 defined, II:147, II:276
 discussed, II:157–164
 illustration of, II:158–163
 with stable distributions, II:163–164
 usefulness of, II:158
 use of, I:39–40, II:142, II:168–169
- Principal components, defined, II:148, II:159
- Principal components analysis (PCA), I:556
- Prior elicitation, informative, I:152–153, I:159
- Prior precision, I:158
- Priors, I:153, I:165–167, I:168, I:171–172
- Probabilistic decision theory, II:719–721, II:729
- Probabilities
 in Bayesian framework, I:140, I:144, I:146–148
 conditional, I:117, II:517–518, III:477
 formulas for conditional, I:108t
 interpretation of, II:123
 in models, II:299
 posterior, I:140, I:144
 prior, I:140, I:144
 prior beliefs about, I:147
 realistic, III:596–597
 as relative frequencies, III:152
 risk-adjusted, I:264
 risk neutral, I:58–59, I:59, I:102, I:104, I:111–114, I:115–116, I:117, III:594–596
- Probability density function (PDF), III:384–385
- Probability distributions
 binomial, III:186t
 continuous, III:578
 for drawing black balls, III:176–177
 inverting the cumulating, III:646
 for prepayment models, III:598
 for rate of return, I:7t, I:9t
 use of, III:638, III:645–646
- Probability-integral transformation (PIT), III:365
- Probability law, III:161
- Probability measures, III:157–159, III:594–597
- Probability of default (PD). *See* default probabilities
- Probability theory, II:133, II:700–701
- Probit regression models, II:348–349, II:350
- Processes
 absolute volatility of, III:474
 exponential, III:498
 martingale, I:119, I:262–263, III:509, III:517
 non-decreasing, III:503–505
 normal tempered stable, III:504–505
 predictable, II:132–133
 subordinated, III:387–388
 weakly stationary, II:360–361
- Process maps, III:94
- Proctor & Gamble, cash flows of, II:567–568, II:568t, II:571–573, II:573t
- Product transitions, III:66, III:71–73
- Profit, riskless, I:480
- Profitability ratios, II:555–557, II:563
- Profit margin ratios, II:555–556
- Profit opportunities, I:261
- Programming, linear, I:29, I:32–33
- Programming, stochastic
 defined, III:123–124
 in finance, III:125–126
 general multistage model for
 financial planning, III:128–132
 use of scenario trees in, III:131–132
vs. continuous-time models, III:127–128
vs. other methods in finance, III:126–128
- Projected successive over relaxation (PSOR) method, II:677
- Projections, as-was, usefulness of, II:38
- Propagation effect, III:351
- Prospectus prepayment curve (PPC), III:54–55, III:56
- Protection, buying/selling of, I:230–231
- 100 PSA (Public Securities Association prepayment benchmark), III:51–52, III:55
- Pseudo-random numbers, generation of, III:647

- PSPs (performance-seeking portfolios), *I:36, I:37*
- Public Securities Association (PSA) prepayment benchmark, *III:51–55, III:51f, III:62–63*
- Pull to par value, *I:216*
- Pure returns, *II:51*
- Put-call parity, *I:437*
for American-style options, *I:446–448, I:452–453, I:452t*
for European options, *I:499*
for European-style options, *I:444–446, I:445t, I:451, I:451t*
perfect substitutes in European-style, *I:445t*
relations of, *I:446*
- Put-call parity relationship, *I:445, I:446, I:485*
- Put options, *I:439*
- Puts, American-style
early exercise of, *I:444, I:450–451*
error on value of, *II:677t, II:678t*
lower price bound, *I:443–444, I:450*
numerical results for, *II:677–678*
- Puts, European-style
arbitrage trades, *I:443t*
lower price bound, *I:443, I:450*
- Pyrrho's lemma, *II:330, II:331*
- Q-statistic of squared residuals, *II:422*
- Quadratic objective, two-dimensional, *I:29f*
- Quadratic programming, *I:29, I:33–34*
- Quadratic variation, *III:474*
- Quantiles
development of regression, *II:356*
methods, *II:354–356*
plot (QQ-plot) of, *III:272*
use of regression, *II:353–354, II:356–357*
- Quantitative methods, *II:483*
- Quantitative portfolio allocation, use of, *I:17–18*
- Quantitative strategies, backtesting of, *I:201*
- Quintile returns, *II:97–98*
- Quotes
delayed, *II:454*
discrepancies in, *II:453–454*
histograms from simple returns, *II:458f*
methods for sampling, *II:457–460*
mid-quote closing, *II:460f*
mid-quote format, *II:456*
mid-quote time-interpolated, *II:460f*
quantile plots of, *II:459f, II:461f*
- R^2 , adjusted, *II:315–316*
- Radon-Nikodym derivative, *I:111, I:130, I:133–134, III:510–511, III:515*
- Ramp, loans on, *III:52*
- Randomized operational time, *III:521*
- Randomness, *I:164, III:534–537, III:580*
- Random numbers
clusters in, *III:649–650*
generation of, *III:645–647*
practicality of, *III:647*
reproducing series of, *III:646*
simulations of, *III:650f*
- Random walks
advanced models of, *I:92–94*
arithmetic, *I:82–84, I:97, II:125*
for Brownian motion, *III:478–479*
computation of, *I:83, I:85, I:87, I:90*
correlated, *I:92–93, II:502–503*
defined, *III:486*
in forecastability, *II:127*
generation of, *I:85*
geometric, *I:84–88, I:89, I:97*
and linear nonstationary models, *II:508*
multivariate, *I:93*
parameters of, *I:87–88*
polynomial fitting of, *II:704f*
simulation of, *I:87*
and standard deviation, *II:385*
500-step samples, *II:708f*
strict, *II:126*
use of, *II:132, III:474*
variables in, *I:83–84*
- Range notes, valuing, *I:252*
- RAS Asset Management, *III:624*
- Rate-and-term refinancing, *III:66*
- Rating agencies, *I:300, III:44*
effect of actions of, *I:367–369*
role of, *I:362*
- Rating migration, *I:362, I:367–369*
- Rating outlooks, *I:365–366*
- Ratings
maturity of, *I:301*
- Ratings-based step-ups, *I:352*
- Rating transitions, *I:368, I:368t, I:381*
- Ratios
analysis of, *II:575–576*
classification of, *II:545–546*
defined, *II:545*
quick (acid test), *II:554*
scales of, *II:487*
- Real estate prices, effect of, *III:44*
- Real yield duration, calculation of, *I:286*
- Receipts, depository, *II:36*
- Recoveries, in foreclosures, *III:75*
- Recovery percentages, *III:30–31*
- Recovery rates
calibration of assumption, *I:537–538*
for captive finance companies, *I:366–367*
and credit risk, *I:362*
dealing with, *I:334n*
on defaulted securities, *I:367t*
drivers of, *I:372*
modeling of, *I:316–317*
random, *I:383*
relationship to default process, *I:372, I:376*
time dimension to, *I:366–377*
- Rectangular distribution, *III:219–221*
- Recursive out-of-sample test, *II:236*
- Recursive valuation process, *I:244*
- Reduced form models, usefulness of, *I:412*
- Redundant assets/securities, *I:51*
- Reference entities, *I:526*
- Reference priors, *I:159–160n*
- Refinancing
and ARMs, *III:72*
categories of, *III:48*
discussion of, *III:68–69*
rate-and-term, *III:68*
speed of, *III:25–26*
threshold model, *III:18*
- Refinancing, paths of rates, *III:8t*
- RefiSMM(Price) function, *III:25–26*
- Regime switching, *I:173n*
- Regression
binary, *III:364*
properties of, *II:309–310*
spurious, *II:384, II:385*
stepwise, *II:331*
- Regression analysis
results for dummy variable regression, *II:348t*
usefulness of, *II:305*
use in finance, *II:316–328*
variables in, *II:330*
- Regression coefficients, testing of, *I:170*
- Regression disturbances, *I:164*
- Regression equations, *II:309–310*
- Regression function, *II:309*
- Regression models, *I:168–169, I:170–172, II:302*
- Regressions
estimation of linear, *II:311–314*
explanatory power of, *II:315–316*
linear, *II:310–311*
and linear models, *II:308–311*
pitfalls of, *II:329–330*
sampling distributions of, *II:314*
spurious, *II:329*
- Regression theory, classical, *II:237*
- Regressors, *II:308–310, II:311, II:330*

- Reg T (Treasury Regulation T), *I:67*
- Relative valuation analysis
 hypothetical example of, *II:40–45*
 hypothetical results, *II:40t*
 implications of hypothetical,
II:41–42
 low or negative numbers in, *II:42–43*
- Relative valuation methods
 choice of valuation multiples in,
II:38–39
 usefulness of, *II:45*
 use of, *II:33–34, II:45*
- Replication, *I:526*
- Reports, *II:200–201, II:283–286*
- Research, process of quantitative,
II:717f
- Residuals, *II:220, II:328–329*
- Restructuring, *I:528–530, I:529, I:529t, I:530, I:537*
- Return covariance matrix formula,
II:141
- Return distributions, *III:333f, III:388–392*
- Return effects, *II:47–48, II:51, II:51f*
- Return generating function, *II:256*
- Return on assets, *II:547–548, II:548–550*
- Return on equity (ROE), *II:37–38, II:41–42, II:548, II:550*
- Return on investment ratios,
II:547–551, II:548, II:563
- Returns
 active, *II:115*
 arithmetic *vs.* geometric average,
II:598
 defined, *II:598*
 estimated moments of, *II:204*
 estimates of expected, *I:190–191*
 ex ante, *I:7*
 excess, *I:66, I:67, I:74*
 expected, *I:71–72, II:13–14, II:112*
 ex post, *I:6*
 fat tails of conditional distribution,
II:753n
 finite variance of, *III:383–384*
 forecasting of, *II:111–112, II:362*
 historical, *II:285f, III:389t*
 monthly *vs.* size-related variables,
II:52t
 naïve, *II:51, II:53f*
 naïve *vs.* pure, *II:52f, II:53–54*
 Nasdaq, Dow Jones, bond, *II:365f*
 pure, *II:51, II:53f, II:54t*
 robust estimators for, *I:40–41*
 rolling 24-month, *II:229f*
 systematic *vs.* idiosyncratic, *II:173*
 time-series properties of, *II:733–734*
- Returns to factors, *II:248*
- Return to maturity expectations
 hypothesis, *III:569*
- Return volatility, excess and DTS,
I:396–397
- Reverse optimization, *I:203n*
- Riemann-Lebesgue integrals, *III:483*
- Riemann-Stieltjes integrals, *I:122, III:473–474, III:487*
- Riemann sum, *II:743–744*
- Risk. *See also* operational risk
 alternative definitions of, *III:350*
 analyzing with multifactor models,
II:184–188
 assessment of, *III:640–641*
 asymmetry of, *III:350–351*
 budgeting of, *II:115, II:286–287*
 of CAPM investors, *I:73–74*
 changes in, *II:368, III:351*
 coherent measures of, *III:327–329*
 collective, *II:470*
 common factor/specific, *II:258*
 controlling, *I:397*
 correlated, *II:271t*
 correlated *vs.* isolated, *II:271*
 counterparty, *I:478, I:479*
 decomposition of, *II:250–253, II:257–261, II:265*
 and descriptors, *II:140*
 downside, *III:382*
 effect of correlation of asset returns
 on portfolio, *I:11–12*
 effect of number of stocks on,
II:249f
 estimation of, *I:40*
 in financial assets, *I:369*
 forecasting of, *II:112–113*
 fundamental, *II:199*
 funding, *II:199*
 horizon, *II:199*
 idiosyncratic, *II:178, II:188, II:188t, II:283, II:285t, II:291*
 idiosyncratic *vs.* systematic, *I:40–41*
 implementation, *II:199*
 including spread in estimation of,
I:399
 indexes of, *II:140, II:256*
 interest rate, *I:521–522, III:4*
 issue specific, *II:283t*
 liquidity, *II:199*
 main sources of, *II:211*
 market price of, *III:579, III:588, III:591*
 model (*See* model risk)
 modeling, *III:11*
 momentum, *II:181t*
 as multidimensional phenomenon,
III:350
 noise trader, *II:199*
 perspective on, *II:91–92*
 portfolio, *I:7–10, I:9–10, I:11, II:180t*
 repayment, *III:48*
 price movement costs, *II:69*
 quantification of, *I:4, I:7–8*
 realized, *II:118*
 reinvestment, *III:4–5*
 relativity of, *III:350*
 residual, *II:258–259*
 by sector, *II:185t*
 in securities, *I:73*
 sources of, *II:173–174, II:251f, II:274, II:281–282*
 systematic, *II:186*
 tail, *I:384, I:385*
 true *vs.* uncertainty, *II:721*
 in a two-asset portfolio, *I:8*
 in wind farm investments, *I:563–564*
- Risk analysis, *II:268–286, II:273t, II:274t, II:275t*
- Risk aversion, *I:404*
 in analysis, *III:570*
 coefficient for, *I:59*
 functions, *III:339f*
 of investors, *I:191*
 and portfolio management, *I:37*
- Risk-based pricing, *III:70*
- Risk decomposition
 active, *II:259–260, II:259f*
 active systematic-active residual,
II:260, II:260f
 insights of, *II:252*
 overview of, *II:261f*
 summary of, *II:260–261*
 systematic-residual, *II:258–259, II:259f*
 total risk, *II:258, II:258f*
- Risk exposures, *I:394, I:521*
- Risk factors
 allocation of, *I:398*
 constraints on, *II:63–64*
 identification of, *II:256*
 macroeconomic, *I:415–416*
 missing, *II:693*
 systematic, *II:268, II:474*
 unsystematic, *II:474*
- Riskiness, determining, *I:145*
- Risk management
 internal models of, *III:289–290*
 in investment process, *II:104*
 portfolio, *III:643–644*
 in portfolio construction, *II:303*
 and quasi-convex functions, *I:28*
- Risk measures, safety-first, *III:352, III:354–356, III:357*
- RiskMetrics™ Group
 approach of, *III:322–323*
 comparison with FTSE100 volatility,
III:413f
 methodology of, *III:412–413*
 software of, *III:413*
 website of, *III:412*

- Risk models
 - applications of, *II:286–290*
 - comparisons among, *II:747–751*
 - defined, *II:692*
 - equity, *II:172–173, II:192–193, II:255, II:264*
 - indicator, *III:93–94*
 - and market volatility, *II:748*
 - multifactor, *II:257–258*
 - principal of, *II:292n*
 - and uncertainty, *II:724*
 - use of, *II:171–172, II:268, II:290*
- Risk neutral, use of term, *III:593–594*
- Risk neutral density (RND)
 - concept of, *II:521*
 - fitting data to models of, *II:526–527*
 - generally, *II:527*
 - parametric models for, *II:523–525*
- Risk oversight, *II:303*
- Risk premiums
 - for default, *III:599*
 - importance of, *III:587*
 - quantifying, *III:580–581*
 - of time value, *I:513*
 - as a variable in discount bond prices, *III:581*
 - variables, *I:403, I:405*
- Risk reports
 - credit risk, *II:278–281*
 - detailed, *II:272–286*
 - factor exposure, *II:275–283*
 - implied volatility, *II:282*
 - inflation, *II:282*
 - issue-level, *II:283–285*
 - liquidity, *II:282*
 - prepayment risk, *II:281*
 - risk source interaction, *II:281–282*
 - scenario analysis, *II:285–286*
 - summary, *II:272–275*
 - tax-policy, *II:282–283*
- Risk tolerance, *II:720–721, II:725, II:729f*
- Risky bonds, investment in, *II:726–729*
- Robot analogy, *III:594*
- Robust covariance matrix, *II:446*
- Robust optimization, *II:83, III:141–142*
- Robust portfolio optimization, *I:17–18, I:193, III:138–142*
 - effect on performance, *III:144*
 - need for research in, *III:145–146*
 - practical considerations for, *III:144–145*
 - in practice, *III:142–144*
- Rolling windows, use of, *II:371*
- Roots
 - complex, *II:632–634, II:636–637*
 - in homogenous difference equations, *II:642*
 - real, *II:630–632, II:635–636*
- Ross, Stephen, *II:468, II:475*
- Rounding, impact of, *III:306n*
- Roy CAPM, *I:67, I:69, I:70*
- Ruin problem, development of, *II:470–471*
- Runge-Kutta method, *II:650–652, II:651f, II:652f*
- Russell 1000, *II:213, II:236–237*
- Saddle points, *I:23, I:23f, I:30*
- Sales, net credit, *II:557–558*
- Samples
 - effect of size, *I:158–159, I:159f, III:407*
 - importance of size, *III:152*
 - and model complexity, *II:703–707*
 - in probability, *III:153*
 - selection of, *II:716*
- Sampling
 - antithetic, *I:383*
 - importance, *I:384, III:648–649*
 - stratified, *II:115, III:648*
- Sampling error, *III:396*
- Samuelson, Paul, *I:556, II:468, II:473–474*
- Sandmann-Sondermann model, *I:493*
- Sarbanes-Oxley Act (2002), *II:542*
- Scalar products, *II:625–626*
- Scale parameters, *I:160n*
- Scaling laws, use of, *III:280*
- Scaling vs. self-similarity, *III:278–280*
- Scenario analysis
 - constraints on, *III:130*
 - factor-based, *II:189–192, II:193*
 - for operational risk, *III:93*
 - usefulness of, *II:179*
 - use of, *II:288–290, III:378*
- Scenarios
 - defined, *III:128*
 - defining, *II:189*
 - generation of, *III:128–132*
 - network representation of, *III:129f*
 - number needed of, *III:640–641*
- Scholes, Myron, *II:468, II:476*
- Schönbucher-Schubert (SS) approach, *I:329–331*
- Schwarz criterion, *II:387, II:389*
- Scorecard Approach, *III:100n*
- Scott model, *II:681–682*
- SDMs (state dependent models), *I:342, I:351–352*
- Secrecy, in economics, *II:716*
- Sector views, implementation of, *II:182–184*
- Securities
 - alteration of cash flows of, *I:210*
 - arbitrage-free value of, *I:261*
 - baskets of, *I:483–484*
 - convertible, *I:462*
 - creating weights for, *II:102–104, II:103f*
 - evaluation of, *I:50*
 - fixed income, *I:209–210, II:268*
 - formula for prices, *I:107*
 - non-Treasury, *I:222–223, I:223t*
 - of other countries, *I:226*
 - payoffs of, *I:49–50, I:116–117, I:121–122*
 - pricing European-style, *III:642*
 - primary, *I:458*
 - primitive, *I:51*
 - private label (*See* MBS (mortgage-backed securities), nonagency)
 - ranking of, *I:200–201*
 - redundant, *I:124*
 - risk-free, *I:115*
 - selection of, *I:225–226*
 - structured, *I:564, I:565–566*
 - supply and demand schedule of, *III:626f*
 - valuing credit-risky, *III:645*
 - variables on losses, *I:370*
- Securities and Exchange Commission (SEC)
 - filings with, *II:532*
- Security levels, two-bond portfolio, *I:382t*
- Selection, adverse vs. favorable, *III:76–77*
- Self-exciting TAR (SETAR) model, *II:405*
- Self-similarity, *III:278–280*
- Selling price, expected future, *II:19–20*
- Semimartingales, settings in change of time, *III:520–521*
- Semi-parametric models
 - tail in, *II:744–747*
- Semiparametric/nonparametric methods, use of, *II:522*
- Semivariance, as alternative to variance, *III:352*
- Sensitivity, *III:643–644*
- Sensitivity analysis, *I:192, II:235*
- Sequences, *I:378, III:649–651, III:650*
- Series, *II:299, II:386, II:507–508, II:512*
- SETAR model, *II:425–426*
- Set of feasible points, *I:28, I:31*
- Set operations, defined, *III:153–154*
- Sets, *III:154*
- Settlement date, *I:478*
- Settlements, *I:526–528*
- Shareholders
 - common, *II:4*
 - equity of, *II:535*
 - negative equity of, *II:42*
 - preferred, *II:4–5*
 - statement of equity, *II:541*

- Shares, repurchases of, *II:207, II:210f, II:211, II:215–216, II:227*
- Sharpe, William, *I:75, II:468, II:474*
- Sharpe-Lintner CAPM (SL-CAPM), *I:66–67, I:75, I:78n*
- Sharpe ratios, *I:40, I:62, I:193*
- Sharpe's single-index model, *I:74–75*
- Shipping options, pricing of, *I:565*
- Shortfall, expected, *I:385–386*
- Short positions, *I:67*
- Short rate models, *III:543–545, III:545–550, III:552–554, III:557, III:604–610*
- Short rates, *III:212–213, III:541, III:549, III:595–596*
- Short selling
 constraints on, *I:67*
 effect of constraints on, *I:17, I:191–192, II:461*
 effect of on efficient frontiers, *I:17f*
 example, *I:480–481*
 as hedging route, *I:409*
 in inefficient markets, *I:71f*
 and market efficiency, *I:70–71*
 net portfolio value, *I:433t*
 and OAS, *I:259*
 and real estate, *II:396–397*
 in reverse cash-and-carry trade, *I:483*
 for terminal wealth positions, *I:460–461*
 using futures, *I:432–433*
- Shrinkage
 estimation of, *I:192, I:194–195, I:201–202, III:142*
 optimal intensity of, *I:202n–203n*
 use of estimators, *II:78*
- δ -algebra, *III:15, III:157*
- δ -fields
 defined, *III:508*
- Signals (forecasting variables), use of
 in forecasting returns, *II:111–112*
 evaluation of, *II:111–112*
- Similarity, selecting criteria for, *II:35*
- Simulated average life, *III:12*
- Simulations
 credit loss, *I:378–380*
 defined, *III:637*
 efficiency of, *I:384*
 financial applications of, *III:642–645*
 process of, *III:638*
 technique of, *III:444–445*
- Single firm models, *I:343–352*
- Single monthly mortality rate (SMM), *III:50–51, III:58*
- Skewness
 defined, *III:238–239*
 and density function, *III:204–205*
 indicating, *III:235*
 and the Student's *t*-distribution, *III:387*
 treatment of stocks with, *I:41*
- Sklar's theorem, *I:326, III:288*
- Skorokhod embedding problem, *III:504*
- Slackness conditions, complementary, *I:32*
- SL-CAPM (Sharpe-Lintner CAPM), *I:66–67, I:75, I:78n*
- Slope elasticity measure, *III:315, III:317*
- Smith, Adam, *II:468, II:472*
- Smoothing, in nonparametric methods, *II:411–412*
- Smoothing constant, *III:409–410*
- Smoothly truncated stable distribution (STS distribution), *III:245–246*
- Smooth transition AR (STAR) model, *II:408–409*
- Sobol sequences, pricing European call options with, *III:445–446*
- Software
 case sensitivity of, *III:434*
 comments in MATLAB code, *III:427*
 developments in, *II:481–482*
 macros in, *III:450–452, III:450f, III:460, III:466*
 pseudo-random number generation, *III:646–647*
 random number generation commands, *III:645–647*
- RiskMetrics Group, *III:413, III:644*
- simulation, *III:651f*
 for stable distributions, *III:344, III:383*
 stochastic programming applications, *III:126*
 use of third party, *II:481*
- Solutions, stability of, *II:652–653*
- Solvers, in MATLAB, *III:435*
- Space in probability, *III:156, III:157*
- Sparse tensor product, *II:673*
- S&P 60 Canada index, *I:550–552, I:550t, I:553f*
- Spearman, Charles, *II:153–154*
- Spearman model, *II:153–154*
- Spearman's rho, *I:327, I:332, I:336n*
- Splits, in recursive partitioning, *II:376–377*
- Spot curves, with key rate shifts, *III:313f, III:314f*
- Spot price models, energy commodities, *I:556–557*
- Spot rates
 arbitrage-free evolution of, *I:557–558*
 bootstrapping of curve, *I:217–220*
 calculation of, *III:581*
 and cash flows in OAS analysis, *I:259*
 changes in, *III:311, III:312f, III:312t*
 computing, *I:219–220*
 under continuous compounding, *III:571*
 defined, *III:595*
 effect of changes in, *I:514, III:313–314, III:314t*
 and forward rates, *III:572*
 models of, *III:579–581*
 paths of monthly, *III:9–10, III:10t*
 theoretical, *I:217*
 Treasury, *I:217*
 uses for, *I:222*
- Spot yields, *III:565, III:566, III:571*
- Spread analysis, *II:290t*
 table of, *II:290t*
- Spread duration, beta-adjusted, *I:394*
- Spreads
 absolute and relative change volatility, *I:396f*
 change in, *I:392, I:393, I:394f, I:399*
 determining for asset swaps, *I:227–228*
 level vs. volatility of, *I:397*
 measurement of, *II:336–337*
 measure of exposure to change in, *I:397*
 nominal, use of, *III:5*
 option-adjusted, *I:253–255, I:254f*
 reasons for, *I:210–211*
 relative vs. absolute modeling, *I:393*
 volatility vs. level, *I:394–396, I:395f*
 zero-volatility, *III:5*
- Squared Gaussian (SqG) model, *III:547–548*
- Square-root rule, *III:534*
- SR-SARV model class, *II:370*
- St. Petersburg paradox, *III:480*
- Stability
 notion of, *II:667*
 in Paretian distribution, *II:739–741*
 property of, *II:740–741, III:236–237, III:244–245*
- Stable density functions, *III:236f*
- Stable Paretian model, α -stable distribution in, *II:748*
- Standard Default Assumption (SDA) convention, *III:59–60, III:60f*
- Standard deviations
 and covariance, *I:9*
 defined, *III:168*
 mean, *III:353*
 posterior, *I:155*
 related to variance, *III:203–204*
 rolling, *II:362–363*

- Standard deviations (*Continued*)
 and scale of possible outcomes,
III:168f
 for tail, *III:341*
- Standard errors. *See also* errors
 for average estimators, *III:400–402*
 defined, *III:399*
 estimation of, *III:640*
 of the estimator, *III:400*
 for exponentially weighted moving
 averages (EWMA), *III:411–412*
 reduction of, *III:648*
- Standard normality, testing for,
III:366–367
- Standard North American contract
 (SNAC), *I:529*
- Standard & Poors 500
 auto correlation functions of, *II:389t*
 cointegration regression, *II:390t*
 daily close, *III:402f*
 daily returns (2003), *III:326f*
 distributions of, *III:384f*
 error correction model, *II:391t*
 historical distributions of, *III:390f*
 index and dividends (1962–2006),
II:388f
 parameter estimates of, *III:385t*,
III:387t, *III:388t*
 return and excess return data
 (2005), *II:316–317t*
 stationarity test for, *II:389t*
 time scaling of, *III:383f*
 worst returns for, *III:382t*
- State dependent models (SDMs), *I:342*,
I:351–352
- Statement of stockholders' equity,
II:541
- State price deflators
 defined, *I:103*, *I:129–130*
 determining, *I:118–119*, *I:124*
 formulas for, *I:107–108*, *I:109–110*
 in multiperiod settings, *I:105*
 and trading strategy, *I:106*
- State prices
 and arbitrage, *I:55–56*
 condition, *I:54*
 defined, *I:101–102*
 and equivalent martingale
 measures, *I:133–134*
 vectors, *I:53–55*, *I:58*, *I:119*
- States, probabilities of, *I:115*
- States of the world, *I:457–458*, *I:459*,
II:306, *II:308*, *II:720*
- State space, *I:269n*
- Static factor models, *II:150*
- Stationary series, trend *vs.* difference,
II:512–513
- Stationary univariate moving average,
II:506
- Statistical concepts, importance of,
II:126–127
- Statistical factors, *II:177*
- Statistical learning, *II:298*
- Statistical methodology, EWMA,
III:409
- Statistical tests, inconsistencies in,
II:335–336
- Statistics, *II:387*, *II:499*
- Stein paradox, *I:194*
- Stein-Stein model, *II:682*
- Step-up callable notes, valuing of,
I:251–252
- Stochastic, defined, *III:162*
- Stochastic control (SC), *III:124*
- Stochastic differential equations
 (SDEs)
 binomial/trinomial solutions to,
III:610–613
 with change of time methods,
III:523
 defined, *II:658*
 examples of, *III:523–524*
 generalization to several
 dimensions with, *III:490–491*
 intuition behind, *III:486–487*
 modeling states of the world with,
III:127
 for MRAM equation, *III:525–526*
 setting of change of time, *III:521*
 solution of, *III:491–493*
 steps to definition, *III:487*
 usefulness of, *III:493*
 use of, *II:295*, *III:485–486*,
III:489–490, *III:536*, *III:603*,
III:619
- Stochastic discount factor, *I:57–58*
- Stochastic integrals
 defined, *III:481–482*
 intuition behind, *III:473–475*
 in Ito processes, *III:487*
 properties of, *III:482–483*
 steps in defining, *III:474–475*
- Stochastic processes
 behavior of, *I:262*
 characteristic function of, *III:496*
 characteristics of, *II:360*
 continuous-time, *III:496*, *III:506*
 defined, *I:263–264*, *I:269n*, *II:518*,
III:476, *III:496*
 discrete time, *II:501*
 properties of, *II:515*
 representation of, *II:514–515*
 and scaling, *III:279*
 specification of, *II:692–693*
- Stochastic programs
 features of, *III:124*, *III:132*
- Stochastic time series, linear,
II:401–402
- Stochastic volatility models (SVMs)
 with change of time, *III:520*
 continuous-time, *III:656*
 discrete, *III:656–657*
 importance of, *III:658*
 for modeling derivatives,
III:655–656
 multifactor models for, *III:657–658*
 and subordinators, *III:521–522*
 use of, *III:653*, *III:656*
- Stock indexes
 interim cash flows in, *I:482*
 risk control against, *II:262–263*
- Stock markets
 bubbles in, *II:386*
 as complex system, *II:47–48*
 1987 crash, *II:521*, *III:585–586*
 dynamic relationships among,
II:393–396
 effects of crises, *III:233–234*
 variables effects on different sectors
 of, *II:55*
- Stock options, valuation of long-term,
I:449
- Stock price models
 binomial, *III:161*, *III:171–173*, *III:173f*
 multinomial, *III:180–182*, *III:181f*,
III:184
 probability distribution of
 two-period, *III:181t*
- Stock prices
 anomalies in, *II:111t*
 behavior of, *II:58*
 correlation of, *I:92–93*
 and dividends, *II:4–5*
 lognormal, *III:655–656*
 processes of, *I:125*
- Stock research, main areas of, *II:244t*
- Stock returns, *II:56*, *II:159f*
- Stocks
 batting average of, *II:99*, *II:99f*
 characteristics of, *II:204*
 common, *II:4*, *II:316–322*
 cross-sectional, *II:197*
 defined, *II:106*
 defining parameters of, *II:49*
 determinants of, *II:245f*
 execution price of, *III:626*
 fair value *vs.* expected return, *II:13f*
 finding value for XYZ, Inc., *II:31t*
 information coefficient of, *II:98f*
 information sources for, *II:90f*
 measures of consistency, *II:99–100*
 mispriced, *II:6–7*
 quantitative research metrics tests,
II:97–99
 quintile spread of, *II:97f*
 relative ranking of, *I:196–197*
 review of correlations, *II:101f*

- sale/terminal price of, *II:5*
 short selling of, *I:432–433*
 similarities between, *II:245f*
 sorting of, *II:215*
 testing of, *II:95, II:96f*
 that pay no dividend, *II:17*
 use of, *II:90*
 valuation of, *II:6, II:8–9, II:14, II:18–19*
 weightings of, *II:101f*
- Stock selection**
 models for, *II:197*
 in quantitative equity investment process, *II:105*
 quantitative model, *II:94–95*
 for retail sector, *II:94f*
 strategies for, *II:195*
 tree for, *II:379–381, II:380f*
- Stopping times, *II:685***
Straontonovich, Ruslan, *II:470*
Strategies, backtesting of, *II:235–236*
Stress tests, *I:412, I:417, I:418, III:93, III:596–597*
Strike price, *I:509, I:514*
Strong Law of Large Numbers (SLLN), *I:270n, III:263–264*
Structural breaks, *I:167, III:274–275*
Student's *t* distribution
 applications to stock returns, *III:215–216*
 and AVaR, *III:334–335*
 classical, *II:734–738*
 density function of, *II:735*
 discussion of, *III:213–216*
 distribution function of, *III:215f*
 for downside risk estimation, *III:386–387*
 fitting and simulation of, *II:737–738*
 heavy tails of, *I:160n, I:176, II:747–748, II:751, III:227–228*
 limitations of, *II:736*
 in modeling credit risk, *I:387–388*
 normals representation in, *I:177–178*
 skewed, *II:736–737, II:753n*
 skewness of, *III:390*
 standard deviation of, *I:173n*
 symmetry of, *III:387*
 tails of, *III:392*
 use of, *I:153–154, I:172n, III:234*
- Student's *t*-test, *II:219***
Sturge's rule, *II:495*
Style analysis, *II:189*
Style factors, *II:247*
Style indexes, *II:48*
Stylized facts, *II:503–504*
Subadditivity property, *III:328*
Subordinated processes, *I:186n, III:277, III:521–522*
- Successive over relaxation (SOR) method, *II:677***
Summation stability property (Gaussian distribution), *II:732–733*
Supervisory Capital Assessment Program, *I:300, I:412*
Support, defined, *III:200*
Survey bias, *I:293*
Survival probability, *I:533–535*
Swap agreements, *I:434, I:435–436n*
Swap curves, *I:226, II:275–276*
Swap rates, *I:226, III:536f*
Swaps
 with change of time method, *III:522*
 covariance/correlation, *I:547–548, I:549–550, I:552*
 duration-matched, *I:285*
 freight rate, *I:558*
 modeling and pricing of, *I:548–550*
 summary of studies on, *I:546t*
 valuing of, *I:434–435*
- Swap spread (SS) risk, *II:278, II:278t***
Swaptions, *I:502–503, III:550*
Synergies, in conglomerates, *II:43–44*
Systematic risk, *II:290*
Systems
 homogenous, *II:624*
 linear, *II:624*
 types of, *II:47, II:58*
- Tailing the hedge, defined, *I:433***
Tail losses
 in loss functions, *III:369–370*
Tail probability, *III:320*
Tail risk, *I:377, I:385, II:752*
Tails
 across assets through time, *II:735–736*
 behavior of in operational losses, *III:111–112*
 in density functions, *III:203*
 dependence, *I:327–328, I:387*
 Gaussian, *III:98–99, III:260*
 heavy, *II:734–744, III:238*
 modeling heaviness of, *II:742–743*
 for normal and STS distributions, *III:246t*
 power tail decay property, *II:739, III:244*
 properties of, *III:261–262*
 tempering of, *II:741*
- Takeovers, probability of, *I:144–145***
Tangential contour lines, *I:29–30, I:30f, I:32f*
Tanker market, *I:565*
TAR-F test, *II:426*
TAR(1) series, simulated time plot of, *II:404f*
- Tatonnement, concept of, *II:468***
Taxes
 and bonds, *I:226*
 capital gains, *II:73*
 cash, *II:573*
 for cash/futures transactions, *I:484*
 complexity of, *II:73–74*
 deferred income, *II:535, II:538*
 effect on returns, *II:83–84, II:84, II:85n*
 in financial statements, *II:541*
 impact of, *I:286–287*
 incorporating expense of, *II:73–75*
 managing implications of, *III:146*
 and Treasury strips, *I:218*
- Tax policy risk, *II:282–283***
Technology, effect of on relative values, *II:37*
Telescoping futures strategy, *I:433*
Tempered stable distributions
 discussions of, *III:246–252, III:384–386*
 generalized (GTS), *III:249*
 Kim-Rachev (KRTS), *III:251–252*
 modified (MTS), *III:249–250*
 normal (NTS), *III:250–251*
 probability densities of, *III:247f, III:248f, III:250f, III:252f*
 rapidly decreasing (RDTS), *III:252*
 tempering function in, *III:254, III:258n*
- Tempered stable processes, *III:499–501, III:500t, III:512–517***
Tempering functions, *III:254, III:255t*
Templates, for data storage, *II:204*
Terminal profit, options and forwards, *I:438f, I:439f*
Terminal values, *II:45*
Terminology
 of delinquency, default and loss, *III:56*
 of prepayment, *III:49–50*
 standard, of tree models, *II:376*
- Term structure**
 in contiguous time, *III:572–573*
 continuous time models of, *III:570–571*
 defined, *III:560*
 eclectic theory of, *III:570*
 of forward rates, *III:586*
 mathematical relationships of, *III:562*
 modeling of, *I:490–494, III:560*
 of partial differential equations, *III:583–584*
 in real world, *III:568–570*
- Term structure modeling**
 applications of, *III:584–586*
 arbitrage-free, *III:594*

- Term structure modeling (*Continued*)
 calibration of, III:580–581
 discount function in, III:565
 discussion of, III:560–561
- Term structure models
 approaches to, III:603–604
 defined, I:262, I:263
 discrete time, III:562–563
 discussion of, III:561–562
 of interest rates, I:314
 internal consistency checks for, III:581
 with no mean reversion, III:613–616
 for OAS, I:265–267
 quantitative, III:563
 static *vs.* dynamic, III:561–562
- Term structures, III:567–568, III:570, III:579, III:587
- Tests
 Anderson-Darling (AD), III:112–113
 BDS statistic, II:423–424, II:427
 bispectral, II:422–423
 cointegration, II:708–710
 Kolmogorov-Smirnov (KS), III:112–113
 monotonic relation (MR), II:219
 nonlinearity, II:426–427, II:427^t
 nonparametric, II:422–424
 out-of-sample *vs.* in-sample, II:236
 parametric, II:424–426
 RESET, II:424–425
 run tests, III:364
 threshold, II:425–426
 for uniformity, III:366
- TEV (tracking error volatility), II:180, II:186, II:272–274, II:286–287
- Theil-Sen regression algorithm, II:440–442, II:443–446, II:444^t
- The Internal Measurement Approach (BIS), III:100n
- Theoretical value, determination of, III:10–11
- Théorie de la Spéculation (The Theory of Speculation)* (Bachelier), II:121–122, II:469
- Theory of point processes, II:470–471
- Three Mile Island power plant crisis, II:51–52
- Three-stage growth model, II:9–10
- Threshold autoregressive (TAR) models, II:404–408
- Thresholds, II:746–747
- Through the cycle, defined, I:302–303, I:309–310
- Thurstone, Louis Leon, II:154
- Tick data. *See* high-frequency data (HFD)
- Time
 in differential equations, II:643–644
 physical *vs.* intrinsic scales of, II:742
 use of for financial data, II:546–547
- Time aggregation, II:369
- Time decay, I:509, I:513, I:521^f
- Time dependency, capture of, II:362–363
- Time discretization, II:666, II:679
- Time increments
 models of, I:79
 in parameter estimation, I:83
- Time intervals, size of, II:300–301
- Time lags, II:299–300
- Time points, spacing of, II:501
- Time premiums, I:485
- Time series
 autocorrelation of, II:331
 causal, II:504
 concepts of, II:501–503
 continuity of, I:80
 defined, II:501–502, II:519
 fractal nature of, III:480
 importance of, II:360
 multivariate, II:502
 stationary, II:502
 stationary/nonstationary, II:299
 for stock prices, II:296
- Time to expiry, I:513
- Time value, I:513, I:513^f, II:595–596
- TIPS (Treasury inflation-protected securities)
 and after-tax inflation risk, I:287
 apparent real yield premium, I:293^f
 effect of inflation and flexible price CPI, I:292^f
 features of, I:277
 and flexible price CPI, I:291^f
 and inflation, I:290, I:294
 performance link with short-term inflation, I:291–292
 real yields on, I:278
 spread to nominal yield curve, I:281^f
 volatility of, I:288–290, I:294
vs. real yield, I:293–294
 10-year data, I:279–280
 yield of, I:284
 yields from, I:278
- TLF model, strengths of, III:388–389
- Total asset turnover ratio, II:558
- Total return reports, II:237^t
- Total return swaps, I:540–542, I:541–542
- Trace test statistic, II:392
- Tracking error
 actual *vs.* predicted, II:69
 alternate definitions of, II:67–68
 defined, II:115, II:119
 estimates of future, II:69
 as measure of consistency, II:99–100
 reduction of, II:262–263
 standard definition, II:67
 with TIPS, I:293
- Tracking error volatility (TEV). *See* TEV (tracking error volatility)
- Trade optimizers, role of, II:116–117
- Trades
 amount needed for market impact, III:624
 cash-and-carry, I:487
 crossing of, II:75
 importance of execution of, III:623, III:631
 measurement of size, III:628
 in portfolio construction, II:104, II:116–117
 round-trip time of, II:451
 size effects of, III:372, III:630
 speed of, II:105
 timing of, III:628–629
- Trading costs, II:118, III:627–628, III:631–632
- Trading gains, defined, I:122, I:123
- Trading horizons, extending, III:624
- Trading lists, II:289^t
- Trading strategies
 backtesting of, II:236–237
 categories of, II:195
 in continuous-state, continuous-time, I:122
 development of factor-based, II:197–198, II:211
 factor-based, II:195, II:232–235
 factor weights in, II:233^f
 in multiperiod settings, I:105
 risk to, II:198–200
 self-financing, I:126–127, I:136
- Trading venues, electronic, II:57
- Training windows, moving, II:713–714
- Tranches, III:38, III:39^t, III:45
- Transaction costs
 in backtesting, II:235
 in benchmarking, II:67
 components of, II:119
 consideration of, II:64, II:85–86n
 dimensions of, III:631
 effect of, I:483
 figuring, II:85n
 fixed, II:72–73
 forecasting of, II:113–114
 incorporation of, II:69–73, II:84
 international, III:629
 linear, II:70
 and liquidity, III:624–625
 managing, III:146
 measurement of, III:626
 piecewise-linear, II:70–72, II:71^f

- quadratic, *II:72*
 in risk modeling, *II:693*
 types of, *III:623*
- Transformations, nonlinear,
III:630–631
- Transition probabilities, *I:368, I:381t*
- Treasuries
 correlations of, *III:405t*
 covariance matrix of, *III:406t*
 curve risk, *II:277t*
 discount function for, *III:564–565*
 futures, *I:482*
 inflation-indexed, *I:286*
 movements of, *III:403f*
 on-the-run, *I:227, III:7, III:560*
 par yield curve, *I:218t*
 spot rates, *I:220*
 3-month, *II:415–416, II:416f*
 volatility of, *III:404–406, III:406t*
- Treasury bill rates, weekly data, *I:89f*
- Treasury inflation-protected securities (TIPS). *See* TIPS (Treasury inflation-protected securities)
- Treasury Regulation T (Reg T), *I:67*
- Treasury securities, *I:210–211*
 comparable, defined, *III:5*
 in futures contracts, *I:483*
 hypothetical, illustration of duration/convexity, *III:308–310, III:308t*
 maturities of, *I:226*
 options on, *I:490*
 par rates for, *I:217*
 prediction of 10-year yield, *II:322–328*
 valuation of, *I:216*
 yield of, *II:324–327t*
- Treasury strips, *I:218t, I:220–221, I:286, III:560*
- Treasury yield curves, *I:226, III:561*
- Trees/lattices
 adjusted to current market price, *I:496f*
 bushy trees, *I:265, I:266f*
 calibrated, *I:495*
 convertible bond value, *I:274–275*
 extended pricing tree, *III:23f*
 from historical data, *III:131f*
 pruning of, *II:377*
 stock price, *I:274*
 three-period scenario, *III:131f*
 trinomial, *I:81, I:273, I:495–496*
 use of in modeling, *I:494–496*
- Trees/lattices, binomial
 building of, *I:273*
 for convertible bonds, *I:275f*
 discussion of, *I:80–81*
 interest rate, *I:244*
 model of, *I:273–275*
- stock price model, *III:173*
- term structure evolution, *I:495f*
 use of, *I:114–115, I:114f*
- Trends
 deterministic, *II:383*
 in financial time series, *II:504*
 and integrated series, *II:512–514*
 stochastic, *II:383, II:384*
- Treynor-Black model, *I:203n*
- Trinomial stochastic models, *II:11–12*
- Truncated Lévy flight (TLF), *III:382, III:384–386*
 IDD in, *III:386*
 time scaling of, *III:385f*
- Truncation, *III:385–386*
- Truth in Savings Act, *II:615*
- T*-statistic, *II:240n, II:336, II:350, II:390*
- Tuple, defined, *III:157*
- Turnover
 assessment of, *III:68*
 defined, *III:66*
 in MBSs, *III:48*
 in portfolios, *II:234, II:235*
- Two beta trap, *I:74–77*
- Two-factor models, *III:553–554*
- Two-stage growth model, *II:9*
- U.K. index-linked gilts, tax treatment of, *I:287*
- Uncertainties
 and Bayesian statistics, *I:140*
 in measurement processes, *II:367*
 modeling of, *II:306, III:124, III:131–132*
 and model risk, *II:729*
 quantification of, *I:101*
 representation of, *III:128*
 time behavior of, *II:359*
- Uncertainty sets
 effect of size of, *III:143*
 in portfolio allocation, *II:80*
 selection of, *III:140–141*
 structured, *III:143–144*
 in three dimensions, *II:81f*
 use of, *III:138, III:140*
- Uncertain volatility model, *II:673–674*
- Underperformance, finding reasons for, *II:118*
- Underwater, on homeowner's equity, *III:73*
- Unemployment rate
 as an economic measure, *II:398*
 application of TAR models to, *II:405–406*
 characteristics of series, *II:430*
 forecasts from, *II:433*
 performance of forecasting, *II:432–433, II:432t*
 and risk, *II:292n*
- test of nonlinearity, *II:431, II:431t*
- time plot of, *II:406f, II:430f*
- Uniqueness, theorem of, *III:490*
- Unit root series, *II:385*
- Univariate linear regression model, *I:163–170*
- Univariate stationary series, *II:504*
- U.S. Bankruptcy Code. *See also* bankruptcy
 Chapter 7, *I:350*
 Chapter 11, *I:342, I:350*
 Utility, *I:56, II:469, II:471, II:719–720*
- Validation, out of sample, *II:711*
- Valuation
 arbitrage-free, *I:216–217, I:220–222, I:221t*
 and cash flows, *I:223*
 defined, *I:209*
 effect of business cycle on, *I:303–304*
 fundamental principle of, *I:209*
 with Monte Carlo simulation, *III:6–12*
 of natural gas/oil storage, *I:560–561*
 of non-Treasury securities, *I:222–223*
 relative, *I:225, II:34–40, II:44–45*
 risk-neutral, *I:557, III:595–596, III:601*
 total firm, *II:21–23*
 uncertainty in, *II:15*
 use of lattices for, *I:240*
- Value
 absolute *vs.* relative basis of, *I:259–260*
 analysis of relative, *I:225*
 arbitrage-free, *I:221*
 book *vs.* market of firms, *II:559–560*
 determining present, *II:600–601*
 formulas for analysis of, *II:238–239*
 identification of relative, *I:405*
 intrinsic, *I:484–485*
 present, discounted, *II:601f*
 relative, *I:405, II:37–38*
vs. price, *I:455n*
- Value at risk (VaR). *See also* CVaR (credit value at risk)
 in backtesting, *II:748*
 backtesting of, *II:749f, III:325–327, III:365–367*
 boxplot of, *III:325f*
 and coherent risk measures, *III:329*
 conditional, *III:332, III:355–356, III:382*
 deficiencies in, *I:407, III:321, III:331–332, III:347*
 defined, *II:754n, III:319–322*
 density and distribution functions, *III:320f*

- Value at risk (VaR) (*Continued*)
determining from simulation, III:639f
distribution-free confidence intervals for, III:292–293
estimation of, II:366, III:289–290, III:373–376, III:644, III:644t
exceedances of, III:325–326
IDD in, III:290
interest rate covariance matrix in, III:403
levels of confidence with, III:290–291
liquidity-adjusted, III:374, III:376
in low market volatility, II:748
measurements by, II:354
methods of computation, III:323
modeling of, II:130–131, III:375–376
and model risk, II:695
normal against confidence level, III:294f
portfolio problem, I:193
in practice, III:321–325
relative spreads between predictions, II:750f, II:751f, II:752f
as safety-first risk measure, III:355
standard normal distribution of, III:324t
use of, II:365
vs. deviation measures, III:320–321
- Value of operations, process for finding, II:30t
- Values, lagged, II:130
- Van der Korput sequences, III:650
- Variables
antithetic, III:647–648
application of macro, II:193n
behavior of, III:152–153
categorical, II:333–334, II:350
classification, II:176
declaration of in VBL, III:457–458
dependence between, II:306–307
dependent categorical, II:348–350
dependent/independent in CAPM, I:67
dichotomous, II:350
dummy, II:334
exogenous *vs.* endogenous, II:692
fat-tailed, III:280
independent and identically distributed, II:125
independent categorical, II:333–348
interactions between, II:378
large numbers of, II:147
macroeconomic, II:54–55, II:177
in maximum likelihood calculations, II:312–313
mixing of categorical and quantitative, II:334–335
nonstationary, II:388–393
as observation or measurement, II:306
random, I:159n
in regression analysis, II:330
separable, II:647
slope, III:553
split formation of, III:130f
spread, II:336
standardization of, II:205
stationary, II:385, II:386
stationary/nonstationary, II:384–386
stochastic, III:159–164
use of dummy, II:335, II:343–344
- Variables, random, II:297
 α -stable, III:242–244, III:244–245
Bernoulli, III:169
continuous, III:200–201, III:205–206
on countable spaces, III:160–161, III:166
defined, III:162
discrete, III:165
infinitely divisible, III:253
in probability, III:159–164
sequences of, I:389
on uncountable spaces, III:161–162
use of, I:82
- Variance gamma process, III:499, III:504
- Variance matrix, II:370–371
- Variances
addressing inequality of, I:168
based on covariance matrix, II:161t, II:163t, II:164f
conditional, I:180
conditional/unconditional, II:361
in dispersion parameters, III:202–203
equal, I:164
as measure of risk, I:8
in probability, III:167–169
reduction in, III:647–651
unequal, I:167–168, I:172
- Variances/covariances, II:112–113, II:302–303, III:395–396
- Variance swaps, I:545–547, I:549, I:552
- Variational formulation, and finite element space, II:670–672
- Variation margins, I:478
- Vasicek model
with change of time, III:523–524
for coupon-bond call options, I:501–502
distribution of, I:493
in history, I:491
for short rates, III:545–546
use of, I:89, I:497
valuing zero-coupon bond calls with, I:499–500
- VBA (Visual Basic for Applications)
built-in numeric functions of, III:456
comments in, III:453
control flow statements, III:458–460
debugging in, III:461
debugging tools of, III:461, III:477
example programs, III:449–452, III:461–466
in Excel, III:449, III:450f
FactorialFun1, III:455–456
functions, user-defined, III:463f
functions in, III:477
generating Brownian motion paths in, III:463–465
If statements, III:459
For loops, III:458–459
methods (actions) in, III:452–453
modules, defined, III:455
as object-oriented language, III:452, III:466
objects in, III:452
operators in, III:459–460
Option Explicit command, III:458
pricing European call options, III:465–466
programing of input dialog boxes, III:460–461
programming tips for, III:454–461
properties in, III:453
random numbers in, III:464–465
subroutines and user-defined functions in, III:466–477
subroutines *vs.* user-defined functions in, III:455–457
use of Option Explicit command, III:458
user-defined functions, III:463f
user interaction with, III:460–461
variable declaration in, III:457–458
With/End structure in, III:453–454
writing code in, III:453–454
- Vech notation, II:371–372
- VEC model, II:372
- Vector autoregressive (VAR) model, II:393
- Vectors, II:621–622, II:625–626, II:628
- Vega, I:521
- Vichara Technology, III:41–42, III:43t
- Visual Basic for Applications (VBA). *See* VBA
- Volatilities
absolute *vs.* relative, III:404–405
actual, I:514
aim of models of, I:176
analysis of, II:270–272
and ARCH models, II:409

- assumptions about, *III:7*
 calculation of, *II:272, III:534t*
 calculation of daily, *III:533–534*
 calibration of local, *II:681–685*
 clustering of, *II:359, II:716, III:402*
 confidence intervals for, *III:399–400*
 constant, *III:653*
 decisions for measuring, *III:403–404*
 defined, *III:533, III:653*
 with different mean reversions,
III:538f
 of the diffusion, *I:125*
 effect of local, *III:609*
 effect on hedging, *I:517–518*
 of energy commodities, *I:556–557*
 estimation of, *II:368–369*
 in EWMA estimates, *III:410–411*
 exposure to, *II:252f, II:252t*
 forecasts of, *I:179–180, II:172,*
II:367–368
 in FTSE 100, *III:412–413*
 historical, *I:513, III:534, III:654*
 hypothetical modelers of, *III:408*
 implied, *I:513–514, II:282, II:662,*
III:654
 in interest rate structure models,
I:492
 jump-diffusion, *III:657*
 level-dependent, *III:654–655,*
III:656
 local, *II:681, II:682–683, III:655*
 as a measure, *I:545, II:373*
 measurement of, *I:393, III:403–406*
 minimization of, *II:179*
 in models, *II:302*
 models of, *II:428*
 in option pricing, *I:513–514*
 patterns in, *I:395*
 in random walks, *I:84*
 and risk, *II:270*
 in risk-neutral measures, *III:587*
 smile of, *III:557*
 and the smoothing constant,
III:409–410
 states of, *I:180–181*
 stochastic, *I:94, I:547, I:548,*
III:655–658, III:656, III:658
 stochastic models, *II:681*
 time increments of, *I:83*
 of time series, *I:80*
 time-varying, *II:733–734*
 types of, *III:658*
 vs. annual standard deviation,
III:534
 Volatility clustering, *III:242, III:388*
 Volatility curves, *III:534–535,*
III:535t
- Volatility measures, nonstochastic,
III:654–655
 Volatility multiples, use of,
III:536
 Volatility risk, *I:509*
 Volatility skew, *III:550, III:551f,*
III:555–556, III:654
 measuring, *III:550*
 Volatility smile, *II:681, III:555–557,*
III:556f, III:654, III:656
 Volatility swaps, *I:545–547, I:552*
 for S&P Canada index (example),
I:550–552
 valuing of, *I:549*
 Volume-weighted average price
 (VWAP), *II:117, III:626–627*
 VPRs (voluntary prepayment rates)
 calculation of, *III:76*
 in cash flow calculators, *III:34*
 defined, *III:30*
 impacts of, *III:38*
- W. T. Grant, cash flows of, *II:576*
 Waldrop, Mitchell, *II:699*
 Wal-Mart, *II:569, II:570f*
 Walras, Leon, *II:467, II:468–469,*
II:474
 Waterfalls, development of, *III:8*
 Weak laws of large numbers (WLLN),
III:263
 Wealth, *I:460t, III:130*
 Weather, as chaotic system, *II:653*
 Weibull density, *III:107f*
 Weibull distributions, *III:106–107,*
III:112, III:229, III:262, III:265,
III:267, III:268
 Weighting, efficient, *I:41–42*
 Weights, *II:115, II:185t, II:231–232,*
II:724
 Weirton Steel, cash flows of,
II:577f
 What's the hedge, *I:300, I:303, I:306,*
I:417. See also hedge test
 White noise. *See noise, white*
 Wiener processes, *I:95, I:491, I:497,*
III:534–535, III:579, III:581
 Wilson, Kenneth, *II:480*
 Wind farms, valuation of, *I:563–564*
 Wold representation, *II:506*
 Working capital, *II:551*
 concept of, *II:567*
- XML (eXtensible Markup Language),
 development of, *II:482*
- Yield and bond loss matrix, *III:41f*
 Yield curve risk, *III:307, III:316–317*
- Yield curves
 horizon, *III:585*
 initial consistency with, *III:544*
 issuer par, *I:238t, I:244t*
 nonparallel, *III:309–310*
 parallel shifts in, *III:308–309*
 par-coupon, *III:585*
 reshaping duration, *III:315–316*
 in scenario analysis, *II:290*
 SEDUR/LEDUR, *III:316, III:317*
 shifts in, *III:586*
 slope of, *III:315*
 in term structures, *III:560*
 in valuation, *I:235*
- Yields
 calculation of, *II:613–618*
 comparison across countries, *I:226*
 dividend, *II:4*
 on investments, *II:617–618, II:619*
 loss-adjusted, *III:36, III:40*
 and loss matrix analysis, *III:40–41*
 projected, *III:37f, III:38f*
 real, *I:278–280, I:280f*
 rolling, *I:258–259*
- Yield spreads
 computation of, *I:226*
 determining, *I:373–374*
 for different rating grades, *I:374t*
 in Merton model, *I:305–306*
 over swap and treasury curves,
I:226–227
- Zero-coupon bonds
 assumptions about, *I:261*
 calculations using CIR model, *I:502t*
 calculations using Vasicek model,
I:502t
 defaultable, *I:317, I:335n*
 default-free, *I:318*
 development of valuation model
 for, *III:582–583*
 equations for, *III:554*
 future market price for, *I:492–493*
 lattices for, *I:266f*
 market for, *I:264*
 and martingales, *I:262*
 PDEs of, *I:268–269n*
 pricing of, *I:316*
 term structure model for, *III:584*
 value of, *III:572–573*
 valuing, *I:213, I:499–501, I:499t*
 Zero coupon rates, *III:546–547*
 Zero coupon securities, *I:218*
 Zero one distribution, *III:169–170*
 Zero volatility spread, *III:11–12*
 Zipf's law, *III:263, III:269*
 Z-scores, *II:191, II:240n*